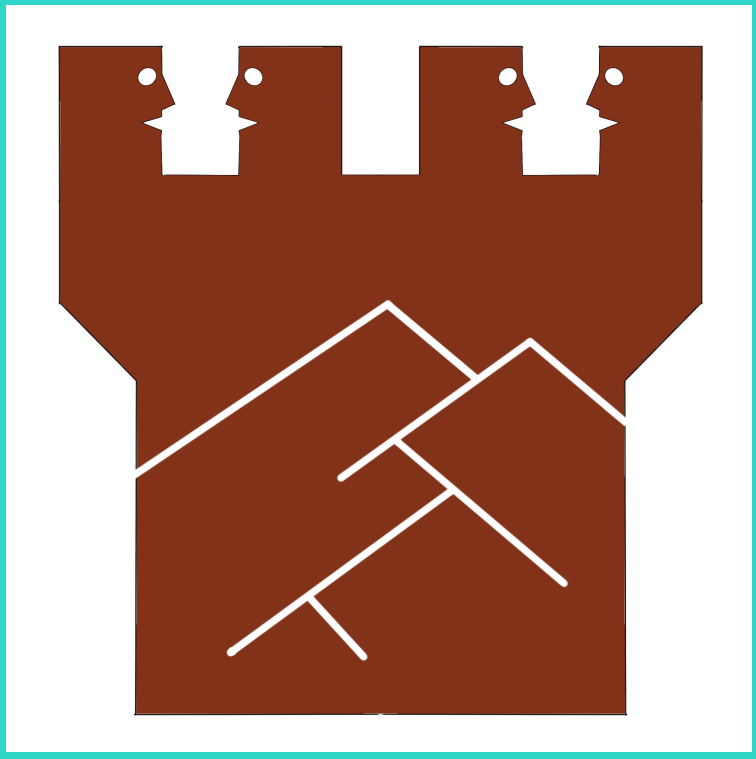
SemDial 2024

Trento Logue



Proceedings of the 28th Workshop
On the Semantics and Pragmatics of Dialogue

Held at
University of Trento, Italy
September 11-12 2024

Raffaella Bernardi, Eleni Breitholtz, & Giuseppe Riccardi (eds.)

UNIVERSITÀ DI TRENTO
CIMeC Center for Mind/Brain Sciences

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.semdial.org/

Co-presidents: Ellen Breitholtz and Julian Hough

Anthologists: Christine Howes, Casey Kennington and Brielen Madureira

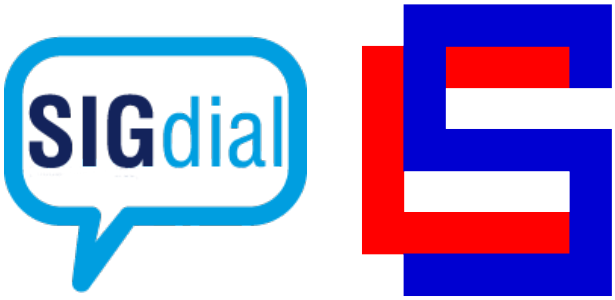
Webmasters: Janosch Haber, Julian Hough

TrentoLogue Website

https://event.unitn.it/semdial2024/

TrentoLogue Sponsors

TrentoLogue Endorsements



Preface

CIMeC, the Center for Mind/Brain Sciences at the University of Trento, Italy, is proud to host SemDial 2024, the international workshop on the semantics and pragmatics of dialogue. CIMeC, based in the town of Rovereto, 24 Km south of the university's main location in Trento and a short distance from the north tip of Lake Garda, is one of Italy's most important venues for cognitive science and neuroscience, hosting cutting-edge research facilities for brain imaging (fMRI, MAG, TMS, two-photon microscopy) in the newly refurbished area of Manifattura Tabacchi. Since its inception, it also includes groups working on computational models of human language that combine insights from the generative theoretical perspectives with predictive ML-based models of the combination of language and vision. SemDial fits right in with this tradition, and we expect it to provide stimulating input for the students in our international Master in Cognitive Science and PhD program.

SemDial 20204 - Trentologue - received 34 full paper submissions, 14 of which were accepted as full papers after a peer-review process, during which each submission was reviewed by a panel of at least two experts. The poster abstracts had 35 submissions from a combination of recommended pre-accepted re-submissions of long papers and a further call for research in progress and short papers - 32 of these poster abstracts were presented. All accepted full papers and poster abstracts are included in this volume.

We would like to extend our thanks to the authors who contributed their work and to our Programme Committee members for their very detailed and helpful reviews!

Trentologue features three keynote presentations by Uri Hasson, Professor at the Department of Psychology and the Neuroscience Institute, Princeton University; Azzurra Ruggeri, Associate Professor at the Department of Cognitive Science, Central European University, Vienna and at Technical University Munich; Bernardo Magnini, Senior Researcher in Computational Linguistics at the Fondazione Bruno Kessler, Trento. We are honoured to have them in this year's SemDial and we thank them for their participation. Abstracts of their contributions are also included in this volume.

The event is endorsed by the Erasmus Mundus Program in Language and Communication Technologies.

Raffaella Bernardi, Eleni Breitholtz, and Giuseppe Riccardi

Rovereto

September 2024

Programme Committee

Raffaella Bernardi (chair)	University of Trento
Ellen Breitholtz (chair)	University of Gothenburg
Giuseppe Riccardi (chair)	University of Trento
Jedediah Allen	Bilkent University
Maxime Amblard	Université de Lorraine
Mattias Appelgren	University of Gothenburg
Ron Artstein	University of Southern California
Timo Baumann	OTH Regensburg
Raffaella Bernardi	University of Trento
Raffaella Bernardi	University of Trento
Maria Boritchev	Télécom Paris
Harry Bunt	Tilburg University
Heather Burnett	CNRS-Université de Paris 7
Eve V. Clark	Stanford University
Robin Cooper	University of Gothenburg
Mathilde Dargnat	Nancy University and ATILF-CNRS
Valeria de Paiva	Samsung Research America and University of Birmingham
Amandine Decker	Université de Lorraine, University of Gothenburg
Paul Dekker	ILLC/University of Amsterdam
Simon Dobnik	University of Gothenburg
Arash Eshghi	Heriot-Watt University
Raquel Fernández	University of Amsterdam
Victor Ferreira	University of California San Diego
Kallirroi Georgila	University of Southern California
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	University of Gothenburg, King's College London
Amy Han Qiu	University of Gothenburg
Julian Hough	Swansea University
Christine Howes	University of Gothenburg
Nikolai Ilinykh	University of Gothenburg
Amy Isard	University of Hamburg
Casey Kennington	Boise State University
Staffan Larsson	University of Gothenburg
Pierre Lison	Norwegian Computing Center
Andy Lücking	Goethe University Frankfurt
Bernardo Magnini	FBK-irst
Danielle Matthews	The University of Sheffield
Chiara Mazzocconi	Institute of Language, Communication, and the Brain, Laboratoire Parole et Langage - Aix-Marseille University
Philippe Muller	IRIT, Toulouse University
Bill Noble	University of Gothenburg

Dimitri Ognibene	Università degli Studi di Milano Bicocca
David Pagmar	University of Gothenburg, University of Stockholm
Paul Piwek	The Open University
Massimo Poesio	Queen Mary University of London
Laurent Prévot	Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309
Matthew Purver	Queen Mary University of London
James Pustejovsky	Computer Science Department, Brandeis University
Robert Ross	Technological University Dublin
David Schlangen	University of Potsdam
Gabriel Skantze	KTH Royal Institute of Technology
Massimo Stella	University of Trento
Matthew Stone	Rutgers University
Peter Sutton	Universitat Pompeu Fabra
Jakub Szymanik	University of Trento
David Traum	ICT USC
Enric Vallduvi	Universitat Pompeu Fabra
Nigel Ward	UTEP
Grégoire Winterstein	Université du Québec à Montréal
Roberto Zamparelli	University of Trento

Local Organizing Committee

Raffaella Bernardi	CIMeC, University of Trento
Jakub Szymanik	CIMeC, University of Trento
Roberto Zamparelli	CIMeC, University of Trento
Vanessa Maria Caleca	CIMeC, University of Trento

Table of Contents

Invited Talks

Full Papers

Poster Abstracts

Invited Talks

Deep language models as a cognitive model for natural language processing in the human brain

Uri Hasson
Princenton University
hasson@princeton.edu

Naturalistic experimental paradigms in cognitive neuroscience arose from a pressure to test, in real-world contexts, the validity of models we derive from highly controlled laboratory experiments. In many cases, however, such efforts led to the realization that models (i.e., explanatory principles) developed under particular experimental manipulations fail to capture many aspects of reality (variance) in the real world. Recent advances in artificial neural networks provide an alternative computational framework for modeling cognition in natural contexts. In this talk, I will ask whether the human brain's underlying computations are similar or different from the underlying computations in deep neural networks, focusing on the underlying neural process that supports natural language processing in adults and language development in children. I will provide evidence for some shared computational principles between deep language models and the neural code for natural language processing in the human brain. This indicates that, to some extent, the brain relies on overparameterized optimization methods to comprehend and produce language. At the same time, I will present evidence that the brain differs from deep language models as speakers try to convey new ideas and thoughts. Finally, I will discuss our ongoing attempt to use deep acoustic-to-speech-to-language models to model language acquisition in children.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Emergence and Developmental trajectory of Ecological Active Learning

Azzurra Ruggeri
Technical University Munich and
Central European University
ruggeri@mpib-berlin.mpg.de

The internet has made information available at our fingertips at all times: Search engines, accessed via our computers, tablets, or smart phones, allow us to look up things whenever and wherever we want—an enhanced encyclopedia of factual knowledge. This pseudo-infinite space of immediately available knowledge has drastically reduced our need to learn and memorize facts. However, it has increased the urgency to know how to navigate this space effectively. This revolution, triggered by the era of globalization and digitalization, calls for a new science of learning, one that is more focused on how and what to learn—how to effectively ask questions and explore, which sources of information to trust and rely on, how to adapt one's learning strategies to dynamic and multimodal learning environments, how to interpret the information collected, integrating them in one's existing body of knowledge—rather than on standard learning contents. This talk presents the results of recent studies investigating theoretically and empirically the emergence of these abilities, their developmental trajectory across childhood and the factors impacting their success.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Toward collaborative LLMs: Investigating Proactivity in Task-Oriented Dialogues

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Large Language Models (LLMs) promise a huge impact on dialogue generation, including the capacity to mimic human-like collaborative behaviors. However, current data-driven dialogue models present a significant lack of some fundamental properties of collaborative human interaction, such as grounding, clarifying questions, and proactive behavior. Obtaining human-like collaborative behaviors from LLMs reveals itself more complex than expected. In addition, such collaborative phenomena are also poorly investigated from a theoretical point of view, and there is a general need of empirical data, both quantitative and qualitative. In the talk, we focus on proactivity, a characteristic phenomenon of collaborative human-human interaction, where a participant in the dialogue offers the addressee some useful and not explicitly requested information. We report an extensive analysis of proactivity in several task-oriented dialogic corpora, selected with different characteristics. There are several findings from our empirical investigation of proactivity. We found that about 20% of turns in our corpus are proactive turns, showing that this is a very diffused and relevant phenomenon. We collected evidence confirming the non-reactive nature of proactivity, highlighting the presence of a pattern where a turn triggers a reaction in a following turn and a proactive utterance is then added to the turn. Finally, we empirically confirmed that proactivity has a crucial role in recovering from goal-failure situations, contributing to the whole dialogue effectiveness.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Full Papers

Turn-taking dynamics across different phases of explanatory dialogues

Petra Wagner[1,5], Marcin Włodarczak[2], Hendrik Buschmeier[3,5], Olcay Türk[1,5], Emer Gilmartin[4]

[1]Phonetics Workgroup, Faculty of Linguistics and Literary Studies, Bielefeld University
[2]Stockholm University, Stockholm, Sweden
[3]Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University
[4]INRIA Paris, Paris, France
[5]SFB/Transregio 318 Constructing Explainability, Paderborn and Bielefeld, Germany

Abstract

We examined the turn-taking dynamics across different phases of explanatory dialogues, in which 21 different explainers explained a board game to 2–3 explainees each. Turn-taking dynamics are investigated focusing on >19K floor transitions, i.e., the detailed patterns characterizing turn keeping or turn yielding events (Gilmartin et al., 2020). The explanations were characterized by three different phases (board game absent, board game present, interactive game play), for which we observed differences in turn-taking dynamics: explanations where the board game is absent are characterized by less complex floor transitions, while explanations with a concretely shared reference space are characterized by more complex floor transitions, as well as more floor transitions between interlocutors. Also, the speakers' dialogue role (explainer vs. explainee) appears to have a strong impact on turn-taking dynamics, as floor transitions that do not conform with the dialogue role tend to involve more effort, or floor management work.

1 Introduction

1.1 Floor transitions as indicator of different interactions and interaction styles

Floor management, the organization of the back and forth of the conversational floor between interlocutors, is no regular "ping pong game", during which the contributions of the conversation partners are neatly arranged in consecutive turns clearly delimited by minimally overlapping speech or very short pauses. Rather, periods when floor ownership can be clearly determined, with a single speaker producing solo (non-overlapping) speech, are often separated by a succession of shorter utterances, silences, and overlaps. Despite this, the bulk of the existing turn-taking literature has focused on strictly local descriptions of turn-taking centered around individual instances of silence or overlap, thus losing track of these extended patterns of floor negotiation (Sacks et al., 1974; Heldner and Edlund, 2010; Stivers et al., 2009). By not taking into account the diversity and complexity in how the floor is negotiated, research may easily overlook patterns that characterize more monological ("chunking") or more interactive ("chatting") phases of conversations, but also differences in language-specific interaction patterns such as the typical frequency of vocalized feedback or backchanneling (Dingemanse and Liesenfeld, 2022).

To overcome this apparent limitation, Gilmartin et al. (2020) proposed an alternative description of dialogue state in terms of *floor transitions*. Each floor transition consists of two longer intervals of solo speech exceeding some predefined duration (e.g., 1 second) separated by a series of *intervening intervals*: silences, overlaps, or shorter stretches of solo speech. Depending on whether or not they are associated with a speaker change, floor transitions can be furthered classified as between- or within-speaker (BST and WST, respectively). Previous work has demonstrated that both dyadic and multiparty conversations are greatly varied in terms of the floor transition patterns, with the majority of transitions involving more complex patterns of speech and silence than assumed by simple accounts of turn change and retention (Gilmartin et al., 2020; Włodarczak and Gilmartin, 2021; Gilmartin and Włodarczak, 2023).

One point to note is that across different corpora and interactions the vast bulk of floor state transitions between stretches of single party speech (BSTs and WSTs) have been found to involve odd numbers of intervening intervals. This is due to the very low probability of finding *exact* 'smooth switches' in the data – where one speaker starts speaking at exactly the same moment as another stops or where two or more speakers start and stop speaking at the same time.

Analysis of long multiparty casual conversations

has furthermore identified alternating phases differing in the length, composition in terms of speech, silence and overlap, the relative frequencies, and in the distribution of floor state transitions (Gilmartin et al., 2018). This is broadly in line with the findings of conversation analysis of multiparty casual talk, which has noted that conversations comprise a mixture of two different structural subgenres or phases – stretches of highly interactive chat with participation from several speakers, and longer almost monologic chunks (often narrative or expositionary – anecdotes, recounting of experience, . . .) where one speaker dominates and others mostly provide feedback (Eggins and Slade, 1997).

Between-speaker transitions in chat interaction were spread over more intervening intervals than in chunk, thus increasing the frequency of more complex transitions. This could reflect more turn competition, or indeed more backchannels and acknowledgment tokens being contributed by more participants. One-interval transitions comprised the largest class, with a higher proportion of one-interval transitions in chunk than chat, and higher proportions of within speaker than between-speaker one-interval transitions in both, but particularly in monologic chunk.

A comparison of multi-party conversations and the dyadic phone conversations showed less silence and overlap in dyadic conversations (Gilmartin and Włodarczak, 2023). Also, dyadic interactions showed comparatively fewer occurrences of floor transitions with multiple intervals. However, it is unclear whether these results are mainly influenced by the number of speakers participating in the conversation, or whether the lack of the visual channel may have an independent influence: on the phone, speakers may wait for their interlocutor to finish before commencing to speak, and may not give as much verbal feedback in overlap.

1.2 Explanations as a special case of dialogues

Turn-taking has been investigated for dialogue generally (Sacks et al., 1974), for specific types of dialogue (free: Gilmartin et al. 2018, task-oriented: Gravano and Hirschberg 2011, chaired: Larrue and Trognon 1993), and for different types of interaction partners (e.g., children: Garvey and Berninger 1981, artificial conversational agents: Skantze 2021). In this paper, we examine the floor transitions in explanatory dialogues. These constitute a special case of dialogical interaction as they have interesting properties (they are task-oriented and goal-directed, but not too narrow and involve all participants) and are of practical interest and relevance to various fields such as health communication (Collins, 2005), education (Chi, 1996), explainable AI (Rohlfing et al., 2021), or human-robot interaction (Stange et al., 2022). In particular, we expect that successful explanations are not only shaped by an active explainer directed towards a passive explainee, but involve a high level of interaction, bidirectional monitoring and adaptation, or 'co-construction' of an ongoing explanation, with the collaborative goal of reaching understanding (Rohlfing et al., 2021). Fisher et al. (2022) could show for naturally occurring explanatory dialogues between doctors and patients, that explanations may contain both more monological and more dialogical phases, and such phases can be initiated independently of the conversational role. However, it is yet unclear whether and how these explanatory phases can be straightforwardly related to distinct floor transition patterns.

1.3 Research questions

First, we are aiming to discover whether the floor transition patterns we find for explanatory dialogues differ from those found for less constrained, free conversation such as in Switchboard (Godfrey et al., 1992). Second, we are interested in finding out whether floor transitions in explanations can reflect different phases (e.g., chatting vs. chunking) in an ongoing explanation, and how these interact with the different conversational roles (explainer vs. explainee).

2 Methodology

2.1 Dialogue setup

The analyzed data stems from a large corpus of dyadic interactions in German (Türk et al., 2023). The corpus consists of 87 explanatory dialogues, in which an explainer (ER) had the task to explain the board game 'Deep Sea Adventure' (Sasaki and Sasaki, 2014) to several (2–3) randomly chosen explainees (EE) consecutively. That is, each explainer is involved in 2–3 conversations each, thereby possibly adapting their explanation strategy, but also possibly adapting to different conversational partners. Prior to the study, the explainers had (a minimum of) two days to familiarize themselves with the board game rules. Explainers were entirely free in how they explained the board game. However, each explanation dialogue had to contain three phases: initially, there was a phase in which the physical

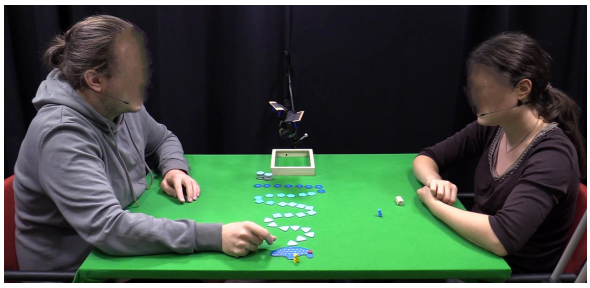


Figure 1: Explanation dialogue setup with the explainer (ER, left) and the explainee (EE, right). The figure shows a dialogue phase with the board game materials present.

board game was not present (gameAbsent). Next, explainers chose a moment at which the actual board game was introduced and the explanation was continued (gamePresent). Last, the explainers were asked to play the board game together with the explainees (gamePlay). This phase may or may not contain aspects of explanation. Explainers were free to choose when to end one explanation phase and begin the next. All interactions were video- and audio-recorded (see Figure 1) using individual head set microphones and multiple camera perspectives.

2.2 Annotations

The explanations were first transcribed with the help of the BAS Web Services (Kisler et al., 2017) or the automatic speech recognition software Whisper (Radford et al., 2022), and then corrected manually using Praat (Boersma and Weenink, 2022). In this annotation step, labels for disfluencies, backchannels, laughter, and audible breathing noises were added. Currently, the corpus is being annotated further for discourse functions, multimodal behaviors, acoustic-phonetic as well as symbolic prosody – but these were not analyzed further in the present study.

2.3 Participants

For the current analysis, we used dialogues from 21 explainers with 2–3 explainees each (75 explanatory dialogues in total). The mean duration of these explanations was M = 26min 47s (SD = 5min 55s). All explainers were German native speaking adults (age: $M = 23.33, SD = 2.58$; 6 male, 14 female, 1 diverse). Not all explainees chose to provide their socio-demographic information. However, they were all recruited based on their report of being a native German speaker. All participants signed a consent form, and the study had been approved by the university Ethics Board.

2.4 Characterizing floor transitions

Using the methodology described in Gilmartin and Włodarczak (2023), the transitions of longer stretches of speech in our data set were characterized as being either examples of *within-speaker transitions* (WST) or *between-speaker transitions* (BST). This yielded a total amount of n = 19 458 floor transitions. For each dialogue, these were further split into the three explanation phases by partitioning the data into three equal parts of overall transitions, the first of which is assumed to roughly correspond to the dialogue phase gameAbsent, the second to the dialogue phase gamePresent, and the third to the dialogue phase gamePlay.

Additionally, each such transition was further characterized with respect to its structural detail: It is determined whether each transition contains stretches of solo speech, silences, or overlaps. Audible breaths, clicks, or laughter occurring on their own were excluded from the speech category and were not taken into account further. Based on the total number of events occurring in between two longer stretches of speech, each floor transition is then given a complexity score. That is, a floor transition that contains a single event in between longer stretches of speech, e.g., a silence, has the transition complexity of 1. With each further event, the complexity score increases.

Transitions types can also be represented with a shorthand notation using uppercase Latin letters to denote individual speakers (*A* and *B* for our dyadic explanations), combinations of letters to denote overlaps, and the letter *X* to denote global silence. Thus, for instance, *A_AB_B* is a between-speaker interval from speaker A to speaker B involving a single overlap, and *A_X_B_AB_A* is a within-speaker transition involving a silent interval, a shorter interval of solo speech by B and an overlap between A and B.

2.5 Analyses

In line with previous research (see Section 1), we expected most floor transitions to show an odd number of intervening intervals, and counted the most frequent patterns for transitions with one and three intervening events. As these counts revealed identical preferred transition patterns across dialogue phases (separately for BSTs and WSTs), we performed χ^2-tests to see whether the patterns distributed differently across the three different dialogue phases.

In order to test whether the transition complexity (measured as the number of individual events occurring between two longer stretches of speech) differed between dialogue phases, transition types, and the dialogue role of the speaker keeping or taking the turn, we calculated non-parametric Kruskal-Wallis tests, followed by post-hoc pairwise comparisons (Dunn tests, Bonferroni corrected). A non-parametric method was chosen, as regression models yielded non-normally distributed residuals.

In order to determine whether the odds for certain transition events (silences, solo speech, overlapping speech) differed between different dialogue phases, we calculated mixed logistic regression models with *silence*, *overlap* and *solo speech* as dependent variables, and dialogue phases (gameAbsent, gamePresent, gamePlay), direction of transition (EE, ER), as well as transition type (BST, WST) as fixed factors, and explainer as random intercepts. We also checked for significant interactions of the fixed factors, and carried out post-hoc pairwise comparisons where these occurred.

All statistical analyses were carried out using R version 4.1.3 (R Core Team, 2022), and the packages tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), and rstatix (Kassambara, 2023). Post-hoc comparisons of factors involved in model interactions were performed using the package emmeans (Lenth, 2022).[1]

3 Results

3.1 Floor transitions across different transition types

In line with earlier research, the vast majority of floor transitions show an uneven number of intervening intervals (see Figure 2). Overall, there are fewer BSTs (n = 5284) than WSTs (n = 14 174). Simple transitions are more likely to be WST, while more complex transitions (>3 intervening intervals) are more likely to be BST (see Figure 2, right). That is, interlocutors invest more floor management work to yield or grab turns, and less to keep them. This tendency is statistically significant ($H(1, 19\,458) = 365.78$, $p < 0.001$), and post-hoc pairwise comparisons showed that this trend is stable for dialogues from 19 out of 21 explainers.

[1]The R-scripts and derived data sets (not the original recordings) can be obtained from the authors upon request.

Table 1: Frequencies of occurrence (raw counts) of floor transitions patterns for 1-interval transitions in BST and WST across the three different explanation phases.

	gameAbs		gamePres		gamePlay	
Pattern	**BST**	**WST**	**BST**	**WST**	**BST**	**WST**
A_X_B	247	2971	575	1899	720	1553
A_AB_B	56	200	171	198	157	118
total	303	3171	746	2097	877	1671

3.2 Floor transition complexities across different explanation phases

The dyadic explanations contain a higher proportion of simple (one-interval) floor transitions than what has been reported for the free dyadic conversations in the Switchboard corpus (Godfrey et al., 1992), especially for the first phase of the game explanations (see Figure 2, left). In later stages, the proportion of simple floor transitions drops strongly, more in line with less constrained conversational data. These differences in complexity across explanation dialogue are statistically significant ($H(2, 19\,458) = 482.53$, $p < 0.001$), and post-hoc pairwise comparisons showed that this trend is stable for dialogues from 19 out of 21 explainers.

3.3 Floor transitions patterns across different explanation phases

The frequencies of occurrence for different floor transition patterns are presented in Tables 1 and 2, separate for BSTs and WSTs. For transitions with one intervening interval (Table 1), the preferred floor transition patterns remain similar across the different dialogue phases for BSTs ($\chi^2(2, 1876) = 0.184$, n.s.), but change for WSTs, with a slightly higher proportion of overlapping transitions in the later dialogue phases, gamePresent and gamePlay ($\chi^2(2, 6953) = 24.62$, $p < 0.001$). For transitions with three intervening intervals (Table 2), the relative distribution of preferred floor transition patterns change significantly, both within BSTs ($\chi^2(2, 1024) = 13.83$, $p < 0.05$) and WSTs ($\chi^2(2, 3325) = 28.36$, $p < 0.05$), but it is difficult to identify a clear-cut pattern in these changes.

Generally, it can be observed that the occurrences of WSTs decrease in course of the dialogue, while the numbers of BSTs increase, indicating a higher level of floor transition related 'work' in the later stages of the explanation.

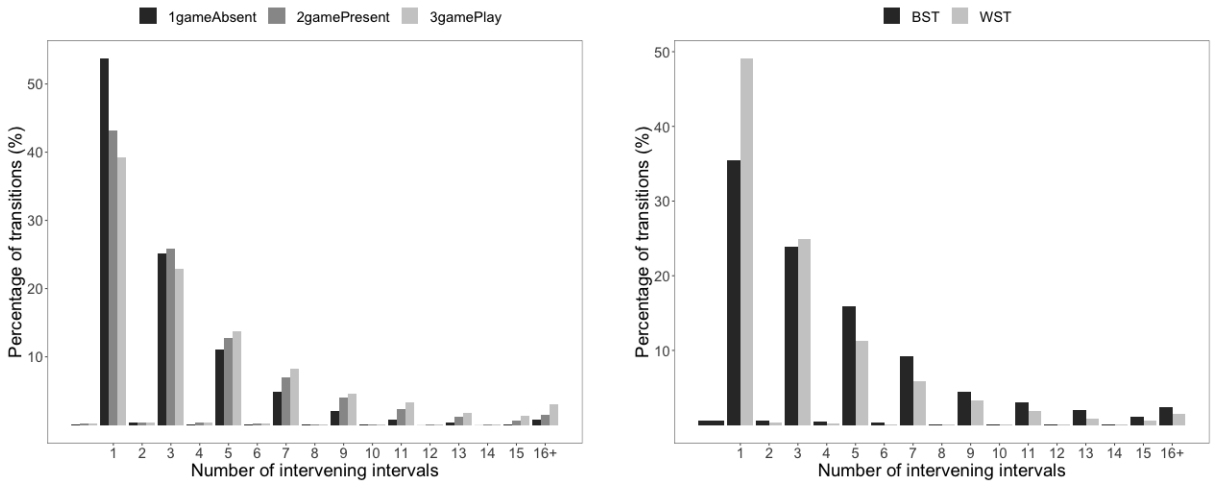


Figure 2: Frequencies of occurrence (%) of transition complexities across the three different dialogue phases (left) and transition types BST and WST (right).

Table 2: Raw counts of the most frequent floor transitions patterns for 3-interval transitions in BST and WST across the three different explanation phases.

	Pattern	gameAbs	gamePres	gamePlay
BST	A_X_B_X_B	43	181	254
	A_X_A_X_B	39	103	126
	A_AB_B_X_B	25	75	68
	A_AB_A_X_B	17	45	48
	total	124	404	496
WST	A_X_A_X_A	733	537	390
	A_X_B_X_A	358	285	234
	A_X_B_AB_A	120	108	101
	A_AB_A_X_A	77	68	54
	A_AB_B_X_A	76	101	83
	total	1367	1099	862

3.4 Floor transitions across different dialogue roles

BST transitions are equally often concerned with transfer of the floor to EEs ($n = 2643$) as to ERs ($n = 2641$), but BSTs to ERs are more complex ($M = 4.62, SD = 4.88$) than those to EEs ($M = 4.1, SD = 4.44$). That is, less complex BSTs tend to correspond to floor transitions to EEs, and more complex BSTs tend to correspond to floor transitions to ERs (see Figure 3, left). These complexity differences are statistically significant ($H(1, 5284) = 17.0, p < 0.001$). In WSTs, this pattern is almost reversed (see Figure 3, right): a lot more WST floor transitions are targeted to ERs ($n = 12\,270$) than to EEs ($n = 1904$), and transitions to ERs have fewer intervening intervals ($M = 3.2, SD = 4.16$) than those to EEs ($M = 4.12, SD = 4.74$). These differences are statistically significant ($H(1, 14\,174) = 128.37, p < 0.001$). Taken together, this indicates that more floor management work is necessary when the floor transitions are not aligned with the assigned dialogue roles, where the explainer's task is to keep the floor (and continue with the explanation), and the explainee's main task is to react and signal understanding, non-understanding, or ask for clarification.

3.4.1 Distributions of overlaps, solo speech, and silences across different game phases

The analysis of preferred floor transition patterns already indicated shifting patterns across different dialogue phases (see Section 2.4). In the following, these tendencies are examined in more detail using mixed logistic regression models.

The regression model for overlaps shows that both game phases and transition types influence the likelihood of overlapping speech (see Figure 4, left). In particular, gamePlay makes overlapping speech less likely ($est = -0.34, se = 0.08, z = -4.1, p < 0.001$) and WST transitions make overlapping speech less likely ($est = -0.14, se = 0.08, z = -17.0, p < 0.001$). Also, there is a significant interaction between dialogue phase and floor transition type, leading to opposite effects for BSTs and WSTs in course of the dialogue: BSTs are losing their stronger likelihood tendency to show overlap in course of the game, showing least overlapping speech during gamePlay, while WSTs are increasing their likelihood to show overlap in course of the game, and are least likely to show overlap during gameAbsent (see Figure 4, left). For BSTs, a pairwise post-hoc comparison showed significant differ-

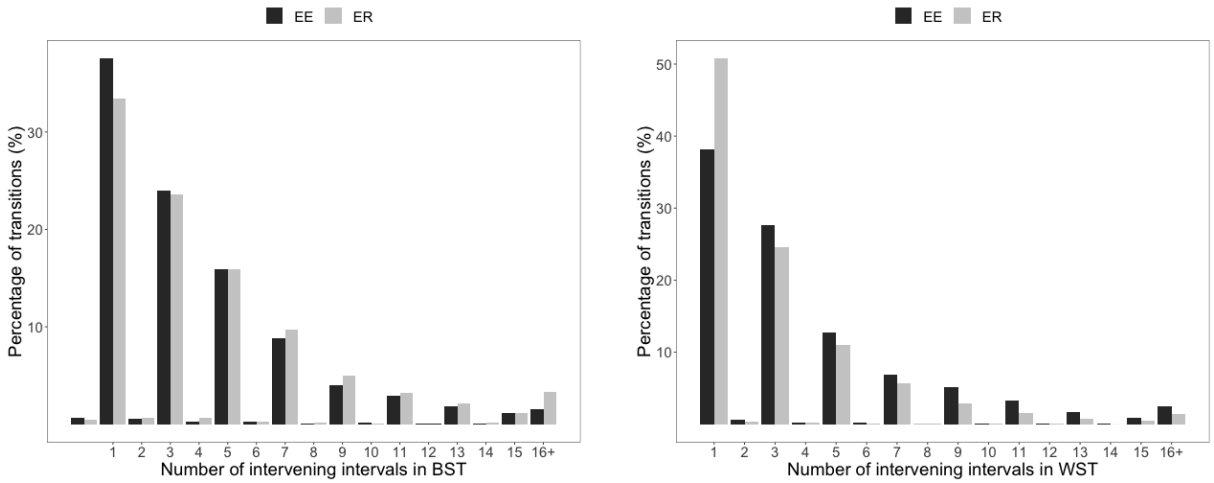


Figure 3: Frequencies of occurrence (%) of transition complexities to ER and EE in BST (left) and WST (right).

ences between gamePlay and the earlier gameAbsent ($est = 0.33, se = 0.08, z = 4.03, p < 0.001$) and gamePresent ($est = 0.21, se = 0.06, z = 3.37, p < 0.01$). For WSTs, a pairwise post-hoc comparison showed significant differences between gameAbsent and the later gamePresent ($est = 0.58, se = 0.05, z = 12.1, p < 0.001$) and gamePlay ($est = 0.73, se = 0.05, z = 2.96, p < 0.001$).

Furthermore (see Figure 4, right), we found that overall, solo speech is less likely to occur in WSTs than in BSTs ($est = -0.6, se = 0.07, z = -7.67, p < 0.001$), and later phases of the dialogue increased the likelihood for solo speech to occur in floor transitions (gamePresent: ($est = 0.19, se = 0.09, z = 2.14, p < 0.05$); gamePlay: ($est = 0.20, se = 0.08, z = 2.38, p < 0.05$). A post-hoc test revealed that the tendency of solo speech to increase in the later stages of the dialogue is largely due to WSTs, which show a significant increase in solo speech between gameAbsent and gamePresent ($est = 0.37, se = 0.04, z = 8.89, p < 0.001$) as well as between gamePresent and gamePlay ($est = 0.19, se = 0.04, z = 4.28, p < 0.001$). For BSTs, this tendency can only be found when contrasting the early gameAbsent and the late gamePlay phases ($est = 2.0, se = 0.08, z = 2.34, p < 0.05$).

As for silences (see Figure 5), the model reveals they have a high likelihood to occur in all floor transitions – in line with the results displayed in Tables 1 and 2. Also, silences are more likely to occur in WST transitions ($est = 2.0, se = 0.08, z = 2.34, p < 0.05$). Due to interactions between the transition types and dialogue phases, we performed pairwise post-hoc comparisons, which revealed that silences are distributed differently for BSTs and WSTs across the dialogue: For WSTs, silences are most likely in the initial gameAbsent and the final gamePlay phase, and differing significantly from gamePresent (gameAbsent-gamePresent: $est = 0.25, se = 0.09, z = 2.59, p < 0.05$; gamePlay-gamePresent: $est = 0.4, se = 0.11, z = 3.73, p < 0.001$). For BSTs, silences are least likely in gameAbsent, and do not differ in their probability to occur in the later phases in the dialogue (gameAbsent–gamePresent: $est = 0.38, se = 0.14, z = 2.69, p < 0.05$; gameAbsent–gamePlay: $est = 0.38, se = 0.14, z = 2.77, p < 0.05$).

4 Discussion

Overall, our results show that explanatory dialogues differ from free dyadic phone conversations in various ways. In particular, they have a higher likelihood to have less complex floor transitions, especially in the first phase of the ongoing explanations, where the physical board game was not yet present and the explanations were made in an 'abstract' fashion. This indicates that floor transition patterns can differentiate between different types of dyadic interactions (phone conversations on a given topic vs. explanations). However, at first glance, this result is not in line with our expectation about explanations being characterized by a high degree of co-construction (Rohlfing et al., 2021). Rather, the floor transitions appear to reveal a strong degree of monologic chunking rather than dialogic chatting. This impression is strengthened by the general prevalence of WSTs (rather than BSTs), and the fact that WSTs rarely coincide with overlaps, but almost always with silences. Also, WSTs to explainers tend

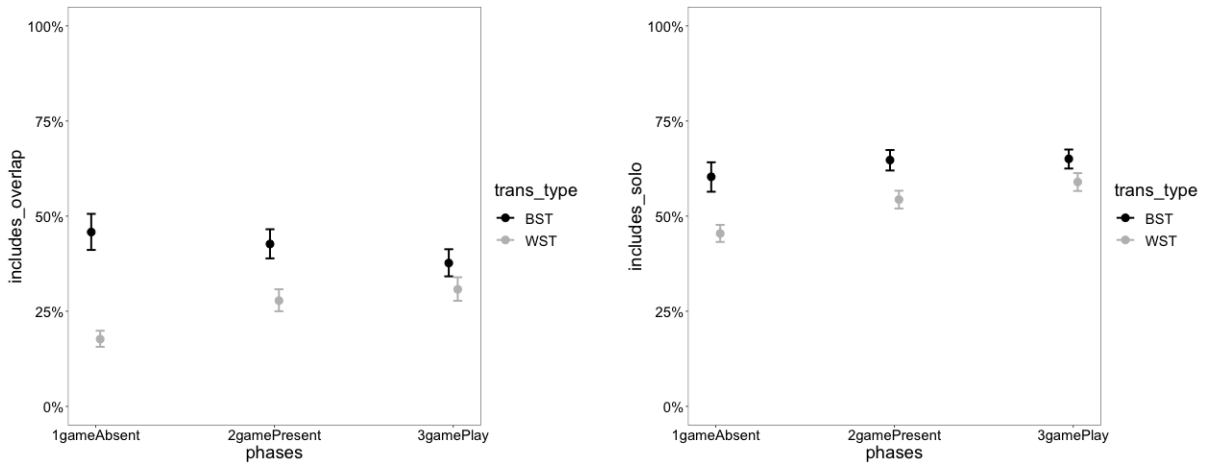


Figure 4: Predicted model probabilities (with 95% CIs) for occurrence of overlapping speech (left) and solo speech (right) across game phases.

to be least complex. Taken together, this gives the impression of an explainer mainly speaking and holding the floor (and not being challenged), and the explainee mostly being in a listening role.

However, we also clearly see that both the floor transition complexities as well as the proportion of BSTs increase in the later phase of the explanation, where the physical board game is introduced as a shared space that interlocutors can refer to, both verbally and multimodally (e.g., by deictic gestures). This is in line with findings by Fisher et al. (2022), who showed that explanations can take more monologic or more dialogic forms, but in our data, this change coincides with a change in situation (visible board game), which probably caused a higher degree of interaction by our interlocutors. A higher degree of co-construction during the gamePresent phase is also indicated by the drop in silences for WSTs, together with a higher proportion of overlaps and complex floor transitions. Currently, we cannot say whether this impact on co-construction can be generalized to other types of explanatory interactions (e.g., doctor-patient, teacher-student), both of which may come with and without a shared physical frame of reference, but our results ask for further analyses across different contextual settings.

It comes as no surprise that the last gamePlay phase in our explanations turned out to be most 'chatty', with a more equal distribution of WSTs and BSTs, and and almost equal proportion of overlap and solo speech in WSTs and BSTs. In earlier, more explanatory phases, BSTs are characterized by considerably more overlap, indicating that more turn 'grabbing' effort is necessary in the explanatory phases than during gamePlay.

In our view, the most interesting finding concerns the interaction between the interlocutors' role (explainee/EE vs. explainer/ER) and floor work necessary in BSTs and WSTs: WSTs to EE were more complex than those to ER, while BSTs to ER were more complex than those to EE. This shows that speakers had to invest more conversational effort whenever they were not conforming to their assigned roles of a predominantly 'speaking explainer' (who tends to have the turn, and may yield it when feedback is needed), or a predominantly 'listening explainee' (who might react with feedback to an explanation, but does not typically keep the turn). We therefore see that dialogue roles influence our floor transition behaviors, and in more equal interactions such as the gamePlay phase, these role-specific behaviors are adapted.

Obviously, our study has several limitations. As conversational data differs across many dimensions, it is difficult to compare results across different settings. Here, we not only compared dyadic free conversations (in American English) to dyadic explanations (in German), but we also compared phone conversations to conversations where interlocutors could see each other, and interact both verbally and non-verbally. We know from prior work that visibility alone has an effect on floor management in instructional dialogues, as visibility decreases overlaps and turn durations, but increases verbal backchanneling (Boyle et al., 1994). It is yet unclear, whether our result for a predominance of monologic interaction in the explanations were not exaggerated, as it currently ignores a large amount

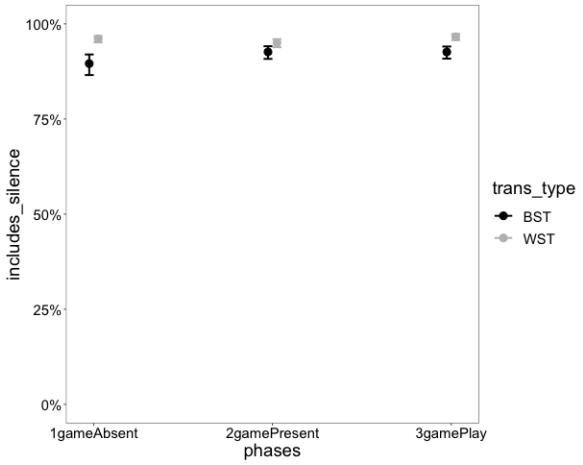


Figure 5: Predicted model probabilities (with 95% CIs) for occurrence of silences across game phases.

of non-verbal feedback behaviors as well as non-verbal cues related to floor management such as gaze, blinking, or head gestures (Malisz et al., 2016; Hömke et al., 2017; Kendrick et al., 2023). It is indeed possible, that interlocutors reduce their usage of gestural floor management cues once the board game is present during the explanation, as they need their hands to carry out the actual movements of the game, need to look at the board game, or use their hands to perform deictic gestures. Because of this, they may switch to a higher proportion of verbalized floor management cues, which we interpreted as more co-constructive interaction. In future work, we will therefore investigate whether non-verbal, gestural floor management follows a similar pattern throughout the various phases of the explanations, or whether verbal floor management compensates if the non-verbal cues cannot be expressed.

Another possibly confounding factor in our data relates to the way that the explanatory dialogue would have evolved without asking our participants to go through various explanatory phases. It is possible, that some of the findings presented here are the result of interlocutors 'warming up' to one another, and becoming more chatty in course of an interaction after a somewhat awkward initial phase. While this cannot be ruled out, we are still confident that this does not explain all our findings, as we see very stable tendencies across many speakers, who also displayed a wide variation in their individual interactive behaviors, or readiness to chat. Also, for silences, we found similarities for the initial and late stages of the conversation, which are difficult to explain if the explanatory phases, as implemented by our design, did not play a role at all.

Lastly, it has to be critically mentioned that our coarse split into three different phases does not properly reflect the three explanation phases. However, as our analysis yielded interesting differences between those three phases, we believe that this approach was successful as a first approximation.

5 Conclusions

Overall, our findings show that explanatory interactions follow turn-taking dynamics that differ from other types of conversational interactions, and shed light on the special turn-taking dynamics in different phases of explanatory interactions. Also, our analysis corroborates the usefulness of floor transitions as a measure for characterizing conversational dynamics and involvement of conversational partners.

Acknowledgments

This research was funded by the German Research Foundation (DFG): TRR 318/1 2021–438445824. We are grateful to two anonymous reviewers who made very helpful comments, and pointed out highly relevant literature we would have missed.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.

Paul Boersma and David Weenink. 2022. Praat: Doing phonetics by computer.

Elizabeth A. Boyle, Anne H. Anderson, and Alison Newlands. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, 37(1):1–20.

Michelene T. H. Chi. 1996. Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology*, 10:33–49.

Sarah Collins. 2005. Explanations in consultations: the combined effectiveness of doctors' and nurses' communication with patients. *Medical Education*, 39:785–796.

Mark Dingemanse and Andreas 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*, pages 5614–5633, Dublin, Ireland.

Suzanne Eggins and Diana Slade. 1997. *Analysing Casual Conversation*. Cassell, London, UK.

Josephine B. Fisher, Vivien Lohmer, Friederike Kern, Winfried Barthlen, Sebastian Gaus, and Katharina J. Rohlfing. 2022. Exploring monological and dialogical phases in naturally occurring explanations. *KI – Künstliche Intelligenz*, 26:317–326.

Catherine Garvey and Ginger Berninger. 1981. Timing and turn taking in children's conversations 1. *Discourse Processes*, 4:27–57.

Emer Gilmartin, Kätlin Aare, Maria O'Reilly, and Marcin Wlodarczak. 2020. Between and within speaker transitions in multiparty conversation. In *Proceedings of Speech Prosody 2020*, pages 799–803, Tokyo, Japan.

Emer Gilmartin, Carl Vogel, and Nick Campbell. 2018. Chats and chunks: Annotation and analysis of multiparty long casual conversations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Emer Gilmartin and Marcin Włodarczak. 2023. Getting from A to B: Complexities of turn change and retention in conversation. In *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, pages 3457–3461, Prague, Czech Republic.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, San Francisco, CA, USA.

Augustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25:601–634.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38:555–568.

Paul Hömke, Judith Holler, and Stephen C. Levinson. 2017. Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, 50:54–70.

Alboukadel Kassambara. 2023. *rstatix: Pipe-friendly framework for basic statistical tests*. R package version 0.7.2.

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*, 378:20210473.

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

Janine Larrue and Alain Trognon. 1993. Organization of turn-taking and mechanisms for turn-taking repairs in a chaired meeting. *Journal of Pragmatics*, 19:177–196.

Russell V. Lenth. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.3.

Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, Stefan Kopp, and Petra Wagner. 2016. The ALICO corpus: Analysing the active listener. *Language Resources and Evaluation*, 50:411–442.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arxiv:2212.04356.

Katharina Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike Buhl, Hendrik Buschmeier, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13:717–728.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Jun Sasaki and Goro Sasaki. 2014. *Deep Sea Adventure (Tabletop Game)*. Oink Games, Tokyo, Japan.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Sonja Stange, Teena Hassan, Florian Schröder, Jacqueline Konkol, and Stefan Kopp. 2022. Self-explaining social robots: An explainable behavior generation architecture for human-robot interaction. *Frontiers in Artificial Intelligence*, 5:866920.

Tanya Stivers, Nick J. Enfield, Penelope Brown, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106:10587–10592.

Olcay Türk, Petra Wagner, Hendrik Buschmeier, Angela Grimminger, Yu Wang, and Stefan Lazarov. 2023. MUNDEX: A multimodal corpus for the study of the understanding of explanations. In *Proceedings of the 1st International Multimodal Communication Symposium*, pages 63–64, Barcelona, Spain.

Hadley Wickham, Mara Averick, Jennifer Bryan, et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Marcin Włodarczak and Emer Gilmartin. 2021. Speaker transition patterns in three-party conversation: Evidence from English, Estonian and Swedish. In *Proceedings of Interspeech 2021*, pages 801–805.

PairwiseTurnGPT: a multi-stream turn prediction model for spoken dialogue

Sean Leishman and **Peter Bell** and **Sarenne Wallbridge**
University of Edinburgh, UK
{s2051283, peter.bell, s.wallbridge}@ed.ac.uk

Abstract

Spoken conversation is characterised by rapid turn transitions and frequent speaker overlaps. However, existing models of turn-taking treat dialogue as a series of incremental turns. We propose PairwiseTurnGPT, a language model that captures the temporal dynamics of lexical content by modelling dialogue as two aligned speaker streams. PairwiseTurnGPT provides a much more nuanced understanding of how lexical content contributes to predicting turn-taking behaviour in speech. By training the model with data configurations containing different turn-taking behaviours, we demonstrate the relative contributions of partial, complete, and backchannel overlaps for accurately predicting the variety of turn ends that occur in spoken dialogue. We also show that PairwiseTurnGPT improves on serialised models of dialogue for predicting turn ends and the more difficult task of predicting when a turn will start.

1 Introduction

Turn-taking—deciding who speaks at what point during an interaction—is a crucial component of successful spoken communication between humans. However, as the example in Figure 1 depicts, it is an intricate task. The organisation between conversational partners has fascinated psycholinguists for decades, particularly how interlocutors achieve such short transitions between turns: gaps between turns typically range from -100 to 500ms (Levinson and Torreira, 2015) (the negative end of the range indicating an overlap between turns). To explain the speed at which turn-taking occurs, Sacks et al. (1974) pioneered the *predictive* model of turn-taking, theorising that people engage in some form of "projection" to determine an appropriate point to begin their own turn while their partner is still speaking. Under this model, speakers construct their speech such that potential turn transition points are foreshadowed to their listener. This raises the question – what features of speech do listeners rely on to predict potential turn ends?

In spoken conversation, turn-end cues stem from both the lexical content and its prosodic realisation. (Ford and Thompson, 1996; Bögels and Torreira, 2015; De Ruiter et al., 2006; Ward, 2019). However, their relative contributions are unknown. While models that leverage lexical and prosodic cues in isolation and combination can learn to predict some turn-taking behaviour, the simplicity of linguistic representations used in such models may obscure the true contributions of lexical content to turn-end prediction (Ward et al., 2018; Roddy et al., 2018). More recently, Ekstedt and Skantze (2020) proposed *TurnGPT*, a model for turn-end prediction that harnesses the power of pre-trained GPT-2 (Radford et al., 2019). TurnGPT achieves a high accuracy in predicting turn-endings, demonstrating the value of lexical information for this task.

TurnGPT has, however, been predominately trained and designed for *written* conversation. Like the GPT-2 model it is based on, the model is limited to a single stream of input. Although written dialogue can be neatly collapsed into a single stream of interleaved turns, compressing spoken conversation in this way disregards much of the nuance of realistic turn-taking behaviour. Whilst TurnGPT has been applied to spoken dialogue, it does so by serialising overlapping utterances into a single stream, sometimes requiring significant reordering or removal of lexical content; Figures 2a and 2b give examples of the TurnGPT formatting.

This paper seeks to better understand the contribution of lexical content to turn-taking in spoken communication by more accurately modelling its temporal dynamics. For this, we propose modelling transcripts as a dual-stream system that resembles their original production much more closely. We present *PairwiseTurnGPT*, a novel architecture capable of modelling these synchronous input streams. Doing so provides insights into how

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

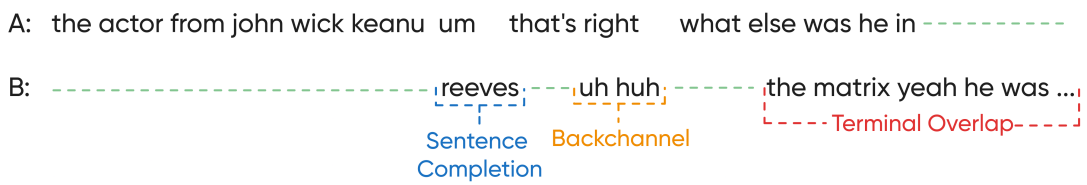


Figure 1: A (synthetic) example of overlaps in dialogue.

well lexical content can predict different types of turn-taking behaviour, including how speakers both end and begin their turns. By comparing training data configurations containing varying degrees of turn-taking complexity (i.e., partial, complete, and backchannel overlaps), we analyse their relative contributions to turn-taking prediction.

2 Background

2.1 Theories of turn-taking

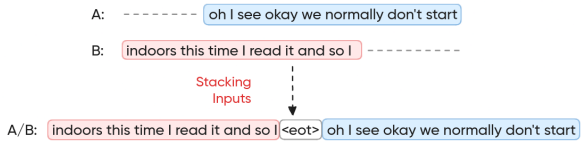
Human turn-taking behaviour has generally been characterised by two processes within the literature: the *reactionary* and the *predictive* approach. The former assumes that participants understand end-of-turn signals and react to them accordingly while the predictive approach posits that listeners predict the end of the turn in advance to time their response. The reactionary approach was pioneered by (Duncan, 1972, 1973, 1974; Duncan and Fiske, 2015) who argued for a precise set of context-free turn-yielding 'signals' which include both vocal and gestural signals (Yngve, 1970).

Others have argued against the general model of a reactionary approach because turn-transitions occur too quickly and turn-yielding signals occur too late within a speaker's utterance for the listener to simply react to an end-of-turn signal (Levinson and Torreira, 2015; Riest et al., 2015). Under the predictive account of turn-taking, the speed of turn transitions is possible because speakers predict appropriate points at which to start their turn (Sacks et al., 1974). This model views turns as combinations of *Turn Construction Units (TCUs)*. TCUs are separated by *Transition Relevance Places (TRPs)* that mark where a turn-transition (turn-shift) can but does not have to occur.

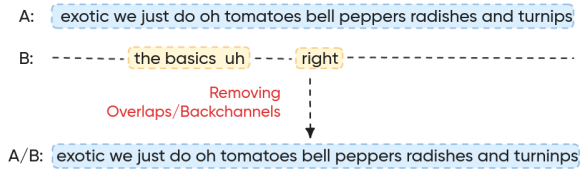
2.2 Behavioural evidence

Early research into turn-yielding signals identified prosodic, syntactic and gestural features that coincide with turn-completions (Duncan, 1972), however defining their contributions has proved complicated. For example, gestural features (Duncan, 1972) and gaze (Kendon, 1967) have shown to be useful cues for turn-taking, but they are action-dependent and more context-sensitive than other features (Clayman, 2012). Numerous works have demonstrated the importance of lexical information for this task. De Ruiter et al. (2006) found that end-of-turn prediction was unaffected by the removal of intonational contours but it was affected by the removal of lexicosyntactic information. Similarly, Magyari and De Ruiter (2012) found that when participants predicted the remaining part of a sentence, this prediction was more accurate if their end-of-turn prediction was also accurate. This suggests that listeners use predicted utterances to determine turn-completion. Pickering and Garrod (2013) also found that listeners imitate the speaker to determine their intention, which they use alongside the speaker's speaking rate to correctly time their own prepared utterance. Findings on importance of lexicosyntactic information align well with the predictive account of Sacks et al. (1974).

Although Ford and Thompson (1996) show that most TRPs occur at syntactic completion points, they theorised that multiple factors are used to determine the completion of TCUs. This theory was tested by Bögels and Torreira (2015) who also sought to refute the claim that intonation had no



(a) Overlapping portions of turns between speaker A and B are ignored by stacking turns in the combined stream. <eot> tokens mark the end of a speaker turn.



(b) Backchannels and completely overlapped turns from speaker B are removed in the combined stream.

Figure 2: The difficulties of serialising spoken dialogue transcripts into a single combined stream.

effect on turn-taking prediction by De Ruiter et al. (2006). This was done by performing the same experiment but with instances of questions with equal syntactic completion points but different turn-shift locations They found that in cases of syntactical ambiguity, lexicosyntactic information is not sufficient for turn-end projection and as such they claim intonation plays a role in disambiguation.

2.3 Computational models for End-of-Turn Detection and Prediction

Models trained to predict turn-taking behaviour are another method for investigating the relative contributions of lexical and acoustic cues. Skantze (2017) show that training with POS tags improves on a purely prosodic model, which supports the importance of syntactic completeness as a cue for turn-end prediction. However, Ward et al. (2018) outperforms Skantze (2017) using only prosodic features. Maier et al. (2017) and Roddy et al. (2018) both employed LSTM RNN models to investigate prosodic and linguistic features in conjunction; Roddy et al. (2018) found that acoustic features are more beneficial and Maier et al. (2017) found that linguistic features performed worse than in their baseline condition. However, linguistic features used in these studies have been simplistic and are unlikely to capture *pragmatic* completeness, a feature deemed crucial by Ford and Thompson (1996).

Ekstedt and Skantze (2020) proposed TurnGPT to harness the strong language modelling of GPT-2. TurnGPT finetunes GPT-2 with a modified objective for dialogue by adding speakers tokens and turn-shift tokens in the model input. The use of a pre-trained language model allows for greater pragmatic and semantic feature representation: TurnGPT is shown to rely not only on syntax but also on the overall pragmatic context of an utterance for turn-end prediction. Jiang et al. (2023) extended the model to condition its predictions on a generated response and found further improvements in end-of-turn prediction performance.

However, these models are not designed specifically for spoken dialogue with much more complex turn-taking behaviour than written dialogue. Transcripts of spontaneous spoken conversations only make up 4% of the training set for TurnGPT. More importantly, the dialogue transcripts are collapsed into a single stream of input for GPT2. To do this, dialogue transcripts are serialised based on turn units; turn units which are fully overlapped or are classified as backchannels, are removed from the transcription, whilst consecutive turn units from the same speaker are concatenated to form each speaker's full turns. The process is depicted in Figure 2a. As well as removing important information about a conversation, the process might also be viewed as fundamentally altering the task of turn-taking prediction in spoken dialogue.

Recently, TurnGPT representations have been used by Wang et al. (2024) in conjunction with an acoustic model to predict backchannel events in spontaneous dialogue transcripts. Like TurnGPT, this model is trained using data serialised into sequential speaker turns; however, backchannels are reintroduced using word-level time stamps. Whilst backchannels are known to be strongly linked to their prosodic realisation (Gravano and Hirschberg, 2011), Wang et al. (2024) found good performance for their prediction using only a language model.

The studies described above provide evidence that lexical information contributes to turn-end prediction, but do not provide a complete picture of its contribution in spoken conversation. We apply powerful modern language models to more realistic representations of turn-taking.

2.4 Characterising overlaps

Overlaps are a frequent and important component of spoken dialogue. Overlaps can occur when speakers mispredict the end of a TCU; however, they can also serve interactional purposes that are lost when serialising spoken dialogue transcripts into a single stream.

Overlaps can be categorised as competitive or cooperative, where speakers are either vying for the floor or aiding one another in the construction of a turn Schegloff (2000). As depicted in the example in Figure 1, cooperative overlap can consist of: *terminal overlap*, where the listener predicts the end of a turn and begins speaking prior to the other speaker finishing their turn; *turn completion* where the listener helps the speaker complete their turn but doesn't intend to take the floor; and *backchannels* such as "uh huh" and "hmm". These typically occur where the speaker requests affirmation from the listener and have their own set of cues, as defined by Clark (1996).

3 PairwiseTurnGPT: A New Approach

Our proposed approach – which we call PairwiseTurnGPT – models each speaker in a conversation

as an independent stream of tokens. We pair tokens across the two streams based on word timing information, enabling effective modelling of the complex interactions between speakers. This avoids the deficiencies inherent in the standard serialised approach (Ekstedt and Skantze, 2020), where turns are interleaved in a manner that erases turn-taking phenomena potentially conveying important information. Though Wang et al. (2024) only incorporate a limited aspect of temporal dynamics, their results demonstrate the value of such information. By aligning streams at the word level, we encode this structure much more explicitly.

3.1 Model Architecture

PairwiseTurnGPT is composed of a GPT-2 stream for each speaker in the dialogue. A diagram is included in Figure 4). Similarly to the spoken dialogue model proposed by Nguyen et al. (2023), GPT-2 weights are shared between the streams. Through a multi-head cross-attention layer in each transformer block, predictions in each stream are conditioned on the conversational history of both speakers. The training objective is the sum of the cross-entropy loss for each speaker streams[1].

To incorporate spoken turn-taking phenomena in PairwiseTurnGPT, we use word-level timings to align the speaker streams. GPT-2's BPE tokenisation functions at the sub-word level, therefore we obtain token timings by uniformly splitting word timings across tokens (Figure 5 depicts an example). Tokens are then aligned in a pairwise manner. For tokens with no significant overlap (defined as an overlapping duration no greater than 50% of the shorter-duration word), an empty `<emp>` token is used to make up the token pair. An example of this alignment is shown in Figure 3a.

3.1.1 Turn-Level Annotation

Pairwise alignment enables our models to learn taking behaviours that involve fine-grained overlap between conversational partners. From the aligned data we identify categories of such turn ending strategies to better understand which behaviours are captured by PairwiseTurnGPT. **Backchannels** involve one speaker interjecting a short utterance such as "hmm", "uh huh" or "yeah" to provide feedback to the speaker (Ward, 2004). We follow (Ekstedt and Skantze, 2020) and define these based on their lexical content[2] and a pause of at least $1s$ between surrounding turns from the same speaker. **Complete Overlap** occurs where one speaker begins and ends their turn before the other speaker finishes theirs, as depicted in Figure 2b. **Yield Turn-Shift** are when one speaker begins their turn before the other speaker finishes theirs (i.e. a partial overlap). Yield turns are those that contain an overlap of $> 0.1s$, or where the other speaker produces an overlap within 2s of the turn ending. **Normal Turn-Shift** turns involve one speaker finishing their turn and the other speaker beginning theirs after a pause. The difference between normal and yield turns is shown in Figure 3b. Appendix C shows how turn types are distributed in Switchboard.

The full alignment process is demonstrated in Figure 3a which includes turn annotation: the determination of the type for each utterance; turn alignment: ensuring each token is aligned appropriately and turn token addition: adding in the end-of-turn token corresponding to the determined turn type.

4 Experimental Setup

4.1 Model

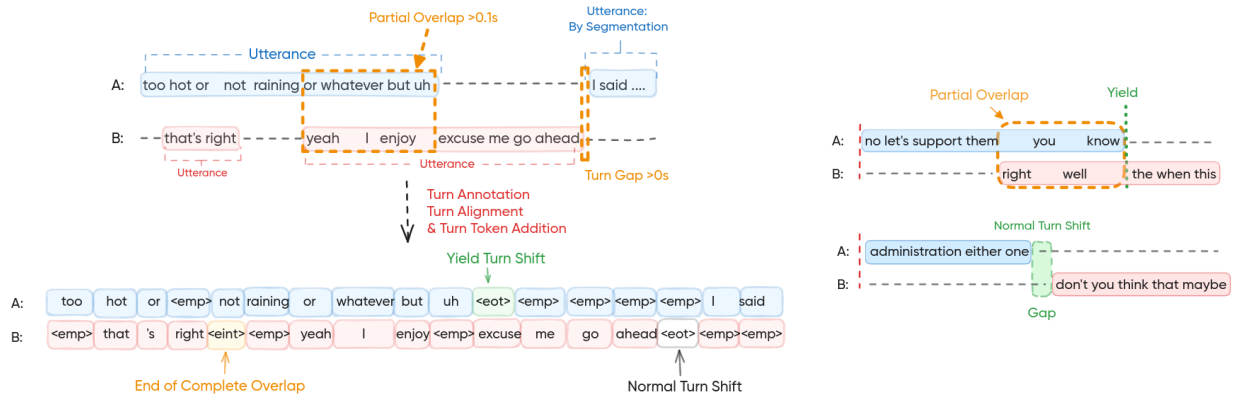
To allow comparison with TurnGPT, we initialise our model with *GPT2-base* which consists of 124M parameters, 12 layers, 12 heads and 768 hidden units. The pre-trained weights were obtained from the OpenAI GPT2 model from the Hugging Face Transformers library (Wolf et al., 2020). The cross-attention weights are initialised using the default Hugging Face method by sampling a unit normal distribution. We fine-tune using the AdamW optimizer with a learning rate of $6.25e^{-5}$ and a weight decay of 0.01. All models are trained for 5 epochs or until the validation loss does not decrease for two consecutive epochs with batches of size 4.

4.2 Data

We train and evaluate PairwiseTurnGPT on the Telephone Speech Switchboard Corpus which consists of 2430 conversations between 542 participants (Godfrey et al., 1992; Deshmukh et al., 1998). Although the dataset is not large, it includes manual transcriptions and manually corrected word timings, making it an ideal base for our study. We remove all annotations of non-verbal vocalisations from the transcripts while partial words, mispronunciations and coinages are replaced with the full

[1]We release our code at https://github.com/Sean-Leishman/PairwiseTurnGPT. This includes code for model training and data preprocessing.

[2]We use the list of candidate backchannel responses defined in Ekstedt and Skantze (2020)



(a) Pairwise data preprocessing: tokens are aligned based on token timing information. B's first utterance is labelled as a complete overlap; the end of A's first utterance is labelled as a "yield" turn shift on account of the partial overlap with B's second utterance, whilst B's second utterance is labelled as a normal turn shift.

(b) Yield & normal turn-shift; labels are based on the amount of overlap with the other speaker's turn.

Figure 3: Data Labelling & Preprocessing

intended word. Switchboard doesn't have a standard evaluation set for this task, so we randomly split the corpus into train, validation and test sets ($[90/5/5]$ proportions, respectively).

4.3 Training data configurations

We train with pairwise data configurations that include varying degrees of turn-taking behaviour to understand their respective contributions to turn-end predictions.

- **Single stream**: As a baseline, we examine the performance of modelling isolated conversation streams. For this, we remove cross-attention and concatenate turns in each stream so no alignment between speakers takes place.
- **Serialised**: We simulate TurnGPT dialogue representations by aligning streams to turns rather than tokens, thus removing all overlap. Content tokens in one stream are always aligned with `<emp>` tokens in the other.
- **Aligned**: Partial overlaps, where a speaker interrupts prior to a turn ending, are included but not complete overlaps or backchannels.
- **Aligned + Overlaps**: Both partial and complete overlaps are included.
- **Aligned + Backchannels**: Partial overlaps and backchannels are included.
- **Aligned + Backchannels & Overlaps**: The fully aligned condition contains all turn-taking phenomena.

4.4 Evaluation Metrics

The end-of-turn prediction task involves mapping next-token prediction to a binary turn end prediction. We discretise the probability of end-of-turn tokens into a binary label using a threshold tuned on the validation set. Turn prediction is evaluated using Balanced Accuracy (bAcc), the mean of the true positive and true negative rates for turn end prediction; it is robust to the unbalanced nature of turn-end tokens and allows more direct comparison to the TurnGPT results. We also evaluate language modelling through token-level perplexity (PPL).

5 Results

5.1 Validating the pairwise architecture

We begin by establishing that the PairwiseTurnGPT architecture performs at a similar level to the original TurnGPT model. We also investigate the effect of the second speaker stream using different training configurations, where no cross-attention layer is introduced. We consider conditions where each stream consists of turns concatenated together (i.e. the single stream condition) to simulate no knowledge of the other speaker; and where each stream includes empty tokens (i.e. serialised without cross-attention), which simulates the temporal aspect of the other speaker but without any lexical content.

Each PairwiseTurnGPT configuration is evaluated using the serialised data configuration in Table 1. TurnGPT and serialised PairwiseTurnGPT achieve nearly identical turn-end prediction performance; however, PairwiseTurnGPT has higher PPL, indicating a weaker internal representation

Model	bAcc ↑	PPL ↓
TurnGPT	0.828	29.3
PairwiseTurnGPT		
Single Stream	0.805	39.3
Serialised w/o CA	0.825	32.9
Serialised	0.828	31.3

Table 1: End-of-turn prediction accuracy and perplexity scores for serialised data across models. PairwiseTurnGPT contains <emp> tokens while TurnGPT does not, so <emp> tokens are not evaluated.

of language. This may be a result of the more demanding training procedure.

We find that a reasonable turn-end accuracy of 0.805 can be achieved using the single-stream configuration. Without knowledge of the other speaker, this model is reliant on the syntactic completeness of the speaker's current utterance and a partial history of the conversation. Removing cross-attention (CA) from the serialised TurnGPT setup demonstrates how model performance is influenced by the other speaker's lexical content. We find that much of the serialised model performance can be achieved without cross-attention (i.e., only using information about when the other speaker is active).

5.2 Training data configurations

We evaluate the effect of training using our data configurations that incrementally approach the original spoken realisation. Rather than evaluating over a serialised configuration, as in Table 1; we evaluate each model on the fully-aligned configuration of the test data; results are shown in Table 2.

Single stream model The turn-end accuracy scores confirm that a speaker's turn ending is at least somewhat predictable from their own conversational history, which contains information regarding the syntactic and pragmatic completeness of the current utterance. As expected, all models trained to condition their predictions on both speaker streams improve over the single stream set up for all turn end types; even a model trained on the serialised data configuration can achieve a 14-point increase in accuracy.

Aligned vs. serialised Table 2 shows that the lossy encoding of the serialised configuration ignores much of the complex turn-taking behaviour in spoken dialogue: training on the aligned configuration produces better overall accuracy than the serialised configuration. The overall improvement comes primarily from the model's ability to predict yielded turn ends; we find a slight decrease in the ability to predict standard turn endings. The partial overlaps in the aligned configurations are a common feature of spoken dialogue; yielded turns constitute 18.11% of turn endings in the dataset (see Table 4). While their inclusion may complicate the prediction of normal turn endings, they are extremely valuable for capturing the true variety of turn ends in spoken conversation. However, in regard to normal turn-ends, performance varies between the serialised and aligned configurations. This suggests that for simpler turn-ends the model could benefit from developing a simpler understanding of turn-taking, without considering additional phenomena.

Effect of backchannels & overlaps Table 2 shows that models trained on aligned configurations, which include backchannels, are the most accurate overall. Only these configurations outperform the serialised data configuration for predicting normal end-of-turns. Their influence may reflect their communicative functions in spoken dialogue: for example, listeners can employ them to inform the speaker of their intention to continue listening (Yngve, 1970). Even from their lexical content alone, our results demonstrate that backchannels are useful cues for turn-taking.

The inclusion of overlaps also improves turn-end prediction over the aligned configuration, however, to a lesser extent. Overlaps may be more difficult to leverage as they make up a smaller proportion of the turn-end tokens in our corpus and are far less constrained than backchannel responses. For example, overlaps can be cooperative or competing acts in dialogue (Schegloff, 2000). If an overlap is competing, it may be less likely that its resolution can be derived from its lexical content alone. However, the inclusion of overlaps in the aligned data achieves the highest yield turn-ending prediction accuracy as the model can better differentiate between complete and partial overlaps.

Although the combination of both features still provides clear improvement over the serialised condition, including all behaviours doesn't necessarily provide an additive benefit for predicting all turn ends. Notably, the accuracy for yield turns in the fully aligned configuration deteriorates compared to that aligned with only overlaps, suggesting

Configuration	bAcc ↑			PPL ↓		
	Normal	Yield	All	Overlap	Non-Overlap	All
Single stream	0.728	0.640	0.710	–	–	–
Serialised	0.868	0.807	0.852	206	**5.63**	7.67
Aligned	0.863	0.927	0.881	48.1	5.77	6.95
+ Backchannel	**0.872**	0.930	0.914	41.3	5.68	6.55
+ Overlap	0.866	**0.936**	0.890	40.2	5.64	6.69
+ Backchannel & Overlap	0.869	0.934	**0.915**	**36.9**	5.67	**6.49**

Table 2: End-of-turn prediction balanced accuracy over turn types. "All" consists of normal, yield, backchannel and complete overlap turn endings. Perplexity is computed on overlapping, non-overlapping, and all tokens of the fully-aligned test set.

that backchannels may blur the distinction between yields and normal turn endings.

Rule-based comparison To probe the necessity of language modelling for this task, we designed a rule-based classifier to predict an end-of-turn whenever two speakers speak simultaneously. This classifier predicts a turn-shift whenever the listener interrupts the current speaker's utterance. Using the setup in Table 2, the classifier achieves a bAcc of 0.890 over yielded turn ends. This is a strong improvement over the serialised model accuracy of 0.807, indicating that an interruption is a significant signal. However, the gap between the fully aligned model with 0.934 accuracy shows that lexical content provides additional predictive power.

Perplexity As expected, the lowest PPL is found for the model trained on the fully aligned data. Although the serialised configuration does not handle overlapping tokens well, it produces the lowest "non-overlap" PPL. This model may be better able to model lexical content as it does not need to learn temporal aspects of overlapping tokens. The introduction of partial overlaps in the aligned configuration allows the model to better represent overlaps; each subsequent data configuration yields further improvement. By representing overlapping portions of the dataset more effectively, the model may learn patterns regarding how overlaps are resolved and lexical features that prompt a listener to produce an overlap.

Similar to the trends in end-of-turn prediction, the inclusion of backchannels in the aligned training data produces a larger overall reduction in PPL than overlaps. However, these configurations produce similar PPL scores across all token sets. The "overlap" subset does not contain backchannels and so while we may expect the backchannel configuration to perform similarly to the aligned configuration, it achieves a PPL that is closer to the aligned with overlaps configuration. This suggests that by learning to represent backchannel turns, the model can extrapolate to overlaps relatively well.

As expected, the fully aligned configuration trained with all types of overlap performs best overall and in the overlap subset of the fully aligned test set. However, this is not the case for non-overlapping tokens where the result is essentially a weighted sum of the PPL resulting from the aligned with backchannels and aligned with overlaps configurations. This is reflected in end-of-turn prediction and suggests a degree of uncertainty when combining two types of features.

5.3 Prediction of turn starts

Thus far, we have examined turn-taking through the lens of turn ends. However, pairwise alignment also allows us to analyse the different strategies people use to initiate a turn. Here, we evaluate how useful lexical content is for determining interjection points for different types of turns.

We predict the beginning of a turn by summing the probability of all non-`<emp>` tokens and producing a binary prediction as was done for the end-of-turn task. Using the fully time aligned test configuration, we evaluate predictions at points where the current token is the `<emp>` token. We consider several turn start strategies: "Normal" is the start of a non-overlapping turn; "Interruption" is the start of an overlap turn in which the interrupted speaker yields the floor; "Overlap" is the start of a completely overlapped utterance; "BC" is the start of a backchannel.

Configuration	bAcc ↑				
	Normal	Interruption	Overlap	BC	All
Serialised	0.702	0.640	0.581	0.592	0.640
Aligned	0.746	0.669	0.592	0.604	0.669
+ Backchannel	0.809	0.763	0.684	0.753	0.763
+ Overlaps	0.774	0.700	0.614	0.647	0.700
+ BC & Overlaps	**0.819**	**0.774**	**0.689**	**0.765**	**0.774**

Table 3: Predicting start points for different turn types.

As expected, Table 3 shows that overlap turn starts are the most difficult type of turn start to predict. The addition of turn-taking phenomena improves prediction across all turn types. Interestingly, the addition of backchannels is far more useful than the addition of overlaps, mirroring our findings from turn-end prediction performance.

Predicting overlap turn starts is worse across all configurations compared to normal turn starts, likely because overlapping turns do not align with a turn end. However, the fact that yielded turn endings can be accurately predicted suggests that lexical content provides an indication of suitable interjection points in conversation. Results are similar for backchannel turn predictions.

Interestingly, the fully time aligned configuration performs best across all turn start types. However, this was not the case for predicting the end of a turn where this configuration was not the most accurate for normal or yield turn ends. We posited that the result over end-of-turns is due to the overall complexity of the training data. However, it seems that by framing the task differently, the model is able to leverage this information.

6 Discussion & Conclusions

By modelling spoken dialogue transcripts as two separate streams of lexical content, our proposed PairwiseTurnGPT provides a much more nuanced understanding of how lexical content contributes to the predictability of turn-taking behaviour than was previously possible. It also improves the accuracy of predicting turn ends over models of dialogue serialised at the level of turns.

We find that both the timing and content of overlaps contribute jointly to increased predictive power. By comparing training data configurations containing different turning-taking behaviours, we demonstrate the relative contributions of partial, complete, and backchannel overlaps for accurately predicting the variety of turn ends and starts that occur in spoken dialogue. Each training data augmentation improves overall turn-end prediction but through different means. Though the underlying intent of backchannel responses is known to be mediated by their prosodic realisation, our results show that the lexical content alone is already a valuable cue for predicting turn ends (Lai, 2009). Overlaps are also useful but to a lesser extent. Though they complicate non-overlapped turn-end predictions, they are crucial for accurately modelling yielded turns. We find that the alignment configuration containing all forms of overlap can muddy the distinction between yields and normal turn endings. Interestingly, training with this data configuration consistently improves predictive performance across turn start types, suggesting that predicting turn starts and ends may benefit from different information. For example, Jiang et al. (2023) has shown that turn starting points are better predicted when conditioned on the content of the upcoming response. Though related, our results highlight the importance of investigating turn-ends and turn-starts as separate prediction tasks.

By allowing for synchronous streams of lexical content, PairwiseTurnGPT provides nuanced insight into how much lexical context contributes to the prediction of turn-taking behaviours in spoken dialogue. This model has the potential to be used in dialogue systems or for gaining deeper insights into human turn-taking behaviour.

Limitations We selected the Switchboard dataset as a representation of extremely natural spoken dialogue and for its manually annotated transcripts and timestamps. However, it is limited in size. Some of our results suggest that the training set may not be sufficiently large to capture the complexities of these interactions fully. In particular, the cross-attention layer that encodes interactions between the two speaker streams is trained from scratch. For example, the difference in accuracy between the serialised conditions with and without cross-attention in Table 1 is less significant than we might expect. Investigating larger training corpora may allow the model to better capture the interaction between both streams. Exploring the predictability of turn-taking in other types of spoken conversation, such as interviews or conversations between friends, could also further illuminate the role of lexical information in turn-taking prediction. We expect the inclusion of prosodic information to improve turn-taking behaviour prediction further and leave this for our future work.

References

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

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

Steven E Clayman. 2012. Turn-constructional units and the transition-relevance place. *The handbook of conversation analysis*, pages 151–166.

Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Johnathon Hamaker, and Joseph Picone. 1998. Resegmentation of switchboard. In *Internation Conference on Spoken Langauge Processing*.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

Starkey Duncan. 1973. Toward a grammar for dyadic conversation. *Semiotica*, 9(1).

Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns1. *Language in society*, 3(2):161–180.

Starkey Duncan and Donald W Fiske. 2015. *Face-to-face interaction: Research, methods, and theory*. Routledge.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990. Association for Computational Linguistics.

Cecilia E Ford and Sandra A Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184.

John J. Godfrey, Edward Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520 vol.1.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023. Response-conditioned turn-taking prediction. In *ACL Findings of the Association for Computational Linguistics*, pages 12241–12248. Association for Computational Linguistics.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Catherine Lai. 2009. Perceiving surprise on cue words: Prosody and semantics interact on right and really. In *Proceedings of Interspeech*, pages 1963–1966.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Lilla Magyari and Jan P De Ruiter. 2012. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, 3:376.

Angelika Maier, Julian Hough, and David Schlangen. 2017. Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems. In *Proceedings of Interspeech*, pages 1676–1680.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative Spoken Dialogue Language Modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. In *Proceedings of Interspeech*, pages 586–590.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. volume 50, page 696 – 735.

Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12121–12125.

Nigel Ward. 2004. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody International Conference*.

Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.

Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *SLT IEEE Spoken Language Technology Workshop*, pages 831–837. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistic Society*, pages 567–578.

A Model architecture

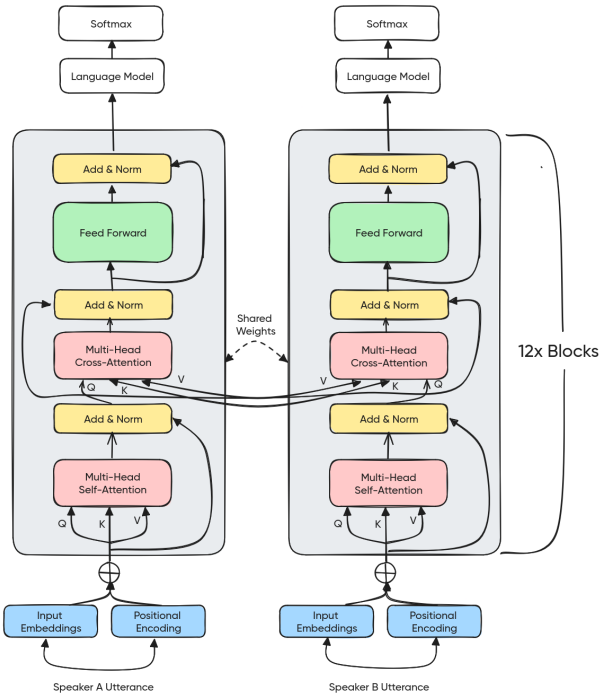


Figure 4: PairwiseTurnGPT Architecture

B Sub-word token alignment

Figure 5 shows how the word ‘uhhuh‘ is decomposed into three sub-word tokens ([“uh”,“h”,“uh”]) under the GPT-2 byte pair encoding tokenizer. The timing of each sub-word token is approximated from the original word-level timestamps by splitting the word duration uniformly across the constituent tokens.

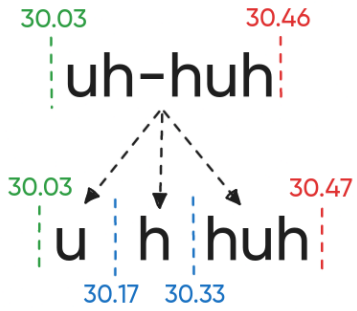


Figure 5: Deriving subtoken alignment.

C Turn type frequency

Turn Type	Token	Count	%
Normal	<eot>	**74522**	**45.19**
Partial Overlap	<yield>	29861	18.11
Overlap	<eint>	16250	9.85
Backchannel	<ebc>	44281	26.85
All	164914	-	

Table 4: The frequency of each turn type in Switchboard using our turn annotation procedure.

Learning Task-Oriented Dialogues through Various Degrees of Interactivity

Sebastiano Gigliobianco
University of Potsdam
Potsdam, Germany
sebastiano.gigliobianco@uni-potsdam.de

Dimosthenis Kontogiorgos
Massachusetts Institute of Technology
Cambridge, MA, USA
dimos@csail.mit.edu

David Schlangen
University of Potsdam
Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

Due to the scarcity of dialogue datasets compared to the vast amount of non-interactive text utilized in large language models, this work aimed to collect dialogues featuring referring expressions in collaborative tasks. In an interactive study, two participants were paired up and presented with the same image of a puzzle. One participant, the instruction giver, had access to an annotated version of the puzzle board, and their task was to find a description that enabled the other participant, the instruction receiver, to identify and select the referent target. The paper investigates whether and to what extent manipulations of the complexity of the task and the degree of interactivity between the users affect the type of referring language that is collaboratively constructed. The results revealed that the aforementioned manipulations had a statistically measurable impact on the type of referring expressions generated by the participants and that interactivity had a major effect on how instructions were collaboratively and iteratively refined.

1 Introduction

The ability to accurately resolve referential entities in text remains challenging. Whether in the domain of information retrieval, question answering, or machine translation, the interpretation and handling of references are fundamental to the coherence of automated text processing systems. Towards addressing these critical challenges, this work presents PentoNav: a dataset containing annotated logs of task-oriented cooperative dialogues.

We developed a collaborative task where two human users were matched in a chat room and were shown a picture of a *Pentomino* puzzle. The key difference between the two users was that one of the participants, the instruction giver (IG), had access to a labelled version of the puzzle with a bounding box around the target piece and had to describe it so that the instruction receiver (IR), who could only see the unlabelled image, could uniquely identify and select the correct piece.

The main question we examined was **whether the complexity of the task and the degree of interactivity between the users have a measurable effect on the type of referring expressions generated**. To study how participants adapt to different settings, we modified the underlying experiment along two main dimensions: *task complexity* and *interactivity between users*.

Our findings indicate that the degrees of interactivity in online interactions have a significant effect on how referring utterances are co-constructed, especially how the feedback of the listener affects the incremental production of referring expressions. Differences were also found in the instruction receiver's task accuracy and response time, as well as in the length of the referring expression produced. Overall, more complex tasks required a higher cognitive load from both participants, indicating that higher task complexity also increases the collaborative effort. Furthermore, a higher degree of interactivity degree also appeared to align with increased accuracy and longer referring expressions.

The resulting dataset (PentoNav) is a publicly available corpus containing 640 *Pentomino* puzzles and descriptions equally distributed among three complexity levels and four experiment designs. PentoNav provides valuable insights into the

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

various strategies participants employ during the collaborative task.

2 Related Work

2.1 Referring Expression Generation

Reference is the linguistic phenomenon in which a noun phrase refers to an entity within a sentence (Stede, 2012). Recent research in the area of referring expression generation has examined how to collect referring expressions generated by humans trying to solve a common task (i.e., the *ReferItGame* (Kazemzadeh et al., 2014)). During such tasks, humans are typically shown pictures of real-world scenes, and generate referring expressions for highlighted objects (Perkins, 2021).

Other datasets combine methods from computer vision and NLP (Loáiciga et al., 2021), investigating phenomena of reference and coreference resolution in task-oriented dialogues with visual support. A lot of referring expression generation work focuses on puzzles such as the *PentoRef* (Zarrieß et al., 2016) and *Pento-DIA Ref* (Sadler and Schlangen, 2023) datasets. Both works use the *Pentomino* puzzle paradigm, that this work also utilizes. In comparison to *PentoRef*, *Pento-DIA Ref* is a synthetic dataset where expressions are generated by the incremental algorithm (Krahmer and van Deemter, 2012).

Some interesting work has been carried out in the area of instruction oriented dialogue. Notable examples contain the Tactical Speaker Identification Speech Corpus (TSID) collected by Graff et al. (1999) or the HCRC Map Task Corpus by the University of Edinburgh (1993). Both corpora feature dialogues between participants tasked with finding a route between two points on a map. Another similar experiment was conducted by Brennan et al. (2013) and differs from the previous studies in that one participant received directions by telephone while searching for target locations on the Stony Brook University campus. Once the target location was reached, the participant had to take a photograph, which was later compared with the target image described by the other participant, alongside the GPS data from the mobile phone.

2.2 Task Complexity

Many studies from the fields of linguistics and cognitive sciences have been conducted to measure the time it takes a person to resolve referential expressions. Elsner et al. (2017), for example, demonstrated how visual complexity measurably affects referring expression generation. During this study, participants were shown abstract scenes containing multiple objects that share some features and were instructed to describe the target piece. Referring expressions were extracted and analyzed, showing how visual complexity can delay or facilitate description generation.

Similarly, Clarke et al. (2013) showed how complicated and cluttered scenes translate to longer referring expressions. The study was conducted by showing participants images from the *Where's Wally* book with a bounding box surrounding the target piece, and they were tasked to write a referring expression for it. The authors were able to find a correlation between the median length of the expression and task complexity showing once again that complexity plays a crucial role when describing objects.

Another setting in which task complexity is commonly used is referential gaze modelling as shown by Alacam et al. (2022) who trained different models on the Eye4Ref work (Alacam et al., 2020) to predict whether a gaze from a participant is directed at a referent object or not. Increasing task complexity was correlated with a decreasing F1-Score.

These studies suggest that task complexity has a measurable effect on people's effort to describe common objects and construct referential expressions, which we use as one of the main dimensions to examine in this study.

2.3 Degrees of Interactivity

While the broad concept that a higher degree of interactivity between participants leads to a higher success rate has been observed in general tasks (Handzic and Low, 2002), de Weck et al. (2019) observed this concept in the field of referring expression generation. In their study, they analyzed the referring expressions of twenty parents telling a story either to their child or to an adult and found an overall wider range of referring expressions when participants talked to children. While not evaluating the strategies itself, the study showed that the interaction setting influences the type of generated referring expressions.

Dialogue is by nature incremental, which means that it's processed step by step as information is delivered (Schlangen and Skantze, 2009). This problem has already been addressed in the past, for example, by Manuvinakurike et al. (2017) who

leverage reinforcement learning to incremental dialogue policy learning in dialogue games and show how this new approach outperforms a human-like baseline system in a collaborative task.

Apart from the incremental nature of human dialogue, the degree of interactivity is also relevant to what type of medium people use to produce referring expressions and how information is distributed to different channels (including the non-verbal channels). Receiving feedback in referring expression generation through backchannels or non-verbal cues has also been shown to affect how references are collaboratively produced (Kontogiorgos, 2022).

Variations in the degree of interactivity play a crucial role: changing the degree of interaction should influence the strategies adopted by people to refer to objects as they may have different ways to receive feedback. This work aims to investigate how these variations affect the production of referring expressions including how task complexity correlates with interactivity.

3 Experimental Setup

Similarly to the *Pento-DIA Ref* dataset (Sadler and Schlangen, 2023), the data collection was conducted by pairing two participants in a chat room with an image of a *Pentomino* puzzle (Figure 1). One of the participants, the instruction giver (IG) was able to see a labelled image with a highlighted piece and had to describe it to the other participant, the instruction receiver (IG), who needed to select it based on the description and the unlabelled image of the same board. To examine strategy differences across diverse settings, four variations of the puzzle's basic design were created, and the complexity of the *Pentomino* boards was modified. During the data collection, the instruction giver was not aware whether the instruction receiver was a human or a computer program.

The participants were recruited using the Prolific platform (Prolific), and the only requirements were proficiency in the English language and being at least 18 years of age. Each participant was only allowed to participate in the study once. The participants were aged between 18 and 58 (on average 27.5 with a standard deviation of 7) and mostly based in Europe. About a third of the participants declared English to be their first language. Out of the 48 participants in total, 27 reported female and 21 male. On average, each participant took 15 minutes to complete the task with a standard deviation of 8 minutes.

The data was collected using SLURK (Götze et al., 2022): an extensible chat server optimized for conducting multi-modal dialogue experiments and data collections, with a framework for creating abstract representations and interfaces to object manipulation tasks.

3.1 Task Complexity

Analogously to work presented by Alacam et al. (2022), participants were shown boards with three different difficulty levels: easy, medium and hard. The complexity of a board is defined by four variables that were used during the process of generation:

- **number of objects:** the total number of objects present on any given board.
- **number of random pieces:** randomly generated pieces are added to the boards to increase variability and prevent generating boards containing only similar pieces.
- **number of similar pieces:** the total amount of pieces on the boards that are grouped based on similar properties. Each piece inside a group shares certain characteristics with other pieces of the same group to add some distractors, thus increasing the complexity of selecting the target object, which is always randomly chosen from one of the grouped pieces.
- **similar pieces per group:** the number of pieces in each group. Grouped pieces share some properties: shape, position, orientation and color. The amount of shared parameters is determined by the difficulty level.

To establish a clear definition of complexity, a pilot study was carried out. Various board settings were explored to identify measurable criteria. The difficulty was measured in terms of speed and number of tokens. The assumption was that a complex task board would require both a higher cognitive load from the participant, and therefore more time to produce it, as well as a higher number of words to describe the target piece.

3.1.1 Pentomino Task Boards

The *Pentomino* boards were generated using the following variables:

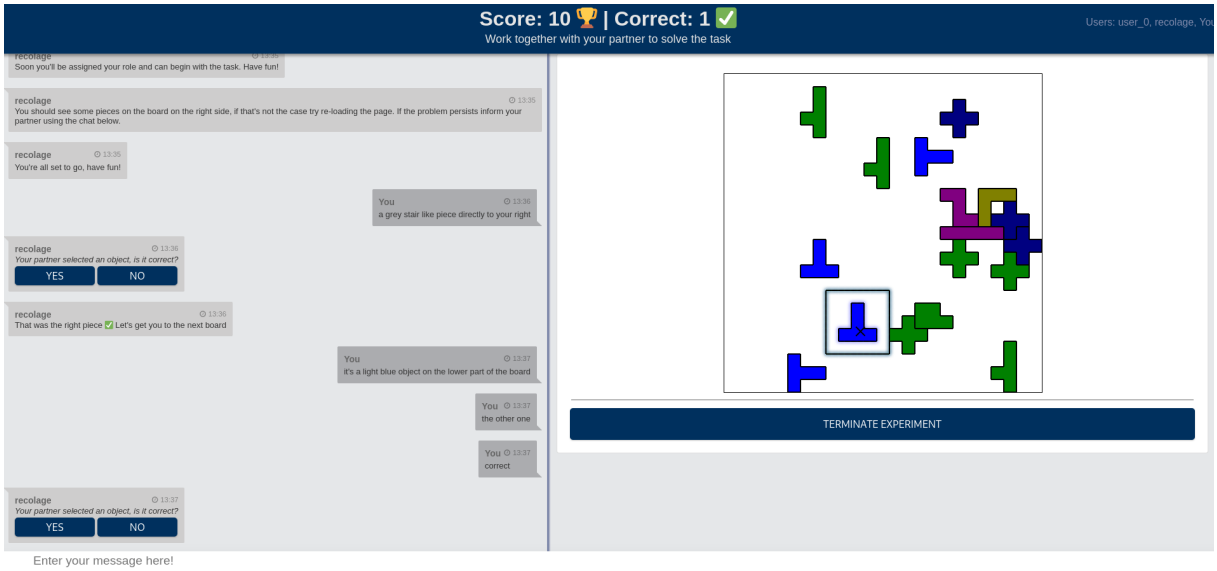


Figure 1: *Interface for the instruction giver*

- **shape:** F, I, L, N, P, T, U, V, W, X, Y, Z.
- **color:** red, orange, yellow, green, blue, cyan, purple, brown, gray, pink, olive green, navy blue.
- **position:** top left, top center, top right, left center, center, right center, bottom left, bottom center, bottom right.
- **orientation:** 0, 90, 180, 270.

As mentioned before, similar pieces shared a predefined amount of variables that were fixed within the group. The only exception was made for the position: during the generation of new objects within a group, there is a 50% chance that an object will be assigned a new position instead of the group position to increase variability. The new position is, however, always adjacent to the group position to maintain similarity.

3.2 Variations in the Degree of Interactivity

Four different experiment designs were developed for this data collection to modify the degree of interactivity in the dialogue. The underlying structure of the experiment remains the same across all variations: the instruction giver has to describe the target object to the instruction receiver, who has to select the object on an unlabelled board.

- **No Feedback**: the first variation removes any means of feedback communication between the users. The IG is only allowed to send one single message to the IR, who can then select the described object with a mouse click. After the first message is sent, the IG is not able to write anything else and the players are not notified by the bot whether the IR's selection was correct.

- **Feedback**: while maintaining the same dynamics of the first variation, this variation allows minimal interaction between the users by notifying users about the outcome of each round.

- **Selection Confirmation**: in this variation, interaction between users is enhanced by having the IG confirm the IR's choice once an object has been selected. Upon selecting the wrong piece, the system allowed the IG to send a new description of the target piece. Points are detracted from the total score every time the wrong object is selected.

- **Gripper**: this variation maximizes interactivity between users by not limiting the number of messages that the IG can send. Moreover, object selection by the IG is achieved by moving a gripper on the board instead of using the mouse. The gripper is fully visible for both users at all time allowing the IG to send additional messages correcting or adding new information to ensure the IR moves in the right direction and selects the correct piece.

4 The Data

During task design, the following factors were taken into consideration:

- **Natural:** the IGs were intentionally not provided with any guidance on what constitutes a helpful or accurate description. This decision aimed to force the IGs to generate their own reference expressions independently, without relying on a predetermined pattern.

- **Diverse:** within the same experiments, some variables were modified, hoping that the IG would come up with different descriptions of the target piece, particularly:

  - **Difficulty level:** more complex task boards should require more complex descriptions to uniquely identify the target object.
  - **Degree of interactivity:** different degrees of interactivity between the users and the interface should have a measurable impact on the type of referring expressions generated.

The resulting dataset is a collection of chat logs. 30 participants took part in the experiment and collected a total of 640 data points equally distributed among the four designs and difficulty levels. A single data point is defined as a combination of a *Pentomino* puzzle, the description provided by the instruction giver and the object selected by the instruction receiver. During the entire data collection, the external participants were always assigned the role of the instruction giver, and one experimenter took the role of the instruction receiver.

Every participant was asked to label 20 boards with the exception of two participants who did respectively 79 and 1 to balance the data points across the experiment's variations. Out of a total of 300 pre-generated *Pentomino* boards, 264 were selected randomly by the system at the beginning of every round. On average, each of the 264 boards was selected 2.5 times, with some boards appearing as often as 7 times.

The complete dataset, together with the raw logs and the scripts used to extract and analyze PentoNav are available on Github.

5 Analysis

5.1 Statistical Analysis

In order to run a statistical analysis of the data, the following features were extracted from the dataset:

- batch position: the order of this board within the 20-boards-batch (extracted to measure order effects).
- interactivity: the degree of interactivity.
- complexity: the complexity level of the board.
- accuracy: whether the IR selected the right object after the IG's description. For the interactivity selection confirmation and gripper, the description is marked as corrected if the IG confirmed the correct selection of the IR.
- target: shape of the target object.
- typing lag: how much time (in seconds) the IG took to start typing the referring expression description of the target.
- description lag: how much time (in seconds) the IG took to send the referring expression description of the target.
- response time: how much time (in seconds) the IR took to select an object after receiving the description from the IG.
- number of tokens: number of tokens in the description.
- number of adjectives: number of adjectives used in the description.
- number of adverbs: number of adverbs used in the description.
- number of nouns: number of nouns used in the description.

Before the feature extraction, the descriptions were first normalized with Pyenchant (Pyenchant) and the linguistic features were extracted with *LFTK* (Lee and Lee, 2023). The normalization step consisted of running the spell checker and replacing wrong-spelled words with the first alternative proposed by Pyenchant.

The scope of this analysis is to find out whether the modifications of the experiments had a statistically significant influence on the generated referring expressions. The statistical analysis was carried out using R and the *lme4 package* (lme4).

interactivity	no feedback			feedback			confirm selection			gripper		
complexity	easy	medium	hard	easy	medium	hard	easy	medium	hard	easy	medium	hard
accuracy	85.25	81.13	89.13	93.75	91.67	89.58	100.00	100.00	100.00	93.75	97.92	97.92
lag to typing	6.25	8.79	7.08	7.44	7.89	8.67	5.48	6.44	7.31	4.93	5.49	5.12
lag to description	15.28	18.26	25.83	25.88	29.54	43.46	26.20	30.73	42.40	24.34	22.17	28.02
reaction time	9.08	8.92	10.00	8.73	8.94	11.17	9.17	9.62	13.33	12.41	13.31	16.65
n tokens	6.31	6.91	9.39	8.30	9.27	12.40	10.70	12.81	16.31	13.08	12.17	14.85
n adjectives	1.31	1.13	1.61	1.81	2.08	2.42	1.83	2.27	2.75	1.62	1.38	1.73
n adverbs	0.15	0.17	0.13	0.22	0.15	0.33	0.47	0.50	0.48	0.31	0.23	0.58
n nouns	1.85	2.08	2.54	2.06	2.19	2.96	2.48	2.71	3.52	3.25	2.92	3.38

Table 1: *Mean values of all variables in all levels of complexity and interactivity.*

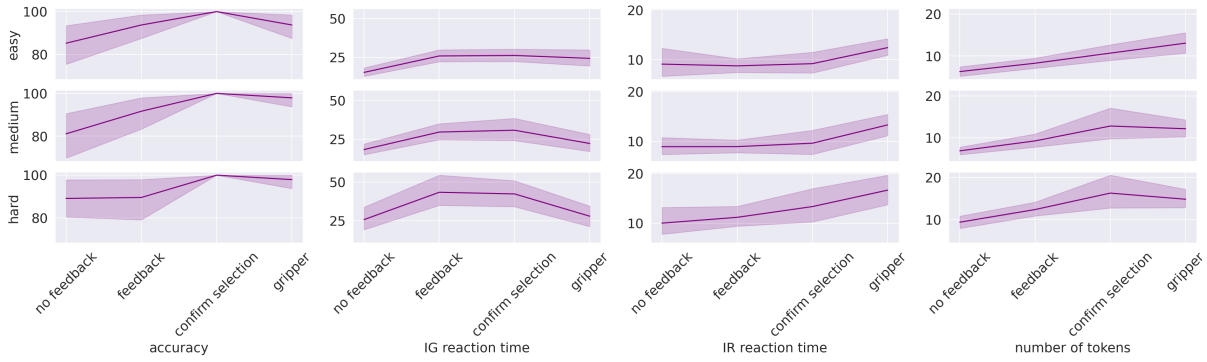


Figure 2: Differences in accuracy, reaction time (IG & IR), and number of tokens per referring expression.

Task Complexity

The results show that the complexity level had a measurable influence on some of the extracted features. The data show that while similar accuracy values can be observed across all three levels, we do see a slight increase in both the time the instruction giver took to both start typing (*lag to typing*) and send the message with the description (*lag to description*). Finally, a substantial increase in reaction time of almost three seconds on the side of the IR can be observed when comparing the easy/medium boards (which have similar values) to the hard scenes.

A look at the linguistic features also indicates that the increasing level of complexity of the board required on average a longer description with an increased number of adjectives and nouns. This initial evidence was also confirmed by training and comparing linear mixed-effects models to fit the data by maximum likelihood with the following parameters:

- fixed effect: complexity level
- random effects: target object, participant, batch position, and interactivity

outcome variable	p-value	χ^2
accuracy	0.7557	0.0968
lag to description	<0.001	41.866
lag to typing	0.03814	4.2989
reaction time	<0.001	11.432
n tokens	<0.001	34.136
n adjectives	<0.001	11.483
n adverbs	0.1123	2.5216
n nouns	<0.001	20.516

Table 2: *Linear mixed effect models: complexity*

All the models were fitted to the data to various outcome variables, which are listed together with the respective *p-values* and χ^2 *values* in *table 2*.

Degree of interactivity

A statistical difference in the data was also found while investigating the effects of the degree of interactivity. The most evident difference can be noted in the accuracy: with increasing levels of interactivity, the accuracy and length of the descriptions also

outcome variable	p-value	χ^2
accuracy	0.04846	3.8939
lag to description	0.7054	0.143
lag to typing	0.05641	3.6399
reaction time	0.02592	4.961
n tokens	0.03131	4.636
n adjectives	0.8686	0.0274
n adverbs	0.07763	3.1138
n nouns	0.06345	3.4448

Table 3: *Linear mixed effect models: interactivity*

outcome variable	p-value	χ^2
accuracy	0.2182	1.5161
lag to description	<0.001	13.18
lag to typing	<0.001	49.852
reaction time	0.4958	0.4638
n tokens	0.9606	0.0024
n adjectives	0.2592	1.2728
n adverbs	0.9336	0.007
n nouns	0.3468	0.8853

Table 4: *Linear mixed effect models: batch position*

increase. Interestingly, the time that the IG needs to start typing decreases while the total time needed to send the description raises from 19.3 seconds in the *no feedback* design to around 32 seconds in both the *feedback* and *selection confirmation* variations to finally fall back to 24.79 seconds in the *gripper* setting. The latter can be tracked down to the fact that in the last setting, the IG was able to send an unlimited number of messages and some users sent multiple shorter messages instead of a longer one, indicating an incremental behavior. While a small increase in the IR's reaction time can be observed when comparing the values of the *no feedback*, *feedback* and *confirm selection* settings, an increase of around 3.5 seconds can be measured in the *gripper* setting. This increase, however, was expected as the gripper is positioned at the center of the board at the beginning of every round and must first be moved on the object that the IR intends to select.

As for the complexity level, linear mixed effect models were also trained to fit the data, and the results are reported in *table 2*. While training the following models, the following parameters were used:

- fixed effect: design
- random effects: target object, participant, batch position, and complexity level

The outcome variables together with the *p-values* and χ^2 *values* are reported in *table 3*.

Batch position

The position of the instance within the batches of 20 boards labelled by the participants also seems to somehow affect the referring expressions with regard to the extracted features. Noteworthy is the effect on the accuracy and the time required by the IG to both start typing and send a message. While for the accuracy, a slight increase can be seen, which indicates that there is a learning effect during the task, this increase only affects the three settings in which the users receive feedback about the piece selected by the IR (feedback, selection confirmation and gripper). The position of the data instance within the batch does not seem to influence the IR's reaction time in any way. With regards to the typing and description lag, on the other hand, a decrease can be measured across all designs.

Linear mixed-effect models were trained with the following parameters:

- fixed effect: batch position
- random effects: target object, participant, interactivity, and complexity

The results are shown in *table 4*.

6 Discussion & Conclusion

In this work, we presented PentoNav: a dataset composed of annotated *Pentomino* puzzles and natural referring expressions generated by the participant to describe one of the objects. The research question postulated at the beginning of this work was whether a manipulation in the degree of interaction between the users and the complexity of the puzzle itself might have an impact on the strategies adopted by the participants to solve the task.

The analysis showed how different degrees of interaction between users, as well as manipulations in task complexity, have a measurable impact on the generated descriptions. Both hypotheses postulated at the beginning of this paper, namely that an increasing level of interaction and puzzle complexity would influence the instruction receiver's accuracy as well as the descriptions generated by the instruction giver, were partially confirmed by the statistical analysis of the data.

One variable that was not considered during planning was the effect of the position of the current data point within the 20 puzzle batches in which the experiment was divided. During the analysis, the position of the data point revealed the learning effect of the participants. This tendency was confirmed by the linear mixed models: while the accuracy increases, the typing and description lag decrease consistently across all designs. This confirms that while progressing through the batch, the participants providing instructions become not only more effective but also faster at generating referring expressions.

6.1 Future work

For the analysis conducted, only a subset of information was extracted from the chat logs. These still contain other information, such as the mouse movements of the instruction receiver on the *Pentomino* board, which can be used to potentially reconstruct the IG's reasoning after receiving a description from the instruction giver. Further insights into the cognitive process of analyzing the description and the board could be offered by the analysis of the instruction receiver's eye movements.

Another interesting application area for this dataset is reinforcement learning. Similarly to the work proposed by Sadler et al. (2023) and Vogel and Jurafsky (2010), an artificial agent can be trained to substitute the instruction receiver and navigate the *Pentomino* board in search of the target piece. Such artificial agents can be deployed online for a subsequent round of data collection, engaging with human participants. The outcome from such agents could be compared to PentoNav to yield valuable insights into the differences between how humans interact with artificial agents versus other human participants.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 423217434 ("RECOLAGE") grant. We would like to thank the anonymous reviewers for their valuable feedback.

References

Özge Alacam, Eugen Ruppert, Amr Rekaby Salama, Tobias Staron, and Wolfgang Menzel. 2020. Eye4Ref: A multimodal eye movement dataset of referentially complex situations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2396–2404, Marseille, France. European Language Resources Association.

Özge Alacam, Eugen Ruppert, Sina Zarrieß, Ganeshan Malhotra, Chris Biemann, and Sina Zarrieß. 2022. Modeling referential gaze in task-oriented settings of varying referential complexity. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 197–210, Online only. Association for Computational Linguistics.

Susan E. Brennan, Katharina S. Schuhmann, and Karla M. Batres. 2013. Entrainment on the move and in the lab: The walking around corpus. In *Cooperative Minds*, Cooperative Minds: Social Interaction and Group Dynamics - Proceedings of the 35th Annual Meeting of the Cognitive Science Society, CogSci 2013, pages 1934–1939. The Cognitive Science Society. Publisher Copyright: © CogSci 2013.All rights reserved.; 35th Annual Meeting of the Cognitive Science Society - Cooperative Minds: Social Interaction and Group Dynamics, CogSci 2013 ; Conference date: 31-07-2013 Through 03-08-2013.

Alasdair Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's wally: The influence of visual salience on referring expression generation. *Frontiers in psychology*, 4:329.

Geneviève de Weck, Anne Salazar Orvig, Stefano Rezzonico, Elise Vinel, and Mélanie Bernasconi. 2019. The impact of the interactional setting on the choice of referring expressions in narratives. *First Language*, 39:014272371983248.

Micha Elsner, Alasdair Clarke, and Hannah Rohde. 2017. Visual complexity and its effects on referring expression generation. *Cogn Sci*, 42 Suppl 4:940–973.

Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. The slurk interaction server framework: Better data for better dialog models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.

David Graff, Douglas Reynolds, and Gerald C O'Leary. 1999. Tactical speaker identification speech corpus (tsid).

Meliha Handzic and Graham Low. 2002. The impact of social interaction on performance of decision tasks of varying complexity. *OR Insight*, 15(1):15–22.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Dimosthenis Kontogiorgos. 2022. *Mutual Understanding in Situated Interactions with Conversational User Interfaces: Theory, Studies, and Computation*. Ph.D. thesis, KTH Royal Institute of Technology.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

lme4. lme4. https://cran.r-project.org/web/packages/lme4/index.html. Accessed: 31/12/2023.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 331–341, Saarbrücken, Germany. Association for Computational Linguistics.

Hugh Perkins. 2021. Texrel: a green family of datasets for emergent communications on relations. *CoRR*, abs/2105.12804.

Prolific. https://www.prolific.com/. Accessed: 31/12/2023.

Pyenchant. Pyenchant. https://github.com/pyenchant/pyenchant. Accessed: 31/12/2023.

Philipp Sadler, Sherzod Hakimov, and David Schlangen. 2023. Yes, this way! learning to ground referring expressions into actions with intra-episodic feedback from supportive teachers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9228–9239, Toronto, Canada. Association for Computational Linguistics.

Philipp Sadler and David Schlangen. 2023. PentoDIARef: A diagnostic dataset for learning the incremental algorithm for referring expression generation from examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2106–2122, Dubrovnik, Croatia. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

M. Stede. 2012. *Discourse Processing*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

University of Edinburgh. 1993. Hcrc map task corpus.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814, Uppsala, Sweden. Association for Computational Linguistics.

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

Behaving according to protocol: How communicative projects are carried out differently in different settings

Ellen Breitholtz & Christine Howes
University of Gothenburg,
Department of Philosophy, Linguistics and Theory of Science; CLASP
ellen.breitholtz@ling.gu.se christine.howes@gu.se

Abstract

There are a number of theories and models for capturing the aspects of organisations that are systematically related to the modes and genres of communication taking place within them. In this paper we will consider the micro-level of organisations and present a model of how similar communicative projects are carried out differently within different activities. Central to our account is the notion of conversational games, which can be seen as strategies for realising communicative projects while assigning speaker roles to dialogue participants.

1 Introduction

It is well established in the literature on organisational communication that the type of organisation affects the type of communication occurring in it, and vice versa (Baker, 2007; Brown and Starkey, 1994; Yates and Orlikowski, 1992). This has lead to a number of theories and models for capturing the aspects of organisations that are systematically related to the modes and genres of communication that take place within them.

In this paper we will consider the micro-level of organisations and present a model of how particular interactions play out within social activities and communicative projects of certain types. In section two we will discuss some categories that have been used to analyse interactions and to define different types of interaction. In Section 3 we will show how these categories can be implemented in a formal theory of dialogue. In Section 4 we consider the dialogue moves involved in a particular conversational game of the type in which a suggestion is made. We consider two different dialogues where this game is being realised in two different ways depending on contextual parameters. Finally, in Section 5, we draw some conclusions

2 Defining interactions

There are several areas of research which aim to categorise interactions in ways that are predictive of their communicative (including linguistic) features. These theories are based on a variety of concepts such as (social) (communicative) *activity* (Allwood, 2000), (communicative) *project*, *frame* (Levin and Moore, 1977; Carlson, 1982), (language) (dialogue) *game* (Lewis, 1979; Ginzburg, 2012), *genre*, etc. In this section we provide a brief overview of some of these concepts and theories.

When defining genres, types of language use etc. a frequently used concept is that of *activity*, as in the activity in the context of which language occurs. A communicative activity can be described as a comprehensive communicative project tied to a socio-cultural situation type reminiscent of the Wittgensteinian concept "form of life" (Allwood, 2000; Malcolm, 1954). On Allwood's account an activity type is characterised by the *goals*, *roles*, *artefacts* and *environment* that are associated with it. The carrying out of an activity consists of a number of sub-goals being completed. These may be more or less communicative in nature. For example, instances of the activity type "Buying/selling coffee in a café" are made up of sub-goals such as "conveying which product one wants to order", "conveying how much the costumer should pay", and, finally, "paying/receiving money".

Linell (2009) also draws on activity types for analysing interactions, but he emphasises the *communicative projects* that make up activity types. Communicative projects are often strongly associated with the sub-goals of particular activities. Thus in the café-interaction, the goal "conveying which product one wants to order" is linked to a project like "establishing an order". Another notion that has been used to define different classes of interactions is the concept of *genre* (see for example Ginzburg, 2016).

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

In the context of literature the concept of genre attempts to capture notions of subject matter, content and vocabulary as well as style, and it is used in a similar way in linguistic theory. Thus the genre "conversation in a bakery" is a monolithic type that involves specific vocabulary as well as grammatical constructions and dialogue moves. This way of thinking about interactions is intuitive and can be useful, and offers a "blueprint" of the characteristics of particular types of interactions. However, what is missing is a way of capturing differences and similarities between interactions that are directly related to specific contextual parameters. For example, an interaction in any shop or institution where a customer or client talks to a sales assistant or receptionist is likely to involve a lot of greetings and good byes, but the vocabulary and level of formality might vary depending on what kind of shop or other institution we consider. Thus we treat 'activity type' and 'project type' as independent categories, where one project type may be embedded in many different activity types – although not necessarily carried out by means of identical conversational strategies. This last fact prompts us, following Breitholtz (2020), to introduce an additional category by means of which to define interactions – that of *conversational game*, reminiscent of Wittgenstein's language game.

We think of a conversational game as a procedure for carrying out a particular project, and depending on the context different games may be available to do that. In addition, a particular game may have different possible moves available at each point in time. For example, making a decision might in some contexts play out in terms of one person suggesting something, another dialogue participant asking a follow up question, receiving an answer and then accepting, or it could play out as one dialogue participant simply telling the other(s) what to do. Which of these is most likely depends to a great extent on the activity the exchange is part of. Consider for example the dialogues below:

(1) BNC HM6 189-192

U I propose that Sir Simon [last or full name], a director retiring by rotation <pause> be and if hereby re-elected a director of the company.

M Put the resolution to the meeting.
Those in favour <pause> any against <pause> thank you.
I declare the resolution carried.

(2) BNC FM2 167-187

A I was thinking of asking Monica if she could record something from the the Model Car Club and

W Yeah.

A their club meetings.

W Oh that'd be okay I think.

V Is that <pause> Monica?

A Yes.

V Erm she's alread already asked her Dad but

A Right.

V but they don't have <pause> meetings. They they meetings take place during the evening er as very sort of <pause> at the <unclear>

A Right. Okay.

V <unclear> it wouldn't work. And we've

A Right.

V missed the A G M.

A Okay.

V It was earlier this year.

A Fair enough.

(1) is an excerpt from a formal meeting where proposals are formally made according to particular rules, while the dialogue in (2) is more informal and several participants are contributing acceptances, clarification requests, etc. However, a game type revolving around a suggestion or proposal is relevant to both of these dialogues. We will refer to this type of game as the *suggestion game*. Basic requirements on such a game are that there are at least two participants, one of whom makes a suggestion regarding some project that is believed to be shared. The other agent (or agents) responds to the suggestion, for example by *accepting* or by *rejecting* the proposal. We refer to whoever makes the first move (that is makes a suggestion) as player 1'. This move may optionally be followed by a motivation for the suggestion, again by player 1. Another player (player 2) may respond to the suggestion by accepting or rejecting the move. Note that this move does not necessarily have to be an actual response. Depending on the level of grounding we find acceptable in the context, abstaining from protesting might be enough to signal acceptance of a given suggestion (as in (1) where agreement and disagreement is signalled by non-verbal means). '

This way of thinking about acceptable moves in a dialogue is analogous to the way syntax is

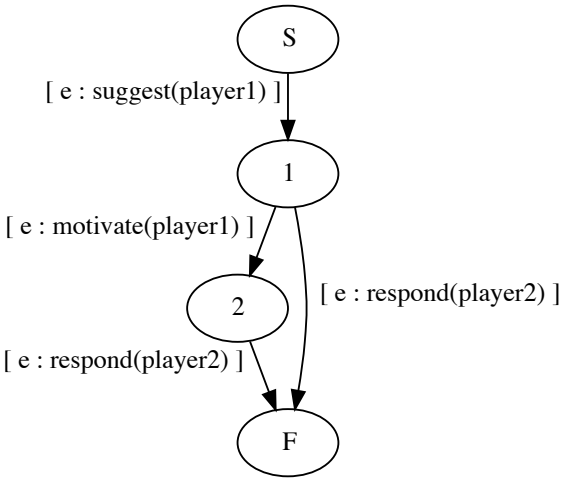


Figure 1: The Suggestion game

viewed in Dynamic Syntax (DS; Kempson et al., 2001; Cann et al., 2005), which is an action centred approach that takes the processes of building up interpretations in interaction as central to how language is used. In recent formulations of DS, the possibilities for future actions are represented as a Directed Acyclic Graph (DAG; see e.g. Howes and Eshghi, 2021), which constrains the space of possible transitions (moves) both within an utterance and between utterances, by restricting the range of (probabilistically) predictable next words or actions. Our formulation of specific games which become relevant at certain points within a dialogue – such as the suggestion game shown in Figure 1 serves the same purpose in making certain follow-up moves more likely in a given context. The types of games available may be further specified according to particular conventions of the organisations in which they occur. For example (1) occurs in the formal business setting of an Annual General Meeting, in which certain conventions apply, such as assenting to (or dissenting from) suggestions by raising your hand, with the majority vote prevailing. In contrast, (2) occurs in a less formal meeting between transcribers of the British National Corpus, and as such is less conventionally structured. This also means that the same sorts of actions (such as a suggestion being made and then either taken up or rejected) play out in different ways, despite being underpinned, in some sense, by the same game.

The moves in Figure 1 would suffice to account for an interaction where Player 1 makes a suggestion – optionally followed by a motivation – to player 2, who responds by either accepting or rejecting it. However, we would also like to allow for less straightforward rounds of the suggestion game, involving things like clarification requests, questions regarding other aspects of the context, etc, as in (2). For example, player 2 might ask for a reason for suggesting φ. This would be perfectly acceptable dialogue behaviour, and players must be allowed, within the suggestion game, to move into games of other types like the *clarification game*.

The ability to move between games reflects the expectations we have when engaging in dialogue – if you ask someone a question, you know that it is likely that you will get a response. However, we can still account for dialogue behaviour which does not conform to one particular game, since we allow dialogue participants to introduce new games – and even new projects.

We therefore want rules which allow for the suggestion game to be played in a number of different ways, including detours into other games. But let us leave that aside for the moment, and just consider the possibilities realised in (2).

3 Updating the Dialogue Game Board

For our model we will use TTR, a type theory with records (Cooper, 2005, 2014, 2023). TTR is a rich type theory, which can account for a range of linguistic phenomena, including many which are particular to dialogue (Cooper, 2005; Ginzburg, 2012; Cooper and Ginzburg, 2015; Lücking, 2016) Two key notions in TTR for dialogue are *Information State Update* (ISU), introduced by (Larsson and Traum, 2000) and (Larsson, 2002), and *Dialogue Gameboard* (DGB) Ginzburg (1994, 1996, 2012). The ISU approach considers the information states of each dialogue participant and how these are updated based on moves in the dialogue. Following Ginzburg, we will model information states as DGBs keeping the "score" of the dialogue in terms of sets of moves, presuppositions, questions, commitments, and other linguistic features which are relevant in the interaction.[1]

TTR is based on the capability in humans (and animals) to perceive and classify the world in terms of categories or *types*.

[1]The notion of game as a metaphor for conversation is not uncommon, see for example Wittgenstein (1953) and (Lewis, 1979).

Formally, the judgement that a particular object, a, is of a certain type T, can be described as $a : T$. The basic type of objects such as humans, animals and things in TTR is *Ind*, the type of individuals.[2]

In TTR *record types* are used in order to represent complex situations which involving many ptypes and individuals. A record type is a structure of pairs of labels and types. Often, the same letters that are used as those used to represent individual variables in other systems – 'x', 'y', etc.– are used as labels associated with the type *Ind*. For *constraints* on the type of situation represented by the record type we use the label 'c' with different subscripts. In (3) we see a record type representing a type of situation where a cat purrs.

(3)
$$\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\text{cat}} : \text{cat(x)} \\ \text{c}_{\text{purr}} : \text{purr(x)} \end{bmatrix}$$

The label 'x' in (3) points to an object of type *Ind*, and there are two constraints on the type of situation, namely that this individual is a cat (c_{cat}:cat(x)) and it purrs (c_{purr}:purr(x)).

In order to account for dialogue in TTR we use a version of the DGB which largely follows previous work (Ginzburg, 2012)[3] Following Cooper (2023) we treat the type of the information state of a conversational participant (the DGB of that participant) as a record type.

We think of the information state of an agent engaged in dialogue as comprising two types of information – that taken to be shared by the dialogue participants (similar to common ground Stalnaker, 1978; Clark et al., 1991) in the situation at hand, and the information taken to be private. As an example, let us assume that the type of an agent's shared information state is T_s and the type of the same agent's private information state is T_p. We see the type of that agent's information state in (4).

(4)
$$\begin{bmatrix} \text{private} : T_p \\ \text{shared} : T_s \end{bmatrix}$$

If we want to represent a "suggestion dialogue" in terms of updates of information states, we need rules handling not only the explicit moves represented in 1, but also *tacit* updates of the DGB. Tacit moves within a game represent inferences and other internal processes. We will now have a look at some of the updates of the DGBs of some of the participants throughout (2).

4 Analysing a suggestion dialogue

4.1 Initial tacit moves

In order to account for communicative games on the DGB we introduce the field *games*.

It is not until the first move is made, and thus considered shared by the conversational participants, that which game is being played can be expected to be shared, and thus appear on the shared game board.

At the beginning of the interaction in (2) the DGB of dialogue participant A is empty apart from the *project*, which we assume to be shared since the necessity of finding some data to record is obvious to both A and their interlocutors (who we shall refer to collectively as B in what follows for the sake of simplicity) in the context of the meeting. We represent a project as a type of event to be brought about by a number of agents. In (5) we see the type of a decision project, $T_{DecisionProject}$. $A_1, \ldots, A_n$ are dialogue participants and *Issue* the thing that is to be decided upon.

(5) $T_{DecisionProject}$ =
$$\left[\text{e} : \text{decide}\left(\{A_1, \ldots, A_n\}, \textit{Issue}\right)\right]$$

To allow representation of sequences of projects, fulfilling some complex goal (linguistic or other), the type *project* on the DGB is list(*RecType*). We would also be able to account for projects suddenly appearing in the information states of dialogue participants due to sudden events, such as "find shelter from the rain".

(6)
$$\begin{bmatrix} \text{private} : \textit{RecType} \\ \text{shared} : \left[\text{project} = [\,[\text{e} : \text{decide}(\{A, B\}, \text{data})]\,] : \text{list}(\textit{RecType})\right] \end{bmatrix}$$

(6) shows the type of the speaker, A's, information state at the beginning of the interaction in (2). For now we are interested only in the information state of dialogue participant A, not that of the listener, dialogue participant B. "Data" represents the issue of which data to collect, in (2).

The first update of the dialogue gameboard is an update of 'private games', that is the repository of conversational games which are salient with respect to a dialogue participant in a given context. Before

[2] *Ind* corresponds to *entity*, *e*, in Montague semantics (Montague, 1973).

[3] For a background on gameboard semantics in TTR the interested reader is referred to Ginzburg (2012) and Cooper (2023).

we move on to how we want to represent this update in TTR, let us have a look at the nature of projects and games in terms of types.

As illustrated in (5) we perceive a project as a record type representing the type of an event where a number of individuals (in this case A and B), jointly perform some action (in this case making a decision) regarding some non-decided-upon issue.

We may think of the development of a conversation as a finite state automaton where the arrows leading from one state to another correspond to the linguistic moves of the conversation, as represented in Figure (1). Instead of focusing on the states between the moves, we could focus on the sequence of moves themselves when defining a conversational game. We would then get a string of move types. The type in (7) for example, is of strings of moves comprising the type of *suggestion game*, $T_{SuggestionGame}$, – a suggestion by player 1 followed by an optional *motivation* by player 1, followed by a *response* (acceptance or rejection) by player 2. We represent move types as record types. A game of the type in (7) is made up of a suggestion, followed by an optional motivation by the dialogue participant who made the suggestion, followed by a response (either an *accept*- or a *reject* move) by the other player.

(7) $T_{SuggestionGame}$=$\left[\text{e : suggest(player1)}\right]^\frown$
$\left[\text{e : motivate(player1)}\right]^{\leqslant 1\frown}$ $\left[\text{e : respond(player2)}\right]$

The notation $\left[\text{e : motivate(player1)}\right]^{\leqslant 1}$ means that the suggestion move is followed by at most one motivation move ($^{\leqslant 1}$). [4] The string in (7) represents the type of a suggestion game on an abstract level – from this type we learn the sequence of move types involved and the relation between the *roles* that are necessary to play the game. However, in order for the game to work as a motor in the dialogue driving the updates, we need to assign the roles of the game to the individuals present in the context. For example, the player who initiates the game by making a suggestion has to be distinct from the player who acknowledges that suggestion.

4.2 Rules for updating private games

There are at least two different scenarios which would lead to an update of private games. First, there is the type of situation where the presence of a project on the DGB causes an agent to search his long term memory for a strategy by which to carry out that project, and load it onto the DGB. The second is when there is already a game on private games that would suffice to carry out the project. Assume for example that A has been thinking since he got out of bed in the morning that he wants to ask Monica to record some meetings from the Model Car Club. He has been meaning to suggest it for a while (or maybe hoping that B will suggest it), thus the suggestion game is activated on his private DGB. When A and B reach the point at which the issue of which data to collect becomes necessary to address, the project appears on the shared DGB. In this case the only update necessary on A's DGB is to place $T_{SuggestionGame}$ first in the list of games, while B has to retrieve the game from long term memory and load it onto private games. The idea here is that the update rules are combined with a control algorithm selecting which rule to apply in a given context. In Figure (2) we see a visualisation of the algorithm controlling the update of private games.

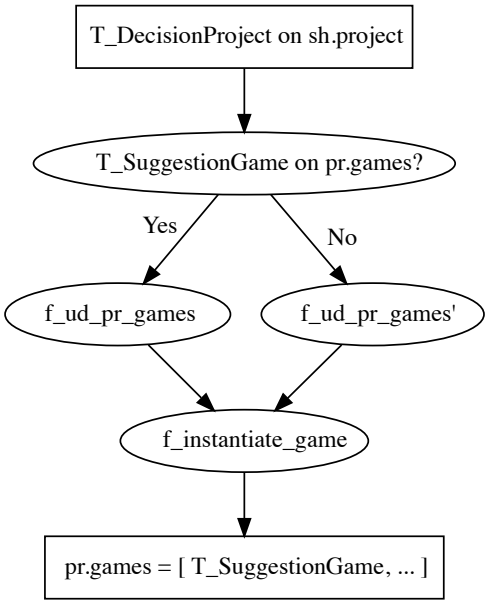


Figure 2: Update of private games

4.2.1 Update private games

We want the first rule $f_{ud\_pr\_games}$ to apply in a context where an agent has a project on her gameboard,

[4] One could argue that a suggestion might be followed by more than one move motivating the suggestion, and it would of course be possible to alter $^{\leqslant 1}$ to $^{\leqslant 2}$ or $^{\leqslant 3}$ or even $^{\leqslant +}$ using the kleene plus to mean one or more (with a corresponding loop in Figure 1) depending on how many motivation moves the model should allow.

but the game first on the list of private games is not relevant to carry out the project. The agent is then licensed to either reraise a game already on private games (but not first on the list) or to load a relevant game from resources onto private games. Now, one question that arises here is what it means to be a relevant game in relation to a particular project. One way of describing this would be in terms of licences in an agent's resources. If an agent has in her resources a link between a type of project T_P and a type of game T_G, she has a licence to carry out a communicative project of type T_P by means of T_G, and may load it onto 'private.games' on her DGB.

Which types of games are relevant to carry out particular types of projects is an empirical question. We think of the update rules licensing the carrying out of a project by means of a particular type of game as reflecting the pragmatic norms of a community. One way of modelling how an agent selects a strategy – for example choosing between an indirect and a direct speech act – would be to extend the model with a probabilistic component (see for example Eshghi and Lemon, 2014). However, in the limited model we are focusing on here, we assume that we have access to only one type of game which is relevant to the project at hand. Moreover, it seems to us that a limited set of project types and game types would suffice to account for a large number of dialogue situations. Thus, for each project type we would introduce a set of postulates defining which games could be relevant to carry out a project of that type. We use the notation relevant_to(T_1, T_2) to represent relevance of T_1 in relation to T_2.

When a communicative project appears on an agent's DGB and the agent initiates carrying out the project there are, as mentioned above, two possibilities. Either there is a game present in the private games field of the DGB by means of which the project can be carried out, or there is not. In the first case we want to make sure that the appropriate game is moved up to the first slot on the list of private games. In the second case, we want to pick an appropriate game from the agent's long term memory, and place it first on the list of private games. The update of 'private.games' thus consists of three rules: $f_{ud\_pr\_games}$ for reraising a game, $f_{ud\_pr\_games'}$ for uploading a game from resources, and – to complete the update – f_{inst_game}. In an instantiated game the roles (player1, player2, etc.) are assigned to dialogue participants from the point of view of the participant on whose gameboard the instantiated game appears. This means in the case of the suggestion game, that when A starts carrying out a decision making project by initiating a game of type $T_{SuggestionGame}$, she has also taken on the role of 'player 1' in that game. In every move type of the instantiated game on her DGB the move to be carried out by player 1 will be assigned to SELF, the ones by player 2 to OTHER.

Even though instantiated games involve assignments of roles to dialogue participants, we still want to be able to treat them as types. For this reason, the type of games is a *join type*. A join type is a disjunction such that, for any two types T_1 and T_2 you can form the join $T_1 \vee T_2$. a : $T_1 \vee T_2$ just in case either a : T_1 or a : T_2. This means that the type of games, T_{Game}, in our theory is a join of the types non-instantiated game, $T_{NonInstGame}$ and instantiated game, $T_{InstGame}$ as defined in (8):

(8) $a : T_{Game}$ iff $a : T_{NonInstGame}$ or $a : T_{InstGame}$

By defining the type of game as a join, we make sure that we can handle situations where, for example, something sudden and unexpected happens, and dialogue participant needs to postpone the initiation of a game already on the DGB. We will look at the instantiation process in more detail further on in this section.

(9) $f_{ud\_pr\_games}$ =

$$\lambda r{:}\begin{bmatrix}\text{pr} : \begin{bmatrix}\text{games} : \text{list}(T_{Game})\end{bmatrix} \\ \text{sh} : \begin{bmatrix}\text{project} = [\ T_{DecisionProject}\ ] : \text{list}(\textit{RecType})\end{bmatrix}\end{bmatrix} .$$
$$\lambda e{:}\begin{bmatrix}\text{g} : T_{SuggestionGame} \\ \text{c}_1 : \text{in(g, } r\text{.pr.games)}\end{bmatrix} .$$
$$\begin{bmatrix}\text{pr} : \begin{bmatrix}\text{games} = [\mu(e\text{.g, } r\text{.pr.games})\ ] : \text{list}(T_{Game})\end{bmatrix}\end{bmatrix}$$

In (9), $f_{ud\_pr\_games}$ takes a situation of the type where there is a decision project on 'shared.project' and, if there is a game of type $T_{SuggestionGame}$ on private games in that record, the function returns a type of situation where that game type is first on 'private.games'.

We think of the update rule $f_{ud\_pr\_games'}$, as seen in (10) as a function from an information state where an agent has a decision project on her gameboard but no game of type $T_{SuggestionGame}$ on the list of games on 'private.games'[5], to an information state where the agent has a decision project on 'shared.project' and a suggestion game first on 'private.games'. In this case the game $T_{SuggestionGame}$

[5]There may be other games on the list of private games, just not the game *suggestion game*.

has to be retrieved from parts of the agent's resources which are external to the DGB.

(10) $f_{ud\_pr\_games'} =$
$\lambda r{:}\begin{bmatrix}\text{pr} : [\text{games} : \text{list}(T_{Game})] \\ \text{sh} : [\text{project} = [T_{DecisionProject}] : \text{list}(\textit{RecType})]\end{bmatrix}.$
$\lambda e{:}\begin{bmatrix}\text{g} : T_{SuggestionGame} \\ \text{c}_1 : \neg\text{in}(\text{g}, r.\text{pr.games})\end{bmatrix}.$
$[\text{pr} : [\text{games} = [e.\text{g} \mid r.\text{pr.games}] : \text{list}(T_{Game})]]$

The functions in (9) and (10) are similar to the update functions discussed by Cooper (2023). In order to obtain the required update of such a function we need to apply it to the *current information state* – that is the information state at the start of the update – of the agent whose information state we seek to capture. Let us consider a scenario where agent A has previously considered suggesting Walnut Street, but was distracted by an event which the agent has just observed. This caused another conversational game, T_{G_X}, to appear on the DGB. His initial information state is thus of the type in (11), which we refer to as $T_{current}$.

(11) $T_{current} =$
$\begin{bmatrix}\text{pr} : [\text{games} = [T_{G_X}, T_{SuggestionGame}] : \text{list}(T_{Game})] \\ \text{sh} : [\text{project} = [T_{DecisionProject}] : \text{list}(\textit{RecType})]\end{bmatrix}$
$s_{current} : T_{current}$

Before we apply the function we need to make sure that the type of the current information state is a subtype of the domain type of $f_{ud\_pr\_games}$. We should point out here that the type of the current information state might very well have other fields such as a shared game, a latest utterance, shared beliefs, etc., and still be a subtype of the domain type of $f_{ud\_pr\_games}$.

In (12) we see the application of $f_{ud\_pr\_games}$ to $s_{current}$, followed by an asymmetric merge of the result of that function application and the type $T_{current}$ of $s_{current}$ (as well as e_1 witnessing the condition that $T_{SuggestionGame}$ is in $s_{current}$.pr.games).

(12) a. $f_{ud\_pr\_games}\ (s_{current})(e_1)=$
$[\text{pr} : [\text{games} = [T_{SuggestionGame}, T_{G_X}] : \text{list}(T_{Game})]]$

b. $T_{current}\ \boxed{\wedge\!\!\cdot}$
$[\text{pr} : [\text{games} = [T_{SuggestionGame}, T_{G_X}] : \text{list}(T_{Game})]] =$
$\begin{bmatrix}\text{pr} : [\text{games} = [T_{SuggestionGame}, T_{G_X}] : \text{list}(T_{Game})] \\ \text{sh} : [\text{project} = [[\text{e} : \text{decide}(\{A_1, A_2\}, \textit{Issue})]] : \text{list}(\textit{RecType})]\end{bmatrix}$

4.2.2 Instantiation of game

After an update putting a game which is a subtype of $T_{SuggestionGame}$ first on the list of private games (either by $f_{ud\_pr\_games}$ or $f_{ud\_pr\_games'}$), we need to instantiate the game, that is associate the roles of the game with the players in this particular situation. To do this we apply the function $f_{inst\_T_{SuggestionGame}}$ to a record assigning the values 'SELF' and 'OTHER' to the roles of the suggestion game.

(13) $f_{inst\_T_{SuggestionGame}} =$
$\lambda r{:}\begin{bmatrix}\text{player1} : \textit{Ind} \\ \text{player2} : \textit{Ind}\end{bmatrix}.\ [\text{e} : \text{suggest}(r.\text{player1})]^\frown$
$[\text{e} : \text{motivate}(r.\text{player1})]^{\leqslant 1\frown}\ [\text{e} : \text{respond}(r.\text{player2})]$

For dialogue participant A in our current example this assignment would be that in (14.)

(14) $r = \begin{bmatrix}\text{player1=SELF} \\ \text{player2=OTHER}\end{bmatrix}$

In (15) we see the application of $f_{inst\_T_{SuggestionGame}}$ to r.

(15) $f_{inst\_T_{SuggestionGame}}(r) =$
$\left[\text{e} : \text{suggest}(\begin{bmatrix}\text{player1=SELF} \\ \text{player2=OTHER}\end{bmatrix}.\text{player1})\right]^\frown$
$\left[\text{e} : \text{motivate}(\begin{bmatrix}\text{player1=SELF} \\ \text{player2=OTHER}\end{bmatrix}.\text{player1})\right]^{\leqslant 1\frown}$
$\left[\text{e} : \text{respond}(\begin{bmatrix}\text{player1=SELF} \\ \text{player2=OTHER}\end{bmatrix}.\text{player2})\right] =$
$[\text{e} : \text{suggest(SELF)}]^\frown[\text{e} : \text{motivate(SELF)}]^{\leqslant 1\frown}$
$[\text{e} : \text{respond(OTHER)}]$

The instantiated suggestion game would in this situation thus be $T_{SuggestionGameInst}$, as seen in (16):

(16) $T_{SuggestionGameInst} = [\text{e} : \text{suggest(SELF)}]^\frown$
$[\text{e} : \text{motivate(SELF)}]^{\leqslant 1\frown}[\text{e} : \text{respond(OTHER)}]$

4.3 Updating the agenda

An important aspect of the notion of conversational game is that players (conversational participants), by identifying an utterance as being part of a particular game, get an idea of which moves are likely to follow and what part they should expect to play over the next few turns of the dialogue. In this sense conversational games may be seen as engines driving dialogues forward. Once a game is loaded onto the gameboard and roles are assigned to individuals in the context, an agent involved in a conversation can at any stage of the game look at her gameboard

and know what options are available if she wants to keep playing the game. Before the update of the agenda, agent A – if playing the suggestion game – has on her private games the instantiated game $T_{SuggestionGameInst}$ which we see in (17).

Now, we want an update rule that would load the first available move of the game which is to be carried out by SELF, onto the agenda. We have a set of rules pertaining to the suggestion game that govern the dynamics of the agenda, which is inherent in the suggestion game in (16). The agenda is part of the 'private'-field of an agent's DGB, and is represented as a record type (move type). Each move type has a label 'e' paired with one of a set of speech act types, e.g. *Suggest*.There are a number of constraints on such move types having to do with the roles of the agents involved in dialogue, c_{actor}. Further, there is a label 'ctnt' for content, which – after the first update of the agenda – will not yet be associated with a specified content.

The first rule to be employed of the rules of the suggestion game is a "starting rule" in (17), stating that if a player has an empty agenda and a suggestion game on his private DGB, he may push a suggestion onto the agenda. We refer to this rule as $f_{update\_agenda\_suggestion}$.

(17) $f_{update\_agenda\_suggestion} =$
$\lambda r{:}\left[\text{pr} : \left[\begin{array}{l}\text{agenda} = [\ ] : \text{list}(\textit{RecType}) \\ \text{games} = [T_{SuggestionGameInst}] : \text{list}(T_{Game})\end{array}\right]\right].$
$\left[\text{pr} : \left[\text{agenda} = [\ \left[\begin{array}{l}\text{e} : \text{suggest(SELF)} \\ \text{ctnt} : \textit{RecType}\end{array}\right]\ ] : \text{list}(\textit{RecType})\right]\right]$

The content of the move type that ends up on the agenda is unspecified. $f_{update\_agenda\_suggestion}$ is applied to a record of the type in (18):

(18) $\left[\begin{array}{l}\text{pr} : \left[\begin{array}{l}\text{agenda} = [\ ] : \text{list}(\textit{RecType}) \\ \text{games} = [T_{SuggestionGameInst}] : \text{list}(T_{Game})\end{array}\right] \\ \text{sh} : \left[\text{project} = [\ [\text{e} : \text{decide}(\{A, B\}, \text{data})]\ ] : \text{list}(\textit{RecType})\right]\end{array}\right]$

We apply the function in (17) to the current information state of the type in (18), and asymmetrically merge the current state type with the result of function application. In (19) we see the type of A's information state after the rule has been applied.

(19) $\left[\begin{array}{l}\text{pr} : \left[\begin{array}{l}\text{agenda} = [\ \left[\begin{array}{l}\text{e} : \text{suggest(SELF)} \\ \text{ctnt} : \textit{RecType}\end{array}\right]\ ] : \text{list}(\textit{RecType}) \\ \text{games} = [T_{SuggestionGameInst}] : \text{list}(T_{Game})\ )\end{array}\right] \\ \text{sh} : \left[\text{project} = [\ [\text{e} : \text{decide}(\{A, B\}, \text{data})]\ ] : \text{list}(\textit{RecType})\right]\end{array}\right]$

The next update rule provided by the conversational game (although this rule is actually general and applicable to any conversational game) is a rule saying that if we have an item on the agenda which is to be performed by SELF and whose content is specified, that is the label 'ctnt' has one specific value ([ctnt=T:*RecType*]), then the agent is allowed to make that move and push the next move onto the agenda. However, at the moment the item on the agenda is not specified in terms of content – the label is just typed *RecType* (ctnt:*RecType*). In order to add a content specific move to the agenda, the agent needs to search her resources for relevant facts and ways of reasoning about the situation and the project at hand.

5 Conclusions

In this paper we consider different approaches to categorisation of interaction according to contextual factors, such as activity type and communicative project. We also consider the micro-level of organisations and present a model of how particular interactions play out within social activities and communicative projects of certain types. Our model can be integrated in a general formal model of dialogue such as Ginzburg (2012). We recognise three categories by which to define interactions – (communicative) activity, (communicative) project and conversational game, which we argue are associated with different aspects of the interaction and to some extent interdependent. These categories are also linked in a principled way to particular fields on the DGB. We illustrated by means of two examples of group decision making how the conversational game a project is associated with can be realised in different ways depending on the activity in which it is embedded. We may think of this as the possibilities afforded by the project being modified by the activity. We showed how the process of identifying a conversational game based on a project at hand, taking on or identifying one's role in the game and carrying out the appropriate moves can be modelled using DGBs modelled in TTR.

Our approach has the advantage that roles in a conversational game are analysed independently from the participants that carry them out in a particular interaction. This enables us to model for example anticipation of contributions by others and co-created utterances.

Acknowledgments

The research in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Cen-

tre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg and ERC Starting Grant DivCon: Divergence and Convergence in Dialogue: The Dynamic Management of Mismatches (101077927)

References

Jens Allwood. 2000. An activity based approach to pragmatics. In Harry C. Bunt & William Black, editor, *Abduction, belief, and context in dialogue: studies in computational pragmatics*, pages 47–78. John Benjamins, Amsterdam.

Kathryn A Baker. 2007. Organizational communication. *Management Benchmark Study*, 1(1):1–3.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Brill, Leiden, The Netherlands.

Andrew D Brown and Ken Starkey. 1994. The effect of organizational culture on communication and information. *Journal of Management studies*, 31(6):807–828.

Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005. *The Dynamics of Language*. Elsevier, Oxford.

Lauri Henrik Carlson. 1982. *Dialogue games: An approach to discourse analysis*. Ph.D. thesis, Massachusetts Institute of Technology.

Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington DC.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Robin Cooper. 2014. How to do things with types. In *Joint proceedings of the second workshop on Natural Language and Computer Science (NLCS 2014) & 1st international workshop on Natural Language Services for Reasoners (NLSR 2014) July*, pages 149–158.

Robin Cooper. 2023. *Type theory and language - From perception to linguistic communication*. Oxford University Press, Oxford, UK.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, pages 375–407. Wiley Blackwell, Oxford.

Arash Eshghi and Oliver Lemon. 2014. How domain-general can we be? Learning incremental dialogue systems without dialogue acts. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, pages 53–61. SEMDIAL.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the Tilburg International Workshop on Computational Semantics*. ITK Tilburg.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. In *Logic, language and computation*, volume 1. CSLI publications, Stanford, CA.

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

Jonathan Ginzburg. 2016. The semantics of dialogue.

Christine Howes and Arash Eshghi. 2021. Feedback relevance spaces: Interactional constraints on processing contexts in Dynamic Syntax. *Journal of Logic, Language and Information*, 30(2):331–362.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell, Oxford.

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

Staffan Larsson and David Traum. 2000. Information state and dialogue management in trindi dialogue move engine tool kit. *Natural Language Engineering*, 6:323–240.

James A. Levin and James A. Moore. 1977. Dialogue-games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420.

David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1):339–359.

Per Linell. 2009. *Rethinking Language, Mind, and World Dialogically: Interactional and Contextual Theories of Human Sense-Making*. Advances in Cultural Psychology: Constructing Human Development. Information Age Publishing.

Andy Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 383–392. IEEE.

Norman Malcolm. 1954. Wittgenstein's philosophical investigations. *The Philosophical Review*, 63(4):530–559.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In Jaakko Hintikka, Julius Moravcsik, and Patrik Suppes, editors, *Approaches to Natural Language*. D. Reidel, Dordrecht.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics*, volume 9 of *Syntax and Semantics*. New York Academic Press, New York, NY.

Ludwig Wittgenstein. 1953. *Philosophical investigations. Philosophische Untersuchungen.* Macmillan, London. Translated by G. E. M. Anscombe.

JoAnne Yates and Wanda J Orlikowski. 1992. Genres of organizational communication: A structurational approach to studying communication and media. *Academy of management review*, 17(2):299–326.

How do Encoder-only LMs Predict Closeness and Respect from Thai Conversations?

Pakawat Nakwijit[1]
[1]Queen Mary University of London
{p.nakwijit, m.purver}@qmul.ac.uk

Attapol T. Rutherford[2]
[2] Chulalongkorn University
Bangkok, Thailand
attapol.t@chula.ac.th

Matthew Purver[1,3]
[3]Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This study explores how encoder-only Language Models (LMs) recognize social relationships from textual data, examining both the models' behaviour and structure. Behaviourally, we analyze word importance, determined by SHAP values, to identify which lexical features—such as pronouns, sentence-final particles, and spelling variations—most influence the model's predictions in different conversational settings. Our findings confirm the use of these lexical features in the model's predictions albeit with varying degrees of contribution. We also validate our results by demonstrating a significant correlation between SHAP values and human evaluations. Structurally, we explore the impact of spelling variations on the structure of the encoder-generated word embeddings in social dimensions; closeness and respect. Using our projection approach, we observe a shift along both social dimensions when spelling variations are introduced in pronouns. Overall, this study sheds light on the mechanisms underlying the encoder model's social relationship recognition and contributes to verifying the alignment between the lexical features used by the model and human intuition.

1 Introduction

Our communication style, including word choice and tone, plays a crucial role in expressing our social identity and relationships, such as closeness and respect (Halliday, 1978; Poynton, 1991). Recognizing these social cues is, however, highly contextual and difficult to identify using traditional methods, particularly in Thai, a language that places a strong emphasis on social harmony and linguistic propriety (Knutson et al., 2003).

The advent of powerful architectures like the Transformer (Vaswani et al., 2017) and its derivatives, such as BERT (Devlin et al., 2019), initiated a new era, achieving remarkable performance across various NLP tasks, including social relationship recognition. However, their complex "black-box" nature renders their inner workings opaque, posing challenges for model interpretation and potentially leading to the generation of harmful content or hallucinations (Weidinger et al., 2021). Therefore, developing explainability mechanisms is critical to elucidate how these models operate enabling users to understand the rationale behind predictions or generated text, fostering trust, accountability, and responsible deployment across various NLP applications (Zhao et al., 2024; Doshi-Velez and Kim, 2017).

This study aims to address these challenges by developing a model proficient in recognizing closeness and respect using encoder-only Language Models (LMs), while simultaneously illuminating the underlying reasoning processes of these models through behaviour and structure aspects of the model. Firstly, we investigated word importance, estimated by SHAP value, to observe what lexical features (including pronouns, sentence-final particles and spelling variation) contribute the most to the model's predictions. We compared them across different conversational settings (private/public conversations, self-reported/perceived labels). In the end, we can confirm that all three lexical features contribute to the model's predictions. It, however, contributes to a different degree in different settings. For instance, first-person pronouns emerge as the primary contributor to the model's predicted closeness across all conversational contexts, surpassing other pronoun types. Conversely, singular pronouns only contribute to perceived closeness. Similarly, words with morphophonemic variation only influence predicted respect within private conversations.

Secondly, we explored the structure of the encoder-generated word embeddings in a social context by projecting the model's word embeddings onto dimensions representing closeness and respect. We presented our work on the investigation of how the introduction of spelling variations affects the

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

model's embeddings. Our findings demonstrate that introducing spelling variations in pronouns does not alter the overall shape of the projected distribution of word embeddings along the dimensions of closeness and respect. However, there is a notable shift towards increased closeness and decreased respect, as confirmed by the Mann-Whitney U test on the mean values. This underscores the model's sensitivity to linguistic nuances in shaping social perceptions.

2 Related Works

In this section, we review various explanation techniques tailored for LMs, categorizing them into two subsections based on their targeted facets of explainability. The first subsection delves into methods designed to provide an explanation from input features to determine the importance of each input token, for a given prediction (Behavioural). The second subsection explores methods that delve into the internal representation of LMs, seeking to discern its correlations with linguistic features (Structural).

2.1 Behavioural Analysis

Behavioural analysis often relies on strategically manipulating model inputs to observe their resulting behaviour. This approach leverages the inherent explanatory power of input features in NLP, where inputs directly correspond to human-interpretable elements like words, sub-words, or characters. By identifying the most influential words, researchers can gain valuable insights into the model's internal decision-making processes.

One prominent approach is Local Interpretable Model-Agnostic Explanations (LIME) by Ribeiro et al. (2016). LIME approximates the behaviour of complex models using a simple model trained locally around specific data points. To provide an explanation for an individual data point, a model, often a linear model due to its simplicity, is trained on data sampled locally around that specific instance. This localized training aims to approximate the behaviour of the original complex model within this restricted region of the feature space. This allows for explanations tailored to an individual instance. The authors demonstrated that explanations generated using LIME can accurately reflect the underlying behaviour of the model. However, LIME's explanatory power is limited to individual instances (local explanations). Additionally, Lundberg and Lee (2017) also highlighted potential shortcomings in LIME, including violations of local accuracy and consistency properties. These limitations can lead to counterintuitive explanations in certain scenarios.

Another method, SHapley Additive exPlanations (SHAP) by Lundberg and Lee (2017), built upon the well-established mathematical concept of Shapley values (Shapley, 1952). SHAP treats input features as contributors to a prediction outcome in a cooperative game. It assigns each feature subset a value reflecting its contribution. This approach offers strong expressiveness, particularly for LMs. Unlike LIME, Lundberg and Lee (2017) demonstrated that it satisfies all desirable properties including local accuracy, missingness and consistency. Additionally, SHAP also allows for global interpretations by averaging its values for each feature across a dataset which have been shown to be consistent with the local explanations (Molnar, 2018; Covert et al., 2020). Notably, Wu et al. (2021) exemplified a successful SHAP application in dataset construction by using it as a guide for their experts in designing counterfactual examples. Hayati et al. (2021) used SHAP to investigate how a model predicts linguistic styles by contrasting lexicons highlighted by humans with those exhibiting high SHAP scores. In this work, we employed SHAP in a comparable manner by aggregating importance scores across three lexical features. These scores are then used to evaluate the significance of each lexical feature across different conversational settings and to assess their alignment with human-annotated scores.

2.2 Structural Analysis

Structural analysis aims to observe linguistic knowledge embedded within the internal representations of the model. It is commonly achieved through probing techniques, which use a simple model, often a logistic regression, to determine whether a target linguistic structure can be predicted from the learned representation. Mohebbi et al. (2021) successfully demonstrated that representations in models like BERT encapsulate linguistically relevant information, encompassing both syntactic and semantic aspects. Their findings also suggest that lower layers predominantly capture word-level syntax, while higher layers excel at encoding sentence-level syntax and semantic knowledge, akin to human language processing. However, Belinkov (2022) argued that conclusions drawn from probing techniques may not always be as robust

as they appear. With sufficiently high-dimensional embeddings, complex probes, and large auxiliary datasets, the probes can seemingly learn to extract any information from any embeddings.

An alternative approach to understanding the model's structure involves examining how the model encodes information within its representations. Torroba Hennigen et al. (2020) extended probing techniques by assessing probe performance on different subsets of dimensions to locate the amount of linguistic information encoded within distinct subsets of dimensions. Their research revealed that many morphosyntactic features are reliably encoded by only a small number of neurons. Kozlowski et al. (2019) adopted a different perspective by projecting embeddings to provide visual explanations. They leveraged the principle that word embeddings should be able to capture semantics as arithmetic relationships between embeddings in a high-dimensional space. Their work illustrated that dimensions induced by pre-trained embeddings correspond to dimensions of cultural meaning (e.g. rich/poor). The projection of words onto these dimensions reflects widely shared stereotypes of social class. For instance, words like "golf" and "tennis" are associated with rich individuals, while "boxing" is linked to lower socioeconomic status. In this study, we adopt a similar approach to investigate how the introduction of spelling variations influences the model's embeddings. This analysis aims to reaffirm that lexical information is effectively represented within the model.

3 Conversation Corpus

The corpus utilized in this study was collected from Nakwijit et al. (2024). The corpus comprises a diverse collection of Thai conversational texts sourced from two sources; 1,234 private conversations specifically curated from their study and 2,496 public conversations from X (formerly Twitter). The corpus is organised into two tasks, including closeness and respect, with three conversational settings, including

- Setting 1: Private Conversations with Self-Reported Relationships (Private-Self)
- Setting 2: Public Conversations with Perceived Relationships (Public-Perceived)
- Setting 3: Private Conversations with Perceived Relationships (Private-Perceived)

They also provided a set of lexicons from 15 lexical features. In this study, we only focus on three lexical features including pronouns, sentence-final particles and spelling variations. Throughout the experiments, we linearized the utterances in a conversation and marked the beginning of each utterance with *[sys]* or *[usr]* to indicate those who initiated the conversation and the respondent. More detailed descriptions of the corpus and lexical features can be found in Appendix A and Appendix B.

4 Social Relationship Models

In this section, we outline our experiments concerning the construction of a social relationship model. Subsequently, the best model according to the F1 score from each setting was selected for further analysis in the subsequent sections of the study.

4.1 Experimental Setup

Before model training and analysis, the corpus underwent standard preprocessing procedures, converting text to lowercase, replacing repeated characters with a *[REP]* token, and tokenizing the text using PyThaiNLP's tokenizer (Phatthiyaphaibun et al., 2023). Following the original paper, we confined our target labels to three levels of closeness and respect, discarding the minority. Labels for closeness and respect were then normalized to a continuous range between -1 and 1, where -1 and 1 denote the lowest and highest degrees of closeness or respect in that setting.

Lastly, we randomly shuffled the corpus and partitioned it into 80% for training, 10% for validation, and 10% for testing. Standard machine learning protocols were followed: training was conducted on the training set, hyperparameters were tuned on the validation set for optimal F1-score, and final metrics were reported based on the test set. The final predictions were discretized back into three labels using thresholds of -0.5 and 0.5 accordingly.

4.2 Selected Models

We experimented with 6 models; 3 simple baselines, and 3 LMs, which are as follows:

Majority-class Model: This model serves as the simplest approach by predicting solely the majority class. It sets a minimum baseline performance that accounts for label imbalances.

Naive Bayes Classifier: It is a probabilistic model based on Bayes' theorem. It operates under the naive assumption of conditional indepen-

dence between individual words, given the class label. This simplification makes it suitable as a baseline model when it is constrained to employ only surface-level lexical information. In essence, it gauges the extent to which closeness and respect levels can be predicted solely based on observable lexicons.

Logistic Regression: An Ordinary Least Squares regression (OLS) model was employed, utilizing 15 linguistic features as predictors such as the number of unique words, number of turns, number of long words, and average number of words per utterance. This model served as a baseline to gauge the predictive power conferred solely by the linguistic features of the conversation.

Fine-tuned XLM-R: It is a multilingual language model designed for understanding and generating text across 100 languages (Conneau et al., 2020).

Fine-tuned WangChanBERTa: It is a monolingual language model trained on a Thai corpus (Lowphansirikul et al., 2021).

Fine-tuned PhayaThaiBERT: It is an extended version of WangChanBERTa via vocabulary transfer to compensate for a lack of foreign vocabulary and orthographic variations in the previous models (Sriwirote et al., 2023).

All three encoders were selected for their status as competitive models, which can leverage pre-trained common-sense knowledge, surface-level lexical information, and broader contextual information. Although they all utilize the RoBERTa architecture (Liu et al., 2019), they vary in terms of their multilingual capabilities (for XLM-R) versus monolingual capabilities (for WangChanBERTa and PhayaThaiBERT) and in the size of their vocabularies, ranging from small (25k words in WangChanBERTa) to large (250k words in PhayaThaiBERT).

We followed the standard fine-tuning practice on WangchanBERTa. The fine-tuning parameters for the model were set as follows:

- Learning rate: 2e-5
- Optimiser: Adam
- Weight decay rate: 0.01
- Number of epochs: 20
- Batch size: 16
- Input max length: 128
- Select the best model with F1 score

Each model was trained five times and reported the average results according to F1 score. The numbers are presented in Table 1.

4.3 Model Performance

A noticeable improvement emerges when additional information is incorporated into the model. The Naive Bayes model, with direct access to surface-level information such as word frequency in a conversation, demonstrates decent performance, achieving F1 scores ranging from 0.43 to 0.56 for closeness and 0.47 to 0.67 for respect—constituting 82% to 90% of the best model's performance. This finding aligns with previous research, suggesting that lexicons alone can serve effectively as social markers (Schwartz et al., 2013). Conversely, linear regression on lexical features yields slightly inferior results, ranging from 0.33 to 0.54 for closeness and 0.31 to 0.46 for respect. Our best model, fine-tuned PhayaThaiBERT, effectively predicts closeness labels with F1 scores ranging from 0.50 to 0.67 and respect labels from 0.43 to 0.75 closely followed by fine-tuned WangChanBERTa and XLM-R.

All LMs surpassed other baselines in nearly all settings, highlighting the importance of pre-trained knowledge, such as contextual representations and common ground knowledge. However, it was evident that XLM-R, as a multilingual model, performed considerably worse than the other two monolingual models. Additionally, vocabulary expansion notably enhanced PhayaThaiBERT's performance over WangChanBERTa in 5 out of 6 settings.

Upon closer examination, all models struggled in two specific settings: *Closeness Setting2: Public-Perceived* and *Respect Setting1: Private-Self*, with F1 scores of only 0.50 and 0.43, respectively. One possible reason for this may be unclear guidelines during data collection, as suggested by the notably low validation agreement observed in *Respect Setting1: Private-Self* (Nakwijit et al., 2024). However, this does not fully explain the models' relative success in other settings, given that the same groups of annotators annotated all labels. Another potential explanation could be that while some settings exhibit consistent and clear linguistic patterns, the constructs of self-perceived respect and perceived closeness are inherently more complex and/or subtle than previously understood. Nevertheless, investigating this matter further falls outside the scope of our study.

5 Understanding Model's Behaviour Through SHAP

In this section, our objective is to ascertain the extent to which each lexicon and lexicon type con-

Model	Task1: Closeness			Task2: Respect		
	Setting 1 Private-Self	Setting 2 Public-Perceived	Setting 3 Private-Perceived	Setting 1 Private-Self	Setting 2 Public-Perceived	Setting 3 Private-Perceived
Baseline						
Majority-class Baseline	0.155	0.206	0.401	0.179	0.276	0.308
Naive Bayes Classifier	0.563	0.435	0.542	0.470	0.678	0.535
Logistic Regression	0.400	0.327	0.542	0.314	0.444	0.463
LMs						
XLM-R	0.604	0.420	0.498	0.200	0.675	0.432
WangChanBERTa	0.657	0.490	0.639	0.313	0.748	**0.761**
PhayaThaiBERT	**0.666**	**0.496**	**0.657**	**0.431**	**0.750**	0.712

Table 1: The f1 performance metrics of our social relationship models in the closeness and respect tasks across three conversational settings

tributes to the model's predictions.

5.1 Methodology

In our analysis, SHAP values were computed using our best model (fine-tuned PhayaThaiBERT). The contribution score for each word in the conversations was calculated. These values were then grouped by their respective lexical features, converted into absolute values, averaged, and subsequently reported in Table 2 and Table 3 for closeness and respect tasks.

5.2 Results and Discussion

Based on the SHAP values, pronouns emerge as a pivotal contributor to the prediction process, exhibiting average SHAP values of 1.13, 4.52 and 1.04 per token for closeness tasks and 1.88, 2.93 and 1.71 per token for respect tasks. These values surpass the baseline derived from random tokens in five of six settings. The numbers also suggest that pronouns with different morphosyntactic features, such as grammatical person and numbers, contribute differently to closeness tasks. Specifically, first-person pronouns contribute in all settings while second-person pronouns are more significant in settings involving private conversations, and third-person pronouns are mainly relevant only in perceived closeness in private conversations. Singular pronouns solely contribute to perceived closeness, while plural pronouns do not exert more influence on closeness than random tokens. Interestingly, pronouns in spelling variation form, which are typically considered as noise, make substantial contributions to predictions in perceived closeness. These findings are even more pronounced in respect tasks, where second-person, singular, and non-standard-written pronouns consistently outperform the random baseline across all settings.

Regarding sentence-ending particles, the findings highlight disparities between two particle subtypes: socially-rated and non-socially-rated. The SHAP values clearly reveal that the model relies on socially-related particles as cues for closeness, while not doing so for the latter subtype. Furthermore, we observe that particles with non-standard spelling influence the model's predictions of closeness and respect more than the random baseline across all three settings, with SHAP values of 1.33, 7.63, and 1.11 for closeness tasks, and 1.54, 2.14, and 0.98 for respect tasks. However, these values are still lower than the SHAP values of pronouns and pronouns with non-standard spellings in four out of six settings.

Spelling variations, on the other hand, do not exhibit high SHAP values across all settings. Its contributions from subtypes of the variations, however, become more pronounced. Morphophonemic variations, for instance, demonstrate SHAP values per token of 1.26, 5.37 and 0.95 in the closeness tasks, and 1.52. 1.90 and 0.86 in the respect tasks. Like pronouns, these values exceed the random baseline in 5 out of 6 settings. Importantly, in those 5 settings, its total contribution even surpasses that of pronouns and sentence-final particles by a considerable margin due to the higher frequency of spelling variations compared to pronouns and particles. This finding underscores the important role of spelling variations, especially in public conversations.

Our observations align closely with the findings obtained from the original work which presents the corpus and analyses it through statistical analysis (Nakwijit et al., 2024). This correspondence may provide evidence that the model leverages analogous linguistic cues to predict the target labels. We,

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.08	125.36	4.07	147.01	0.85	97.91
Pronoun						
All pronoun	1.13	4.05	4.52	9.47	1.60	5.65
» 1st person pronoun	1.25	2.85	5.15	7.73	1.14	2.56
» 2nd person pronoun	1.30	3.29	4.33	7.68	2.04	5.11
» 3rd person pronoun	0.71	1.31	3.47	5.61	1.71	3.14
» Singular pronoun	1.13	4.04	4.52	9.40	1.60	5.65
» Plural pronoun	1.07	1.07	4.30	5.73	0.49	0.49
» Pronoun in non-standard spelling	0.74	1.58	7.62	10.02	1.23	2.44
Sentence-final Particles						
All particles	1.75	8.81	4.16	7.54	0.93	4.68
» Socially-related particles	3.24	10.03	5.08	7.27	1.31	4.08
» Non-socially-related particles	0.85	2.97	3.47	5.45	0.69	2.43
» Particle in non-standard spelling	1.33	1.86	7.63	8.41	1.11	1.56
Spelling Variation						
All spelling variation	1.10	14.48	4.39	19.46	0.86	11.28
» Common misspelt words	0.83	1.29	3.80	5.24	0.80	1.24
» Morphophonemic variation	1.26	10.49	5.37	15.10	0.95	7.91
» Simplified variation	0.90	5.81	3.63	10.79	0.74	4.77
» Repeated characters	0.85	1.82	3.41	4.47	0.54	1.15

Table 2: The average of absolute SHAP values of three lexical features in **closeness tasks** across 3 conversational settings from **fine-tuned PhayaThaiBERT**. The values highlighted in grey denote values exceeding the SHAP values of their respective random baseline

however, obtained different results when applying the same method to fine-tuned WangChanBERTa and XLM-R. The detailed SHAP values for these two models are reported in appendix E.

5.3 Validation with Human Scores

To assess the validity of the explanation, we asked the participation of 13 native Thai-speaking teenagers aged between 18 and 20 years. Each participant was presented with a set of 1000 words selected based on their highest SHAP values and was asked to select one level of closeness/respect that was most closely associated with the given words. These relationship levels were then quantified using numerical values ranging from -2 to 2. Subsequently, we identified the most frequently selected levels among the participants as the final score corresponding to each word. Finally, we calculated the correlation between the human-assigned score and its SHAP value. The results are presented in Table 4. It is important to note that we excluded *Setting 1: Private-Self* because the principle of self-reported labels does not align with our validation methodology.

The findings are presented in Table 4. Our results reveal that, overall, there exists a weak correlation (r=0.20-0.32) between SHAP values and human scores in all tasks, except the perceived closeness in public conversation *(Setting 2: Public-Perceived)* which aligned with the low f1 in the same task found in Table 1.

Notably, pronouns demonstrate a consistent correlation across all settings, in contrast to sentence-final particles and spelling variations, which do not. Specifically, sentence-final particles only show a correlation in the respect tasks within public conversations, while spelling variations correlate in all settings except that task. The absence of correlation in certain instances remains unclear; this may be attributed to insufficient data or potential discrepancies between human perceptions and machine interpretations.

6 Effect of Spelling Variation on Embedding Structure

To build a further understanding of how the model represents social meaning, we adopt an analysis ap-

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Pertoken	Total	Pertoken	Total	Pertoken	Total
Reference						
Average per token	1.24	143.37	1.95	72.22	0.75	86.78
Pronoun						
All pronoun	1.88	6.75	2.93	6.77	1.71	6.27
» 1st person pronoun	1.74	3.98	1.90	2.98	1.62	3.78
» 2nd person pronoun	2.17	5.48	4.04	7.51	1.80	4.60
» 3rd person pronoun	1.88	3.49	1.95	3.48	0.78	1.44
» Singular pronoun	1.88	6.74	2.95	6.78	1.72	6.27
» Plural pronoun	1.14	1.14	1.09	1.34	0.26	0.26
» Pronoun in non-standard spelling	1.81	3.77	2.88	4.15	1.73	3.86
Sentence-final Particles						
All particles	1.16	5.89	1.87	3.60	0.65	3.27
» Socially-related particles	1.35	4.19	2.85	4.12	0.74	2.29
» Non-socially-related particles	1.05	3.69	1.23	2.11	0.60	2.09
» Particle in non-standard spelling	1.54	2.16	2.14	2.52	0.98	1.37
Spelling Variation						
All spelling variation	1.39	18.31	1.71	7.84	0.77	10.10
» Common misspelt words	1.37	2.13	1.74	2.40	0.88	1.36
» Morphophonemic variation	1.52	12.68	1.90	5.62	0.86	7.16
» Simplified variation	1.21	7.84	1.45	4.50	0.65	4.19
» Repeated characters	0.92	1.97	0.72	0.95	0.88	1.88

Table 3: The average of absolute SHAP values of three lexical features in **respect tasks** across three conversational settings from **fine-tuned PhayaThaiBERT**. The values highlighted in grey denote values exceeding the SHAP values of their respective random baseline

Lexical Features	Closeness		Respect	
	Setting 2 Private	Setting 3 Public	Setting 2 Private	Setting 3 Public
Overall	0.059	0.203*	0.315*	0.240
Pronoun	0.238	0.349*	0.498	0.355
Sentence-final Particles	0.037	0.022	0.017	0.442*
Spelling Variation	0.182*	0.299*	0.215*	0.045

Table 4: The correlations between *PhayaThaiBERT*'s SHAP values and human scores for words from three lexical features and its association with closeness/respect. Values with a p-value less than 0.05 are indicated by an asterisk (*).

proach proposed by Kozlowski et al. (2019). The core idea is to observe how closeness/respect are encoded by the model and how the representation changes when there are changes in linguistic features which we presented by the introduction of spelling variations on pronouns.

6.1 Methodology

The analysis consists of 3 steps: calculating the social dimension, projecting word embeddings onto the dimension and observing the social orientation of the words.

Step 1: Calculating the Social Dimension

Each conversation was represented as the average of the hidden embeddings for each token from the last layer of the fine-tuned PhayaThaiBERT. To represent two extreme groups, the embeddings were separated into two opposite groups based on their annotated labels: *Intimate* and *Dislike* for closeness, and *Highly Respectful* and *Disrespectful* for respect. The embeddings were subsequently averaged, and the vector differences from each pair were utilized as social dimensions for closeness and respect, respectively.

Step 2: Projecting Word Embeddings

In this analysis, our focus was specifically on pronouns, given their notable outcomes thus far. We manually chose pronouns with spelling variants as an illustrative example of how the model changes its representation to align with spelling changes and their associated social meanings. The last hidden embeddings of the selected pronouns from all conversations, were projected onto the constructed dimension using cosine similarity.

Step 3: Observing the Social Orientation of the Words

Finally, we examined the social orientation of pronouns by plotting the distribution of projected values. The resulting plots are presented in Figure 1. Additionally, we conducted the Mann-Whitney U test over the mean value to ascertain whether the values in one group are different from those in the other group and reported the corresponding p-values.

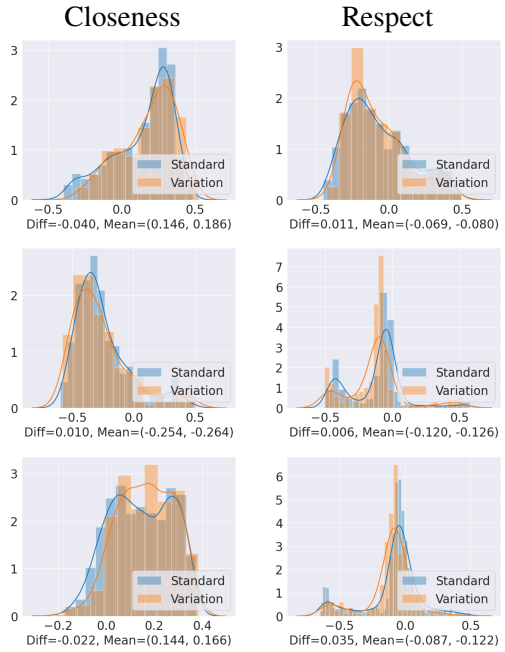


Figure 1: The distribution of social orientation values (cosine similarity) for word embeddings of pronouns and their spelling variations, projected onto dimensions representing closeness and respect from three settings: Private-Self (top), Public-Perceived (middle) and Private-Perceived (bottom)

6.2 Results and Discussion

The Figure 1 shows that, in general, the model represents a pronoun with an embedding that leans toward a closer relationship in private conversation with the average social orientation values of 0.146, and 0.144 for *Setting 1: Private-Self* and *Setting 3: Private-Perceived*. While leaning against a closer relationship in the public one with the average social orientation value of -0.254 for *Setting 2: Public-Perceived*. It, however, consistently leans toward disrespectful relationships across all three settings with the average social orientation values of -0.069, -0.120 and -0.087 for *Setting 1: Private-Self*, *Setting 2: Public-Perceived* and *Setting 3: Private-Perceived* respectively.

Expectedly, our results also suggested that the model represents pronouns and their variants in a similar distribution shape. However, we observed a slight shift in the distribution. The introduction of spelling variation generally makes the model shift toward greater closeness and lesser respect with the differences in mean between the two groups being -0.040*, 0.010, and -0.022* in closeness tasks and 0.011, 0.006*, 0.035* in respect tasks where * indicates when it has p-value less than 0.05. This further confirms that the model can represent social nuance quite nicely.

7 Conclusion

In summary, this research provides valuable insights into the mechanisms guiding encoder-only language models in identifying social relationships from text data. Through a series examination of both behavioural and structural aspects, we illustrated the critical roles played by three lexical features, including pronouns, sentence-final particles, and spelling variation, in shaping model predictions across three conversational settings. By using SHAP, we uncovered nuanced relationships between these lexical features and the behaviour of model predictions. For instance, pronouns of different grammatical persons and numbers contribute differently to tasks involving closeness: first-person pronouns are influential across all settings; second-person pronouns are particularly significant in private conversations; and third-person pronouns mainly affect the perception of closeness in private contexts. Additionally, our results emphasize the importance of spelling variations, often overlooked as linguistic noise, including non-standard forms of pronouns and sentence-final particles, as well as other words written in morphophonemic variations. Lastly, our embedding projection study shows that the models typically represent pronouns as signals of increased closeness and decreased respect. Its embeddings also retain a consistent distribution pattern even when spelling variations are introduced, albeit with a minor shift towards more closeness and less respect suggesting that spelling variation functions as an intensifier of the social meaning. Collectively, these results affirm that encoder-only language models effectively encode and use linguistic information, especially sociolinguistic clues in the lexical features, to a considerable extent.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Joseph R Cooke et al. 1989. Thai sentence particles: forms, meanings and formal-semantic variations. In *Papers in Southeast Asian Linguistics No. 12: Thai sentence particles and other topics*. Pacific Linguistics.

Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gráinne M Fitzsimons and Aaron C Kay. 2004. Language and interpersonal cognition: Causal effects of variations in pronoun usage on perceptions of closeness. *Personality and Social Psychology Bulletin*, 30(5):547–557.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Michael Alexander Kirkwood Halliday. 1978. *Language as social semiotic: The social interpretation of language and meaning*, volume 42. Edward Arnold London.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuphaphann Hoonchamlong. 1992. Some observations on phom and dichan: Male and female 1st person pronouns in thai. *Papers on Tai Languages, Linguistics, and Literatures: In Honor of William J. Gedney on his 77th Birthday*, 16:195–213.

Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.

Thomas J Knutson, Rosechongporn Komolsevin, Pat Chatiketu, and Val R Smith. 2003. A cross-cultural comparison of thai and us american rhetorical sensitivity: Implications for intercultural communication effectiveness. *International Journal of intercultural relations*, 27(1):63–78.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of BERT token representations to explain sentence probing results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christoph Molnar. 2018. A guide for making black box models explainable. *URL: https://christophm. github. io/interpretable-ml-book*, 2(3):10.

Pakawat Nakwijit and Matthew Purver. 2022. Misspelling semantics in Thai. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 227–236, Marseille, France. European Language Resources Association.

Pakawat Nakwijit, Attapol T. Rutherford, and Matthew Purver. 2024. The language of closeness and respect in thai conversations: An analysis of lexical features and spelling variations. Unpublished.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita

Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. Pythainlp: Thai natural language processing in python. *arXiv preprint arXiv:2312.04649*.

Cate McKean Poynton. 1991. *Address and the semiotics of social relations: A systemic-functional account of address forms and practices in Australian English*. phdthesis, University of Sydney.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Lloyd S. Shapley. 1952. *A Value for N-Person Games*. RAND Corporation.

Panyut Sriwirote, Jalinee Thapiang, Vasan Timtong, and Attapol T Rutherford. 2023. Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords. *arXiv preprint arXiv:2311.12475*.

Hanne Surkyn, Reinhild Vandekerckhove, and Dominiek Sandra. 2021. Social media data as a naturalistic test bed for studying sociolinguistic and psycholinguistic patterns in verb spelling errors. In *of the 8th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-Corpora 2021)*, volume 559, page 90.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. 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 6707–6723, Online. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

A Conversation Corpus

The corpus was originally introduced by Nakwijit et al. (2024). It is designed to explore how lexical features interact with social relationships in private and public settings. The construction of the corpus is detailed in the following subsections.

A.1 Setting 1: Private-Self

The authors set up a messaging platform, and crowdsourced participants aged 18-30 to create a chat room and invite another participant for a conversation. Participants selected a seeding topic from the Switchboard corpus (Godfrey et al., 1992) and conversed on this topic for at least 30 turns. After the conversation, they privately filled out a form to identify their relationship in terms of closeness and respect, choosing from *Intimate*, *Close*, *Acquainted*, *Unfamiliar*, *Dislike*, and *Cannot describe* for closeness, and *Highly Respectful*, *Respectful*, *Normal*, *Disrespectful*, and *Cannot describe* for respect.

A.2 Setting 2: Public-Perceived

The authors collected tweets from X (formerly Twitter) based on 53 popular hashtags in 2022. Those tweets were filtered and selected with at least two replies. Each conversation was annotated by three recruited native Thai-speaking teenagers (16-18 years old), who assessed the degree of closeness and respect perceived in the conversation with the same set of labels presented in Setting 1. Each conversation was presented as a dialogue between an initiator (A) and a responder (B), withholding any identifying information about both individuals. Annotators were instructed to provide labels from the perspective of the responder (B). Only conversations with at least two annotators in agreement were retained; the rest were discarded.

A.3 Setting 3: Private-Perceived

The author re-annotated private conversations from Setting 1 by the annotators from Setting 2. The

same procedure and labeling scheme as in Setting 2 were applied during this re-annotation process.

B Lexical Features

Our analysis consider only three lexical features; pronouns, sentence-final particles and spelling variations. The selection of these features was guided by their prominence in sociolinguistic literature, particularly in relation to social factors such as gender, age, and social status in both English and/or Thai.

Pronoun was chosen as it is a well-studied lexical feature known for their social functionality across many languages (Hoonchamlong, 1992; Fitzsimons and Kay, 2004; Kacewicz et al., 2014). Their frequent use and significant role in communication make them a critical feature as a reference baseline.

Sentence-final particle was included because it represents a lesser-known social-related feature. These particles have limited studies due to their observation in a narrower range of languages, primarily East and Southeast Asian languages (Cooke et al., 1989).

Lastly, spelling variation was selected as it represents a recent linguistic pattern that has gained recognition for its potential semantic functions (Surkyn et al., 2021; Nakwijit and Purver, 2022). There are few studies on spelling variations, especially in Thai. Importantly, in this paper, spelling variation is specifically highlighted because of its increasing prevalence in modern conversations driven by the internet and social networks. By examining it, we aim to raise awareness of its importance in contemporary linguistic analysis.

C Social Relationship Models

Here is a detailed description of the input features for our models:

Naive Bayes Classifier: We used word count as input features, discarding terms with a frequency of less than five.

Logistic Regression: We used 15 lexical features as input features, including the number of unique words, number of Thai words, number of long words (more than 7 characters), number of out-of-vocabulary words, number of 1st person pronouns, number of 2nd person pronouns, number of 3rd person pronouns, number of pronouns in non-standard spellings, number of socially-related particles, number of non-socially-related particles, number of sentence-final particles in non-standard spellings, number of common misspelt words, number of morphophonemic variations, number of abbreviations, and number of repeated characters,

For each conversation, we examined each word and identified its lexical type using a dictionary-based approach. The dictionaries for each lexical type were provided by the authors of the corpus. We counted the number of words corresponding to each lexical feature. Finally, the values for each lexical feature were normalized by the total number of words in the conversation.

D Human Validation

In our validation in section 5.3, we intentionally recruited participants aged 18-20. This decision was made to closely match the age range of the participants in the original corpus.

We acknowledge that this decision introduces a bias, potentially affecting the interpretation of results, as language usage can vary across different age groups. However, this age group was our target population because they have grown up with text-only communication technology and are familiar with internet slang and variations, making them ideal candidates for validating our experiments.

During the annotation process, each word was presented without context. The annotators were asked the following question: "ตอบในมุมมองของคนที่ใช้คำๆนี้ในบทสนทนา ถ้าเห็นเขาใช้คำๆนี้แล้ว คิดว่าเขามีความสัมพันธ์อย่างไรกับคนที่เขากำลังพูดด้วย " (translation: Answer from the perspective of the person using this word in the conversation. When you see them using this word, what do you think their relationship is with the person they are speaking to?).

E SHAP Value from LMs

The tables below present the average of absolute SHAP values across all tokens for three lexical features (pronoun, sentence-final particles, spelling variation) in three conversational settings. Values highlighted in grey indicate those exceeding 10% of their respective random baselines, which are calculated from the SHAP values of 100 randomly selected tokens.

E.1 Fine-tuned XLM-R

E.1.1 Closeness

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.07	123.94	1.80	65.14	0.58	67.67
Pronoun						
All pronoun	0.80	3.64	1.29	3.46	0.25	1.13
» 1st person pronoun	0.78	2.14	1.21	2.20	0.23	0.63
» 2nd person pronoun	0.87	2.78	1.23	2.67	0.28	0.88
» 3rd person pronoun	0.49	1.02	1.02	1.92	0.20	0.41
» Singular pronoun	0.80	3.63	1.27	3.37	0.25	1.13
» Plural pronoun	0.23	0.23	2.33	3.00	0.11	0.11
» Pronoun in non-standard spelling	0.47	0.96	1.33	2.01	0.23	0.45
Sentence-final Particles						
All particles	2.98	22.07	1.86	4.27	3.39	25.04
» Socially-related particles	7.11	29.19	2.68	4.40	8.35	34.35
» Non-socially-related particles	0.60	2.98	1.29	2.29	0.51	2.53
» Particle in non-standard spelling	0.85	1.43	2.44	2.81	0.92	1.54
Spelling Variation						
All spelling variation	1.27	23.45	1.65	9.98	0.55	10.13
» Common misspelt words	0.96	1.69	1.50	2.19	0.20	0.35
» Morphophonemic variation	1.70	18.67	2.09	7.44	0.79	8.63
» Simplified variation	0.68	6.22	1.30	5.24	0.24	2.21
» Repeated characters	0.53	1.14	1.50	1.97	0.15	0.32

E.1.2 Respect

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	0.22	25.50	1.37	50.69	0.30	34.33
Pronoun						
All pronoun	0.18	0.83	2.07	6.11	0.16	0.74
» 1st person pronoun	0.15	0.43	1.68	3.16	0.16	0.46
» 2nd person pronoun	0.20	0.64	2.72	6.25	0.17	0.56
» 3rd person pronoun	0.16	0.33	0.77	1.59	0.14	0.30
» Singular pronoun	0.18	0.83	2.10	6.10	0.16	0.74
» Plural pronoun	0.15	0.15	0.64	0.84	0.08	0.08
» Pronoun in non-standard spelling	0.17	0.34	0.75	1.22	0.12	0.26
Sentence-final Particles						
All particles	0.52	3.86	0.96	2.29	0.21	1.53
» Socially-related particles	1.12	4.61	1.40	2.27	0.24	1.00
» Non-socially-related particles	0.17	0.86	0.70	1.36	0.19	0.92
» Particle in non-standard spelling	0.19	0.31	1.03	1.22	0.20	0.34
Spelling Variation						
All spelling variation	0.20	3.63	0.91	5.80	0.19	3.53
» Common misspelt words	0.22	0.39	1.01	1.45	0.19	0.33
» Morphophonemic variation	0.21	2.33	1.04	3.91	0.21	2.32
» Simplified variation	0.19	1.75	0.82	3.48	0.18	1.64
» Repeated characters	0.19	0.40	0.28	0.37	0.19	0.40

E.2 Fine-tuned WangChanBERTa

E.2.1 Closeness

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.34	156.00	2.91	105.40	1.17	135.70
Pronoun						
All pronoun	1.51	5.42	3.68	7.72	1.92	6.78
» 1st person pronoun	1.61	3.69	4.51	6.76	1.67	3.77
» 2nd person pronoun	1.91	4.83	3.76	6.66	2.41	6.05
» 3rd person pronoun	0.94	1.74	2.21	3.57	1.87	3.44
» Singular pronoun	1.52	5.42	3.72	7.73	1.92	6.77
» Plural pronoun	0.46	0.46	1.32	1.76	0.24	0.24
» Pronoun in non-standard spelling	0.90	1.92	6.02	7.91	1.72	3.43
Sentence-final Particles						
All particles	2.87	14.46	3.30	5.99	1.51	7.64
» Socially-related particles	5.24	16.26	3.43	4.91	2.58	8.02
» Non-socially-related particles	1.43	5.00	3.21	5.03	0.86	3.03
» Particle in non-standard spelling	1.98	2.77	7.20	7.94	1.08	1.51
Spelling Variation						
All spelling variation	1.39	18.25	3.36	14.89	1.08	14.23
» Common misspelt words	1.09	1.69	3.14	4.33	1.22	1.89
» Morphophonemic variation	1.64	13.66	4.21	11.83	1.19	9.90
» Simplified variation	1.08	7.00	2.69	7.98	0.87	5.67
» Repeated characters	0.64	1.37	3.39	4.44	0.41	0.88

E.2.2 Respect

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.49	173.20	2.16	80.16	0.46	53.21
Pronoun						
All pronoun	3.57	12.84	2.64	6.11	1.11	4.06
» 1st person pronoun	3.76	8.59	1.86	2.92	1.24	2.89
» 2nd person pronoun	4.17	10.54	3.25	6.05	1.00	2.57
» 3rd person pronoun	3.28	6.10	1.92	3.43	0.44	0.82
» Singular pronoun	3.59	12.85	2.66	6.11	1.11	4.06
» Plural pronoun	0.55	0.55	1.09	1.34	0.22	0.22
» Pronoun in non-standard spelling	3.62	7.52	2.00	2.89	1.07	2.39
Sentence-final Particles						
All particles	1.53	7.77	2.26	4.35	0.49	2.47
» Socially-related particles	2.02	6.27	3.16	4.58	0.69	2.14
» Non-socially-related particles	1.24	4.35	1.67	2.86	0.37	1.29
» Particle in non-standard spelling	1.79	2.52	2.45	2.87	0.39	0.55
Spelling Variation						
All spelling variation	1.91	25.11	1.93	8.86	0.46	6.01
» Common misspelt words	1.53	2.37	2.26	3.12	0.49	0.76
» Morphophonemic variation	2.29	19.08	2.16	6.40	0.51	4.27
» Simplified variation	1.45	9.36	1.64	5.07	0.38	2.44
» Repeated characters	0.66	1.42	1.28	1.69	0.18	0.39

Large Language Models as an active Bayesian filter: information acquisition and integration

Sabrina Patania, Emanuele Masiero, Luca Brini, Valentyn Piskovskyi and **Dimitri Ognibene**
University of Milan - Bicocca
dimitri.ognibene@unimib.it

Gregor Donabauer and **Udo Kruschwitz**
University of Regensburg

Abstract

This study investigates Large Language Models (LLMs) as dynamic Bayesian filters through question-asking experiments inspired by cognitive science. We analyse LLMs' inference errors and the evolution of uncertainty across models using repeated sampling.

Building on Bertolazzi et al. (2023), we trace LLM belief states during repeated queries, finding that entropy decreases with each interaction, signaling reduced uncertainty. However, issues like "resurrection" (reassigning probabilities to invalidated outcomes) and "Bayesian apocalypse" (probabilities approaching zero) reveal significant flaws. GPT-4o consistently outperforms GPT-3 in probabilistic reasoning. These results underscore the need for improved architectures for reliability in high-stakes contexts and suggest a link between token-level and task-level uncertainty dynamics that can be leveraged to enhance LLM performance.

1 Introduction

Large Language Models (LLMs) act as reactive agents, primarily engaging in one-step predictions without explicit planning or deliberation mechanisms. This reactivity, often viewed as a limitation (van Lier, 2023; Floridi, 2023; Wu et al., 2024; Mahowald et al., 2024), does not inherently negate the presence of underlying objectives that the agent may pursue (Brooks, 1991). The behaviour of reactive agents is driven directly by their immediate input, thus their ability to find and select information, and deal with uncertainty has been seen as limited (Kaelbling et al., 1998). Yet, they have been shown to be able to determine their inputs in certain cases (Nolfi, 2002; Bonet, 2010). Indeed, reactive systems can perform effective information-seeking behaviours, crucial under uncertainty such as when communicating with hard-to-interpret agents having different knowledge of the interaction context (Paek and Horvitz, 2000; Ognibene and Demiris, 2013), and traditionally associated with more complex deliberative systems and explicit uncertainty reasoning (Beer and Di Paolo, 2023). Studies have shown that even simple reactive systems can engage in exploration and reduce uncertainty through epistemic actions, even without internal simulations or memory (Ognibene et al., 2013). These findings suggest that LLMs, despite lacking explicit internal reasoning capabilities, may still engage in goal-oriented behaviors and possess information-gathering capabilities.

Uncertainty management is crucially connected with information seeking in normative decision systems and also plays an important role in the computations ascribed to the brain (Friston et al., 2015; Kaelbling et al., 1998). However, how task-level uncertainty is processed in multi-layered deep generative models, particularly auto-regressive LLMs (Brown et al., 2020; Radford et al., 2019), and how they compare to normative systems remains largely unexplored. This is particularly interesting given the breadth of information they contain across disparate domains, in contrast to the limited domain variables usually dealt with by normative models.

LLMs learn the conditioned probability distribution of the next token given the sequence of previous tokens (input context) and produce output tokens sampling sequentially from the learnt distribution model (Radford et al., 2018). While uncertainty regarding the next output may be known and represented by the last layer of the model, the long-term evolution of generation or semantic uncertainty is not immediately available (Farquhar et al., 2024; Kuhn et al., 2023; Chen and Mueller, 2023). In fact, when LLMs are used in chatbots, text generation or other non single token output tasks, the stochastic production of a single output is appended to current context input and participates in the generation of successive outputs. Concatenating stochastic processes may result in extended non factual responses or "hallucinations", starting

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

from one first ambiguous output (Ji et al., 2023; LeCun, 2023).

An important contributing factor is that information about the mental state, knowledge, beliefs, and desires of the writer at the moment of writing the text is not directly available for LLMs during training. This may reduce learning performance (Bianco and Ognibene, 2022) and lead to semantically different next token to have the same probability and contribute to generating hallucinations. Moreover, many LLMs in chatbots appear trained to present overconfident responses even when uninformed and produced in an uncertain state (Chen and Mueller, 2023).

To correct belief tracking (Mrkšić et al., 2017) and uncertainty estimation, i.e. estimating how much an agent knows and does not know about the environment, the domain and the current situation, can be useful to adopt effective uncertainty reduction strategies (Kaelbling et al., 1998; Friston et al., 2015; Taniguchi et al., 2023) such as the generation of questions and clarifications (Varges et al., 2010; Kominis and Geffner, 2017; Tellex et al., 2012). However, models that explicitly reason about uncertainty and lack of knowledge have to face challenging computational complexity due to the expansion of the state space (Kaelbling et al., 1998). Various approximations have been developed also based on neural architectures and reinforcement learning (Ognibene and Baldassare, 2014; Wu et al., 2021; Xu et al., 2022; Wang et al., 2020), and, while these approaches may be particularly data hungry (Schatzmann et al., 2007; Wang et al., 2020), it is worth investigating if and how different LLMs learn similar information-gathering strategies as well as implicitly encode beliefs and uncertainty.

However, unveiling how LLMs may couple information integration and acquisition is challenging due to the limited accessibility and interpretability of LLMs and the stochastic recursive process they use to produce output. Similarly to other approaches (Kuhn et al., 2023; Chen and Mueller, 2023), we use a repeated sampling approach through prompts, or repeated zero shots tests (Brown et al., 2020), to retrieve probabilistic information on the information integration process inside the LLM, i.e. how information in previous dialogue exchanges is integrated into a belief and how this belief determines the output selection strategy. We estimate the evolution of this belief state during the interaction and information acquisition using information theoretic methods, i.e. we measure the entropy of the responses distribution (Ognibene and Demiris, 2013; Friston et al., 2015; Ognibene et al., 2019), an approach already tested analysing information integration in black box models (Lungarella and Sporns, 2006). In other words, we aim to observe if the low-level stochastic process of token production of the LLM can be connected to the dynamic process of task-level information integration. This may later lead to novel and more effective task-level uncertainty management for LLMs.

Building on the experiments from cognitive science on information acquisition strategies (Ruggeri and Lombrozo, 2015) and the analysis of Bertolazzi et al. (2023), our study will computationally trace the belief states of LLMs through repeated queries. By examining the probabilistic responses of these models, we aim to gain deeper insights into their inference processes and uncertainty management. Our approach is inspired by Bayesian filtering, which involves continuously updating the probability distributions of candidate items based on new information from each interaction. This approach allows the models to refine their predictions dynamically, integrating new data to reduce uncertainty over time. When the model is correct, Bayesian models estimations are optimal (Särkkä and Svensson, 2023).

Bayesian filtering, commonly used in dynamic systems for state estimation, applies here as we treat the sequence of interactions as a time series. The model updates its belief state with each dialogue step, combining prior knowledge with new evidence. This method enhances the model's ability to manage and process evolving information, mirroring the cognitive processes involved in human decision-making.

Additionally, we will explore the performance difference between GPT-3 and GPT-4o, investigating how these models handle probabilistic data differently. This study will help frame LLMs as complex systems with significant probabilistic reasoning capabilities, albeit with notable flaws. Addressing these limitations is crucial for enhancing the reliability and accuracy of LLMs, particularly in high-stakes environments such as clinical diagnostics and financial forecasting. In chatbots uncertainty about users' requests, context or domain knowledge affects dialog and could elicit queries aimed at resolving it, but such capabilities are lim-

ited at the moment.

2 Related Work

In recent years, various methods have been proposed to define and quantify uncertainties in the context of Large Language Models (LLMs).

For instance, Chen and Mueller (2023) propose a technique to estimate a numeric confidence score for any LLM output generated by a black-box API. This method involves multiple API calls with varying prompts and sampling temperatures, providing users with a confidence estimate that highlights unreliable outputs. Similarly, Yang et al. (2023) introduce a framework to produce uncertainty-aware LLMs capable of estimating aleatoric, epistemic, or composed uncertainty for each prediction in a model- and data-agnostic manner. Their models learn data-dependent thresholds, enhancing prediction reliability.

Huang et al. (2024) present a unified calibration framework that treats both response correctness and associated confidence levels as distributions. Their approach improves calibration through fine-tuning, integrating relevant documents, and adjusting sampling temperatures. Additionally, Zhang et al. (2024) develop LUQ, a sampling-based uncertainty quantification approach for long texts. LUQ identifies LLMs' lack of confidence in generating factual long texts and proposes the LUQ-ENSEMBLE method, which enhances factuality by ensembling responses from multiple models to select the least uncertain response.

Nevertheless, these methods face limitations in interactive settings due to challenges in measuring the evolution of uncertainty and integrating information throughout interactions (Bertolazzi et al., 2023).

Further contributions addressing these limitations include Ren et al. (2023), who propose a framework for measuring and aligning the uncertainty of LLM-based planners. Their approach ensures that planners recognise their uncertainty and request assistance when necessary, utilising conformal prediction theory to provide statistical guarantees on task completion while minimising human intervention in complex multi-step planning scenarios. This method effectively measures the evolution of uncertainty and integrates information dynamically during interactions. Similarly, Hou et al. (2023) present a method that, instead of ensembling models with different parameters, generates a set of clarifications for the input, processes these through fixed LLMs, and ensembles the corresponding predictions. This approach addresses the integration of information across interactions by generating and processing multiple clarifications.

It is pivotal to highlight that uncertainty handling requires considering both token-level and task-level processes. Information theory approaches, like those discussed by Lungarella and Sporns (2006), can provide valuable insights into how information flow and entropic measures at different levels can be used to analyse and manage these uncertainties effectively.

Our work builds upon these foundational studies by investigating the probabilistic reasoning capabilities of LLMs in dynamic and interactive environments. We aim to fill the gap in understanding how uncertainty evolves throughout interactions and how belief states of LLMs are affected by repeated queries. By comparing the performance of different LLMs, specifically GPT-3 and GPT-4o, we seek to provide deeper insights into their strengths and limitations in managing uncertainty. Our approach leverages repeated sampling and behavioural analysis to develop a comprehensive understanding of LLMs' inference processes, contributing to the broader goal of enhancing the robustness and reliability of LLMs in real-world applications.

3 Experiments

Firstly, we quantified the uncertainty of the model at each step of a dialogue between the Questioner and Oracle. Drawing from the "20 Questions game" (Bertolazzi et al., 2023), we measured the uncertainty associated with both the questions posed and the responses received during the interactions, generated with GPT-3 and GPT-4o.

In order to illustrate a typical interaction in the mentioned game, we provide an example of a prompt supplied to a LLM.

> You will be given of a dialogue of the 20 questions game. You have to list out absolutely all the items from the given candidates set that satisfy each <question, answer> in the given dialogue. The output should strictly use the following template:
> EXPLANATION:

CANDIDATES: item1, item2, item3

Dialogue: target = dalmatian

- *Answerer*: This is the list of candidates: dog, bear, flamingo, hawk, toucan, dalmatian, hippopotamus, chick.

- *Questioner*: Is the item you have assigned an animal?
- *Answerer*: Yes.

- *Questioner*: Is the animal you have assigned a mammal?
- *Answerer*: Yes.

- *Questioner*: Is the mammal you have assigned a carnivore?
- *Answerer*: Yes.

- *Questioner*: Is the carnivorous mammal you have assigned typically found in water?
- *Answerer*: No.

- *Questioner*: Does the carnivorous mammal you have assigned have spots on its body?
- *Answerer*: Yes! That's correct.

Building upon further analyses from the 20 Questions game paper, we implemented an additional method to assess each candidate item's consistency with every question-answer pair throughout the dialogue d_t (with $t \in [0, T]$). Firstly, this methodology allows us to determine which items are systematically excluded at each step t of the dialogue.

To execute this, we employed an additional LLM agent tasked with verifying, given a dialogue d_t up to a certain point t and a candidate item, whether the item was deemed to satisfy all the question-answer pairs of d_t. This query was sampled k times, recording the number of positive occurrences. Consequently, at a given moment t, each i-th item was assigned a probability score computed as follows:

$$p_i(t) = \frac{1}{k} \sum_{j=1}^{k} \delta_{ij}(t) \tag{1}$$

where $\delta_{ij}(t)$ is an indicator function that is 1 if the i-th item satisfies all the question-answer pairs of d_t in the j-th query, and 0 otherwise.

After computing these scores and normalising them, we further calculated the probability distribution across all candidate items. This comprehensive approach provides a dynamic view of the model's uncertainty management and enhances our understanding of the probabilistic reasoning capabilities of LLMs within interactive scenarios, effectively demonstrating the principles of Bayesian filtering by continuously updating beliefs based on incoming data.

Once the probability distributions for each dialogue were calculated, we proceeded to analyse the entropy to assess task-level uncertainty. This analysis involved examining the entropy levels of the distributions at various stages of the dialogue to assess the degree of uncertainty and information gain as the dialogue progressed. The entropy is calculated for the distribution over the items for each dialogue, and then the mean entropy is computed by averaging over all dialogues, resulting in a mean entropy for each step of the dialogue. The mean entropy at each stage t of the dialogue is given by:

$$\overline{H}(t) = \frac{1}{D} \sum_{d=1}^{D} H(p^{(d)}(t)) \tag{2}$$

where $p^{(d)}(t)$ represents the probability distribution of the items at stage t in dialogue d.

By measuring the changes in entropy, we could evaluate how effectively the LLM was processing and refining information through its interactions, and identify any patterns or anomalies in its approach to reducing uncertainty.

We also tested an alternative approach by asking both GPT-3 and GPT-4o to evaluate the validity of items given a dialogue d at step t. Instead of querying each candidate item individually, we presented the entire list of items to the models simultaneously and requested them to identify the valid items (simultaneous approach). This method allows LLM models to consider all options at once, potentially using their comparative reasoning capabilities.

In this approach, the models provided items deemed consistent with the dialogue context up to step t. This method offers a different perspective on assessing candidate items, focusing on the models' ability to process and filter multiple options in parallel.

4 Results

Our analysis identifies a "resurrection" phenomenon, where LLMs reassign non-zero probabilities to outcomes previously deemed invalid. This occurs in about 80% of GPT-3 dialogues. Figures 1 and 2 quantify this for GPT-3. For GPT-4o, Figures 3 and 4 show similar results. Interestingly, the phenomenon appears to be significantly influenced by the approach used, with the simultaneous

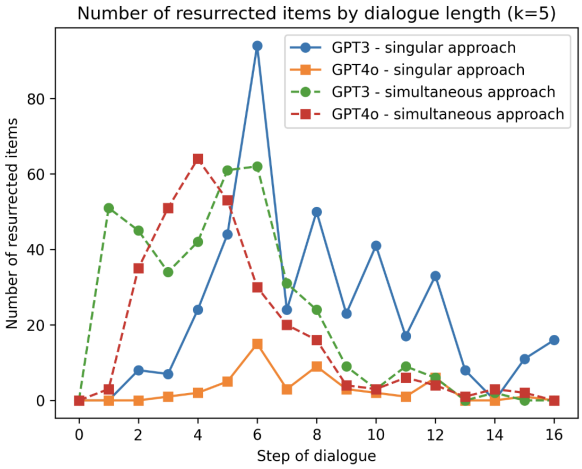


Figure 1: Number of resurrected items for each model and approach as a function of dialogue length on GPT-3 dialogues.

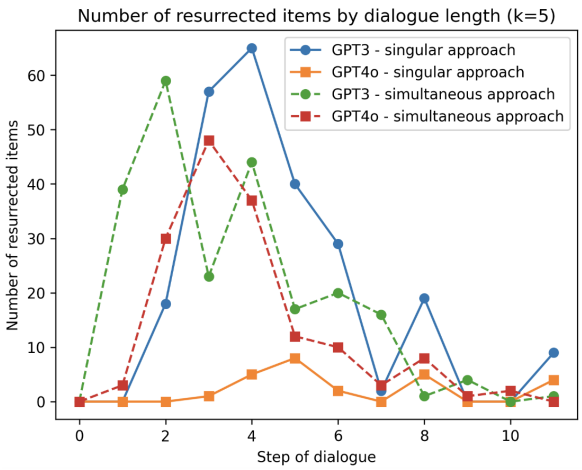


Figure 3: Number of resurrected items for each model and approach as a function of dialogue length on GPT-4o dialogues.

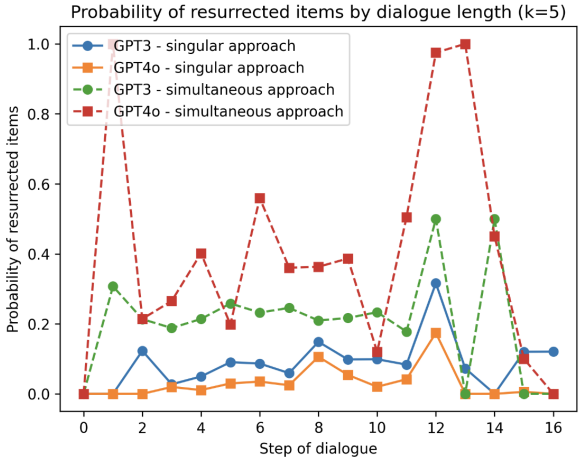


Figure 2: Mean probability absorbed by resurrected items at each step of the dialogue on GPT-3 dialogues.

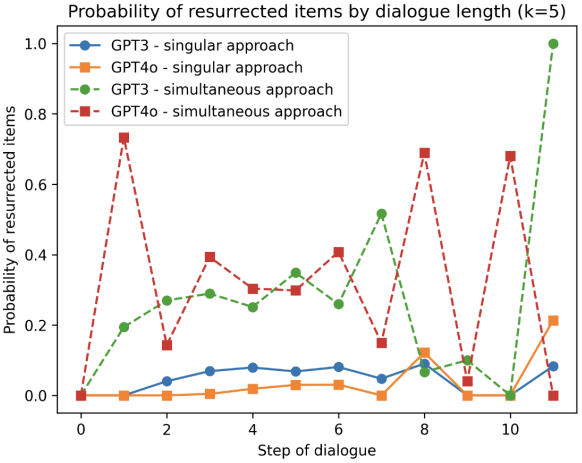


Figure 4: Mean probability absorbed by resurrected items at each step of the dialogue on GPT-4o dialogues.

approach being more affected by the resurrection phenomenon for GPT-4o dialogues.

The duration of the dialogues varies significantly across the dataset. This variation is illustrated in the graph presented in Figure 5, where we compare GPT-3 and GPT-4o for the task, which in this context corresponds to the duration of the dialogue, as a shorter duration implies arriving at a solution with fewer questions.

Figure 6 compares entropy trends for GPT-3 and GPT-4o. The graph includes the ideal entropy curve, which represents the evolution of the probability distribution entropy if the search for the item is carried out optimally, using a binary search approach. GPT-4o outperforms GPT-3, showing less sensitivity to varying k. GPT-3's performance improves with higher k, though it remains more uncertain. Please note that GPT-4o is not tested with $k > 10$ as the results for $k = 5$ and $k = 10$ are very similar, indicating that it is not necessary to increase the sample size.

Figures 7 to 9 detail entropy and cross-entropy results for different models and dialogue sources. While cross-entropy is seemingly a more precise measure of model performance, as it takes into account the correct response, it is important to consider that entropy is a more appropriate measure in our context. This is because we are primarily interested in the overall reduction of uncertainty, and thus in the strategies the model employs to achieve this effectively, rather than its ability to approximate the correct answer. Although, as the figures suggest, these two aspects tend to go hand in hand.

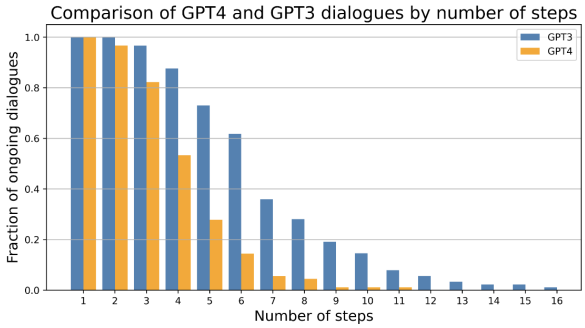


Figure 5: Comparison of dialogue durations for GPT-3 and GPT-4o.

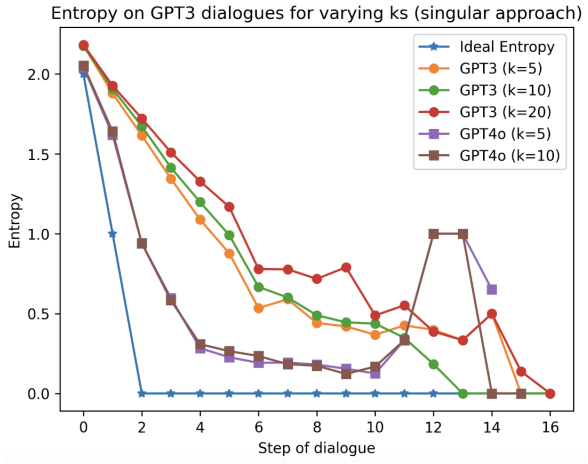


Figure 6: Comparison of entropy trends for GPT-3 and GPT-4o from the beginning of the dialogue, across various k values, with the ideal entropy curve.

Additional insights were gained by analysing dialogue steps in which the model either erroneously confirmed a target (entropy = 0) or generated distributions with only zero values. These instances, observed only in dialogues processed by GPT-4o, were marked by significant ambiguities or errors, often resulting in the incorrect elimination of the true target. This error analysis also extended to GPT-3 and GPT-4o's ability to consistently list valid candidates at each step, revealing differences in their performance throughout the dialogues.

Figures 11 and 12 display the probability of zero distributions by dialogue step for GPT-3 and GPT-4o dialogues, respectively. The results suggest that the phenomenon of Bayesian apocalypse, where all item probabilities approach zero, is more prevalent at specific dialogue steps and is sensitive to the approach used.

Finally, Figures 13 and 14 compare entropy results between simultaneous and singular approaches. The simultaneous approach improves

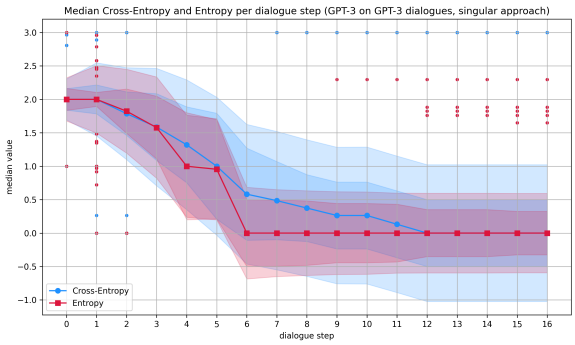


Figure 7: Entropy and Cross-Entropy levels for GPT-3 on dialogues generated by GPT-3.

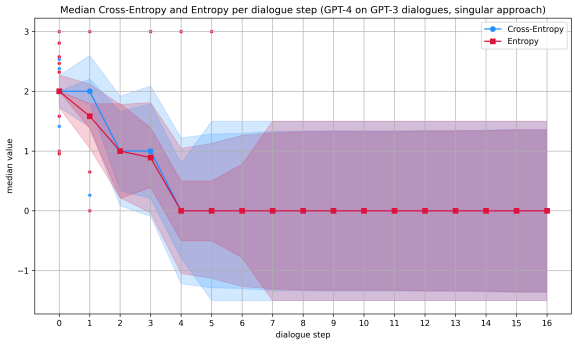


Figure 8: Entropy and Cross-Entropy levels for GPT-4 on dialogues generated by GPT-3.

GPT-3 performance but degrades GPT-4o results at the most significant steps, which are those with the highest number of samples (see Figure 5). GPT-4o consistently outperforms GPT-3, with the simultaneous approach proving to be more beneficial for GPT-3 while negatively impacting GPT-4o performance.

5 Discussion

Our study extends the analysis of Bertolazzi et al. (2023) by computationally tracking LLM belief states through repeated queries. This reveals that entropy decreases with each interaction, and decisions are made when uncertainty is minimised, consistent with normative models (Friston et al., 2015; Ognibene and Demiris, 2013). This suggests a link between the low-level stochastic processes of token production and the integration of higher-level task information. Future work could explore the extraction of uncertainty and information gain predictions from LLM internal states. However, the significant noise in the process suggests that current LLMs, particularly GPT-3, may benefit from targeted training to improve performance.

GPT-4o's better performance compared to GPT-

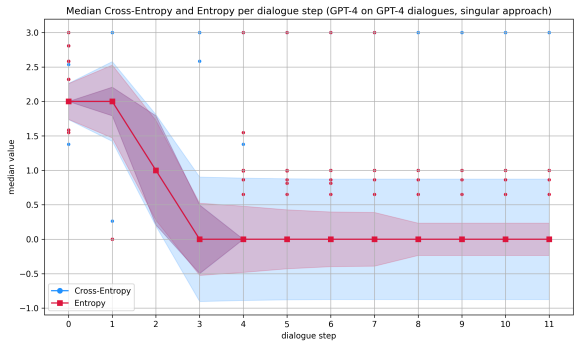


Figure 9: Entropy and Cross-Entropy levels for GPT-4 on dialogues generated by GPT-4.

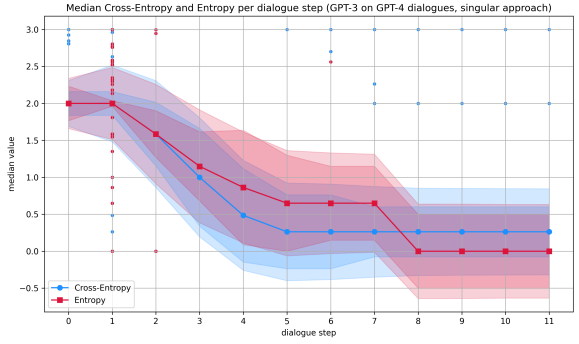


Figure 10: Entropy and Cross-Entropy levels for GPT-3 on dialogues generated by GPT-4.

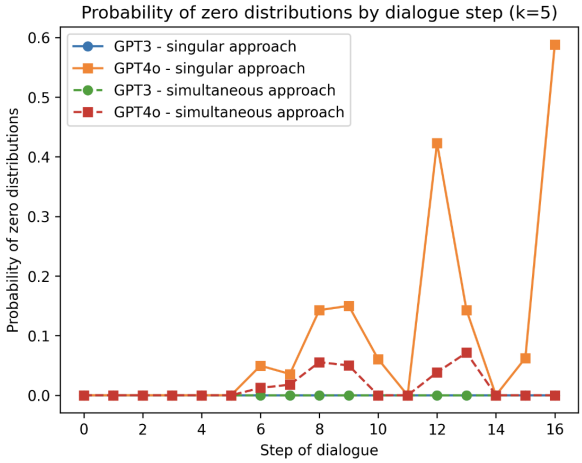


Figure 11: Probability of zero distributions (Bayesian apocalypse) by dialogue length for GPT-3 and GPT-4o on GPT-3 dialogues.

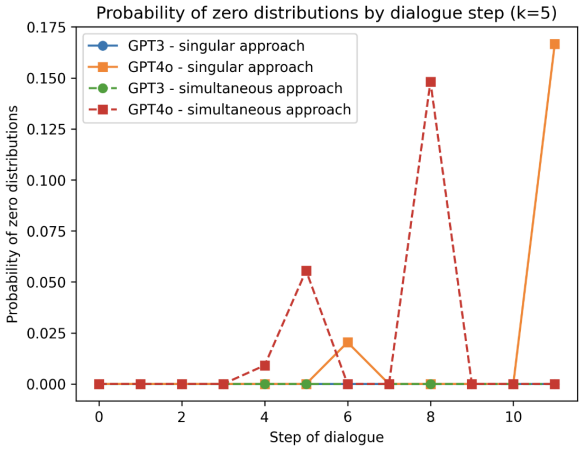


Figure 12: Probability of zero distributions (Bayesian apocalypse) by dialogue length for GPT-3 and GPT-4o on GPT-4o dialogues.

3 on the same dataset likely reflects GPT-3's limitations in retaining and analysing critical information. Although GPT-3 appears to perform better with the alternative approach, this could be misleading. GPT-4o typically reaches solutions faster with shorter dialogues, as shown in Figure 5, where only 25% of GPT-4o dialogues exceed the fifth step, while GPT-3 dialogues often extend to the sixth step. The perceived advantage of GPT-3 in later steps may thus stem from different dialogue lengths rather than actual performance improvements.

The "resurrection" phenomenon, in which LLMs reassign nonzero probabilities to previously invalidated outcomes, is quantified by tracking the frequency and magnitude of these probabilities. As shown in Figure 2, GPT-3's probability for resurrected items slightly decreases over time, indicating how the model handles uncertainty. Figure 4 shows that this phenomenon is less pronounced in GPT-4o. However, the simultaneous approach exacerbates this issue.

The "Bayesian apocalypse", where all probabilities approach zero, results in high uncertainty and challenges in distinguishing valid from invalid hypotheses (Bengtsson et al., 2008). This phenomenon mirrors issues in particle filtering, where insufficient particles fail to represent the true state distribution, leading to similar collapses in probability. This exacerbates model uncertainty and impairs decision-making.

Probability collapses may also indicate hallucinations within dialogues, especially when options are finite. This issue is critical in interactive settings where consistent tracking of candidates is necessary. Hallucinations can lead to incorrect responses and premature elimination of valid options, highlighting the need for better uncertainty management and handling of incomplete or noisy data.

Comparing GPT-3 and GPT-4o, we used both singular and simultaneous sampling approaches. GPT-4o's superior performance likely stems from better information retention and analysis compared

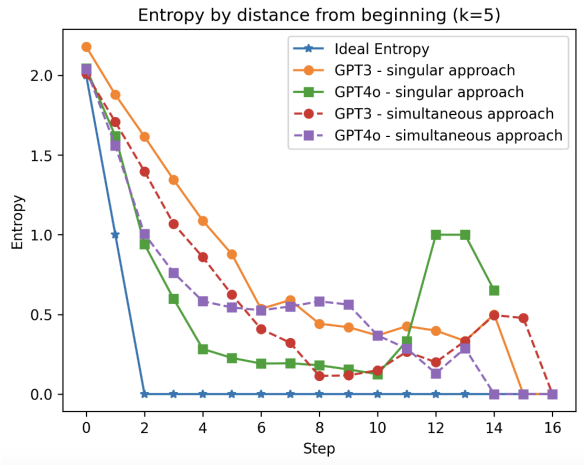


Figure 13: Comparison of entropy trends across models and approaches with GPT-3 dialogues.

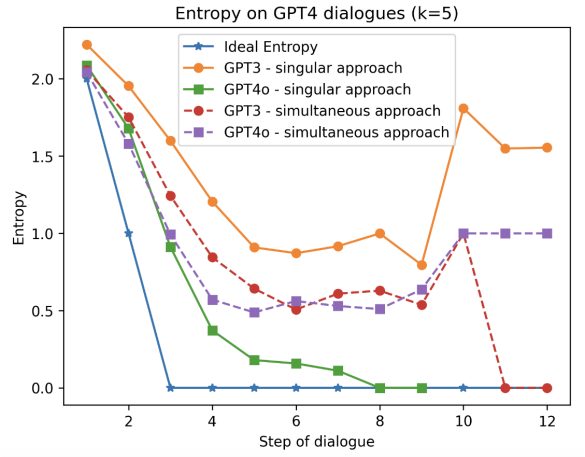


Figure 14: Comparison of entropy trends across models and approaches with GPT-4o dialogues.

to GPT-3. However, GPT-3's apparent advantage with the singular approach after six steps (Figure 13) may be misleading. GPT-4o's shorter dialogues often mean it reaches solutions more quickly, as indicated by Figure 5. Therefore, the perceived benefit of GPT-3 in subsequent steps may reflect differences in dialogue length rather than true performance.

Our findings show that LLMs can effectively explore and identify relevant information without extensive task-specific training, aligning with reactive systems research (Beer and Di Paolo, 2023; Ognibene et al., 2013). Although the entropy reduction approach is effective, current strategies for information integration are still suboptimal. GPT-4o demonstrates more robust performance, but issues in managing probabilistic data and avoiding probability collapses persist. Addressing these challenges is crucial for improving the reliability and accuracy of LLMs, especially for high-stakes applications.

Our results align with Bayesian inference principles, where uncertainty is minimised by updating probability distributions with new evidence. Similarly, LLMs update token predictions based on preceding context, aiming to reduce output uncertainty. Despite lacking explicit task-level uncertainty representation, LLMs dynamically integrate new information, reflecting a Bayesian-like process in their operation.

References

Randall D Beer and Ezequiel A Di Paolo. 2023. The theoretical foundations of enaction: Precariousness. *Biosystems*, 223:104823.

Thomas Bengtsson, Peter Bickel, and Bo Li. 2008. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and statistics: Essays in honor of David A. Freedman*, volume 2, pages 316–335. Institute of Mathematical Statistics.

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. Chatgpt's information seeking strategy: Insights from the 20-questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162.

Francesca Bianco and Dimitri Ognibene. 2022. Robot learning theory of mind through self-observation: Exploiting the intentions-beliefs synergy. *arXiv preprint arXiv:2210.09435*.

Blai Bonet. 2010. Conformant plans and beyond: Principles and complexity. *Artificial Intelligence*, 174(3-4):245–269.

Rodney A. Brooks. 1991. Intelligence without reason. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'91, page 569–595, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large

language models using semantic entropy. *Nature*, 630(8017):625–630.

Luciano Floridi. 2023. Ai as agency without intelligence: on chatgpt, large language models, and other generative models. *Philosophy & technology*, 36(1):15.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. Active inference and epistemic value. *Cognitive neuroscience*, 6(4):187–214.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *ICML 2024*.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Filippos Kominis and Hector Geffner. 2017. Multiagent online planning with nested beliefs and dialogue. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 27, pages 186–194.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Y LeCun. 2023. Do large language models need sensory grounding for meaning and understanding. In *Workshop on Philosophy of Deep Learning, NYU Center for Mind, Brain, and Consciousness and the Columbia Center for Science and Society*.

Max Lungarella and Olaf Sporns. 2006. Mapping information flow in sensorimotor networks. *PLoS computational biology*, 2(10):e144.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Stefano Nolfi. 2002. Power and the limits of reactive agents. *Neurocomputing*, 42(1-4):119–145.

Dimitri Ognibene and Gianluca Baldassare. 2014. Ecological active vision: four bioinspired principles to integrate bottom–up and adaptive top–down attention tested with a simple camera-arm robot. *IEEE transactions on autonomous mental development*, 7(1):3–25.

Dimitri Ognibene and Yiannis Demiris. 2013. Towards active event recognition. In *IJCAI*, pages 2495–2501.

Dimitri Ognibene, Lorenzo Mirante, and Letizia Marchegiani. 2019. Proactive intention recognition for joint human-robot search and rescue missions through monte-carlo planning in pomdp environments. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*, pages 332–343. Springer.

Dimitri Ognibene, Nicola Catenacci Volpi, Giovanni Pezzulo, and Gianluca Baldassare. 2013. Learning epistemic actions in model-free memory-free reinforcement learning: Experiments with a neurorobotic model. In *Biomimetic and Biohybrid Systems: Second International Conference, Living Machines 2013, London, UK, July 29–August 2, 2013. Proceedings 2*, pages 191–203. Springer.

Tim Paek and Eric Horvitz. 2000. Conversation as action under uncertainty. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 455–464.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*.

Azzurra Ruggeri and Tania Lombrozo. 2015. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216.

Simo Särkkä and Lennart Svensson. 2023. *Bayesian filtering and smoothing*, volume 17. Cambridge university press.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, NAACL-Short '07, page

149–152, USA. Association for Computational Linguistics.

Tadahiro Taniguchi, Shingo Murata, Masahiro Suzuki, Dimitri Ognibene, Pablo Lanillos, Emre Ugur, Lorenzo Jamone, Tomoaki Nakamura, Alejandra Ciria, Bruno Lara, et al. 2023. World models and predictive coding for cognitive and developmental robotics: frontiers and challenges. *Advanced Robotics*, 37(13):780–806.

Stefanie Tellex, Pratiksha Thaker, Robin Deits, Thomas Kollar, and Nicholas Roy. 2012. Toward information theoretic human-robot dialog. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia.

Maud van Lier. 2023. Understanding large language models through the lens of artificial agency. *Swedish Artificial Intelligence Society*, pages 79–84.

Sebastian Varges, Silvia Quarteroni, Giuseppe Riccardi, and Alexei V Ivanov. 2010. Investigating clarification strategies in a hybrid pomdp dialog manager. In *Proceedings of the SIGDIAL 2010 Conference*, pages 213–216.

Sihan Wang, Kaijie Zhou, Kunfeng Lai, and Jianping Shen. 2020. Task-completion dialogue policy learning via Monte Carlo tree search with dueling network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3461–3471, Online. Association for Computational Linguistics.

Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2021. Partially observable reinforcement learning for dialog-based interactive recommendation. In *Proceedings of the 15th ACM conference on recommender systems*, pages 241–251.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. 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 1819–1862.

Yunqiu Xu, Meng Fang, Ling Chen, Yali Du, Joey Zhou, and Chengqi Zhang. 2022. Perceiving the world: Question-guided reinforcement learning for text-based games. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 538–560, Dublin, Ireland. Association for Computational Linguistics.

Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, Alexander Amini, and Alejandro Perez. 2023. Uncertainty-aware language modeling for selective question answering.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. Luq: Long-text uncertainty quantification for llms.

Swann's Name: Towards a Dialogical Brain Semantics

Jonathan Ginzburg
CNRS, Université Paris-Cité
Laboratoire de Linguistique Formelle
yonatan.ginzburg@u-paris.fr

Chris Eliasmith
Centre for Theoretical Neuroscience
University of Waterloo
celiasmith@uwaterloo.ca

Andy Lücking
Goethe University Frankfurt
Text Technology Lab
luecking@em.uni-frankfurt.de

Abstract

The paper argues with reference to several examples that dialogical dynamic semantics, the idea that meaning arises from emergent public context, breaks down over extended temporal periods, ignoring as it does individual differences specifically with respect to memory dynamics. We argue, following several recent works, that this highlights the need for a semantics that is brain-based. We offer a sketch for such a semantics by developing a hybrid model that integrates work on memory–oriented dialogue semantics with work in the semantic pointer architecture for functional brain modelling.

1 Introduction

Dialogical dynamic semantics, the idea that meaning arises from emergent public context, can be effective for dialogue over short temporal periods. But over more extended temporal periods, dynamic semantics begins to break down, ignoring as it does individual differences specifically with respect to memory dynamics. Consider the following mundane story: I encounter my neighbour's daughter Swann when she gets locked out and learn her name. Two years pass: I encounter Swann occasionally, as I hear her close the entrance door, but I do not hear her name spoken. One morning I see Chloé, Swann's sister, and wonder: what is Chloe's sister's name? I remember it starts with 'S'. But I cannot remember the name. This lasts for a while. I see a list of names and know that they are not the name. Finally I see the name and recognize it. This *inner dialogue* can also be envisioned as a series of external dialogues:

(1) a. Dialogue 1: Neighbour: *This is Swann.* Me: *Nice to meet you.*

b. Interlude (time passes, events happen)

c. Dialogue 2: (I see Chloé) Me: *How is um* (pause, frowns) *your sister*? Chloé: *Swann*? Me: *Yes.*

(2) provides an additional illustration of the effect of time—dissociation between event-based, individual-based, and metalinguistic information, as exemplified in (2b), a dissociation backed by considerable clinical evidence (Greenberg and Verfaellie, 2010; Bastin et al., 2019).

(2) a. A: *Look, someone's broken the door handle.* B: *Right.* C: *Yeah it's this woman, Sloane.*

b. (a week later) D: *What had happened*? A: *What's her name, I forget, broke the door handle.* D: *and Bill was there too apparently.* B: *Who*? A: *Her partner.* B: *I don't know him.* A: *We met him last week. B: Oh, I see.*

We think cases such as these highlight the need, already outlined in several works (Eliasmith, 2013; Baggio, 2018; Hagoort, 2020; Macnamara and Reyes, 1994; Jackendoff, 2002; Seuren, 2009) for a semantics that is brain-based (where again one can appeal to the (biophysical/biochemical) neuronal and the neuron-network levels): generalizations about behaviour can occur at various levels (Marr, 1982; Bechtel, 2007; Eliasmith and Kolbeck, 2015); appealing also to brain-based levels need not mean that all explanations are most usefully stated at those levels—for instance, as we will see certain rules concerning dialogue coherence.

And yet, we think, nonetheless, that this data enables one to make stronger claims, namely that a brain-based account impacts also on the structure of the cognitive theory one can and should provide. In particular, it requires us to capture (i) the intrinsically associative character of memory (exemplified here by the speaker's thinking of Chloé's

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

sister when seeing Chloé, mirrored by corresponding external dialogue coherence) (ii) dissociative aspects in cognitive states (exemplified by forgetting Swann's name but not Swann and data in (2)), (iii) the pervasive nature of forgetting and the non-redundancy of reproviding (forgotten) information, and (iv) differences in communal memory emergent from individual divergences.

The paper is structured as follows: in section 2 we introduce relevant background about the various neural levels. We develop our account in two stages: in section 3 we apply an externalist, though memory-oriented dialogue framework NeuroKoS (Ginzburg and Lücking, 2022) to the data, which can only offer a partial account; in section 4 we discuss a simple model of the data using the Semantic Pointer brain-modelling framework (Eliasmith, 2013), which offers an account of the aspects which NeuroKos cannot handle.

2 Learning and Forgetting at the Neural Level

2.1 Short-term v. Working Memory v. Long-term Memory

The neuropsychological basis for short-term and long-term memory (STM, LTM) distinctions are both experimental (e.g., ability to recall number sequences or labelled pictures after a single presentation) and based on studies of patients, most notably the patient Henry Molaison (aka H.M.), well known for being high functioning despite lacking the ability to form new (episodic) memories that could persist beyond 45 minutes (Scoville and Milner, 1957; Milner and Klein, 2016; MacKay et al., 2013; Squire and Wixted, 2011). Working memory (WM) is a distinct though closely related notion to short-term memory amounting to 'an actively engaged system used to store information that is relevant to the current behavioral situation.' (Eliasmith, 2013, p. 211). Baddeley (1988, 2012) offered both arguments for the notion of WM and developed an influential framework, M-WM, which postulates a clear structure for WM (on which more below); an alternative to this was proposed by Cowan (2001), who emphasizes the capacity constraints of WM. Both Baddeley's episodic buffer and Cowan's focus of attention are chunk limited buffer stores, and both models by and large agree on a capacity limit of four chunks. An important issue such theories have contended with is whether working memory is a separate system (Baddeley) or merely a temporal slice from a unified memory system (Cowan, on one reading, though ultimately the differences between the frameworks are not large). Norris (2017) argues that STM/LTM are distinct systems given the need for (i) memory for previously unencountered information, (ii) storage of multiple tokens of the same type, and (iii) variable binding (in one sense of the term). Be that as it may, the exact relationship between WM (which is evinced in actual use) and STM/LTM is not fully clear. What is clear is that there are WM/LTM distinctions at neural and neural network levels.

2.2 Short-term and Long-term Learning at the Neural Level

Given the relative ease of access to their neural systems, the solidly established results on learning at the neural level have arisen from various invertebrates and from rodents. As exposited by Kandel et al. (2014) one can distinguish two classes of mechanisms: short/medium term changes in synaptic strength arising from specific patterns of electrical activity or the action of modulatory transmitters; long-lasting synaptic and behavioral memory plasticity requires epigenetic mechanisms—changing gene expression without modifying the underlying DNA: on the one hand the inhibition of miRNA-124 which facilitates the activation of CREB-1, which begins the process of memory consolidation, and on the other hand the delayed activation of piRNA, which leads to the methylation and consequent repression of the promoter of CREB-2. This allows CREB-1 to be active for a longer period of time.

2.3 Short-term and Long-term Learning at the Neural Network Level

As far as LTM goes, it is commonly assumed that memories are not stored in the hippocampus as such, but arise from the interaction of representations based at the hippocampus with neocortical information: sparsely-coded hippocampal neurons referencing and activating the neocortical neurons to re-create the content of an experience (Teyler and Rudy, 2007). Semantic memories, arising by generalisation across the neocortical representations of episodic memories are resistant to hippocampal damage. For a long period, the fact that performance on many explicit tasks is affected by temporally graded retrograde amnesia was explained by assuming that the hippocampus is only a temporary repository for memory whereas the neocor-

tex stores the memory (Squire and Wixted, 2011). More recently, evidence emerged that mediotemporal lobe lesions do not lead to a pattern of retrograde amnesia and also affect non-episodic, semantic memory. Sekeres et al. (2018) propose Transformation Trace Theory (TTT): *transformed* memories (i.e., ones shorn of detail) come to be represented in distributed neocortical networks from where they can be recovered without the involvement of the hippocampus; *detailed* episodic memories are always dependent on the hippocampus. The evidence for this is evidence that once a consolidated memory is reactivated, it can become labile and once again become susceptible to the effects of hippocampal disruption.

This leads to at least the following sources for forgetting, which models of forgetting need to tie into:

1. Non-consolidated short-term memories;
2. Detail modification during activation (Sekeres et al., 2018);
3. Loss as a result of neurogenesis (Weisz and Argibay, 2012; Epp et al., 2016);
4. Weight decay and synapse elimination (Richards and Frankland, 2017).

3 Towards an Account

3.1 Combining Memory and Dialogue GameBoards

As mentioned earlier, we draw on an earlier proposal, the only existing one to our knowledge, for combining externalist dialogue semantics with memory structure (Ginzburg and Lücking, 2020, 2022). But first, a brief explanation of externalist dialogue semantics, as conceived in the framework KoS (Ginzburg, 1994; Larsson, 2002; Purver, 2004; Fernández, 2006; Ginzburg, 2012)—formulated using the logical framework TTR (Cooper and Ginzburg, 2015; Cooper, 2023). Instead of assuming a single context to be operative, a collective notion is emergent (Stephens et al., 2010) from individual *Total Cognitive States* (TCS), one per participant. A TCS has two partitions, namely a *private*, and a *public* one, the DGB.

(3) $\text{TCS} =_{def} \begin{bmatrix} \text{public} : \mathit{DGBType} \\ \text{private} : \mathit{Private} \end{bmatrix}$

Dialogue gameboards (see (4) for the basic structure) track various aspects of the emerging context in terms of concrete real world entities and more abstract ones constructed in TTR. The parameters *spkr* and *addr* together with the addressing condition (at a given time) track verbal turns and mutual engagement; *vis-sit* represents the visual situation of an agent, including his or her focus of attention (*foa*), which can be an object (*Ind*), or a situation or event (*Rec*), relevant *inter alia* for processing gestural answers; *facts* represents the shared assumptions of the interlocutors; uncertainty about mutual understanding that remain to be resolved across participants—*questions under discussion*—are a key notion in explaining coherence and various anaphoric processes (Ginzburg, 2012; Roberts, 1996) and is tracked by the parameter *qud*; 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, which captures expectations arising due to illocutionary acts—one act (querying, assertion, greeting) giving rise to anticipation of an appropriate response (answer, acceptance, counter–greeting), also known as adjacency pairs (Schegloff, 2007); finally, *mood* represents the publicly accessible emotional aspect of an agent that arises by publicly visible actions (such as non-verbal social signals, as well as by verbal exclamations), which can but need not diverge from the private emotional state:

(4) $\mathit{DGBType} =_{def}$

$$\begin{bmatrix} \text{spkr} & : & \mathit{Ind} \\ \text{addr} & : & \mathit{Ind} \\ \text{utt-time} & : & \mathit{Time} \\ \text{c-utt} & : & \text{addressing(spkr,addr,utt-time)} \\ \text{facts} & : & \mathit{Set}(\mathit{Prop}) \\ \text{vis-sit} & = & \left[\text{foa} : \mathit{Ind} \vee \mathit{Rec}\right] : \mathit{RecType} \\ \text{pending} & : & \mathit{List}(\mathit{LocProp}) \\ \text{moves} & : & \mathit{List}(\mathit{IllocProp}) \\ \text{qud} & : & \mathit{POSet}(\mathit{Question}) \\ \text{mood} & : & \mathit{Appraisal} \end{bmatrix}$$

TCSs and in particular DGBs change as a result of private perception and public interaction, which can be described in terms of *conversational rules* (Larsson, 2002). We exemplify here three rules (minor variants of rules in Ginzburg, 2012, Chapters 4,6) that will play a role subsequently. The first exemplifies coherence at the level of Moves, the second the emergence of presuppositions, the third the coherence of clarification questions:

(5) a. **Interlocutor introduction rule**: given that the LatestMove is Introduce(A,B,C), this licenses the next move to be Greet(B,C).

b. **FACTS update following assertion acceptance**: if the LatestMove is Accept(A,p), this licenses FACTS:=FACTS $\cup\{p\}$

c. **Confirmation question emergence**: if A's utterance u is (a sub-utterance of) the maximal element of Pending, QUD can be updated with the question *did A mean c by u?* (c some potential referent/content).

KoS provides a theory of meaning for highly context dependent elements such as non-sentential utterances (6a,b), filled pauses (6c), and non-verbal social signals such as smiles or frowns (6d,e), which figure further below.

(6) a. *yes* $\mapsto$ p (p? is MaxQUD);

b. *right* $\mapsto$ Understand(A,u) (u is MaxPending, A current speaker);

(both Ginzburg, 2012)

c. *um* $\mapsto$ Makes λxMeanNextUtt(spkr,Pending,x) MaxQUD (Ginzburg et al., 2014)

d. smile: Given A as speaker, s as smilable event, $\mapsto$ Pleasant(s,A)

e. frown: Given A as speaker, f as frownable event, q : Question $\mapsto$ Raise(f,q,A)

(both Ginzburg et al., 2020)

The essence of the proposal of Ginzburg and Lücking (2020, 2022) is to tie the externally–oriented data structure used to describe dialogue dynamics, the dialogue gameboard (Ginzburg, 2012), with working and long-term memory. Thus, they propose to 'break up' the dialogue gameboard into WM and LTM components, building on models for WM (Baddeley, 2012) and LTM (Bastin et al., 2019), respectively—see Fig. 1 for a graphical summary. In particular, they proposed to (i) view conversations as episodes tracked in episodic memory, (ii) distinguish within LTM the following components: (a) episodic memory typically associated with the hippocampus, (b) entity-based memory (based in the perirhinal cortex, Bastin et al., 2019), and (c) semantic memory (mainly localized in the posterior region of the left temporal lobe, Saumier and Chertkow, 2002, though the specific regions involved in semantic memory retrieval depend on whether sensorimotor or abstract amodal features are accessed, Reilly et al., 2016).[1]

Characterizing the emergence of LTM is of course highly complex—Ginzburg and Lücking (2022) offered one simplified rule concerning episodic memory, but said nothing about entity and semantic memory. We refine very slightly their rule concerning episodic memory and offer two very simplified rules concerning entity and semantic memory. Events undergo appraisal which leads to both updates in the current emotional makeup of the cognitive state (both in the private and in the public parts) and to creating episodic indices in the hippocampus, which are in effect vertices in a network connecting to percepts of events stored neocortically. We assume that such indices are created for events with positive pleasantness above a threshold or negative pleasantness above a larger threshold—which yields a bias for long-term memory of enjoyable events or of highly unpleasant ones. The rule in (7) creates a fresh index and associates it with the current event in working memory, originating either in Pending or in vis-sit:

(7)
$$\begin{bmatrix} \text{pre} : \begin{bmatrix} \text{e=MaxPending} \vee \text{vis-sit} : \textit{RecType} \\ \text{c1} : \text{Private.Mood.pleasant.affect.pve} \geq \theta_1 \\ \quad \vee \text{Private.Mood.pleasant.affect.nve} \geq \theta_2 \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{n = card(HC-Indices)+1} : \mathbb{N} \\ \text{HC-indices} := \text{HC-Indices} \cup \langle \text{n,pre.e} \rangle \end{bmatrix} \end{bmatrix}$$

Although Tulving (1972) suggested that semantic memory was in some sense prior to episodic, recently it has been common to view both entity and semantic memory as emerging from decontextualized episodic traces (and existing in parallel) (Greenberg and Verfaellie, 2010).

We define an *individual-oriented* subpart of a record type as in (8a) and exemplify it as in (8b):

(8) a. Assume l_1 is a label of the record type i and $i \sqsubseteq \left[\text{l}_1 : \text{Ind}\right]$ and for no other label l_i in

[1] From a formal point of view one might say that an entity–oriented semantics has already been proposed in Irene Heim's File-Change Semantics (Heim, 1982), though in that case the episodes are represented within each individual file, which emerges with the utterance of an indefinite. So there is no dissociation and of course no means to deal with forgetting or associative memory. The same is true for related *mental files* approaches (e.g. Maier, 2016).

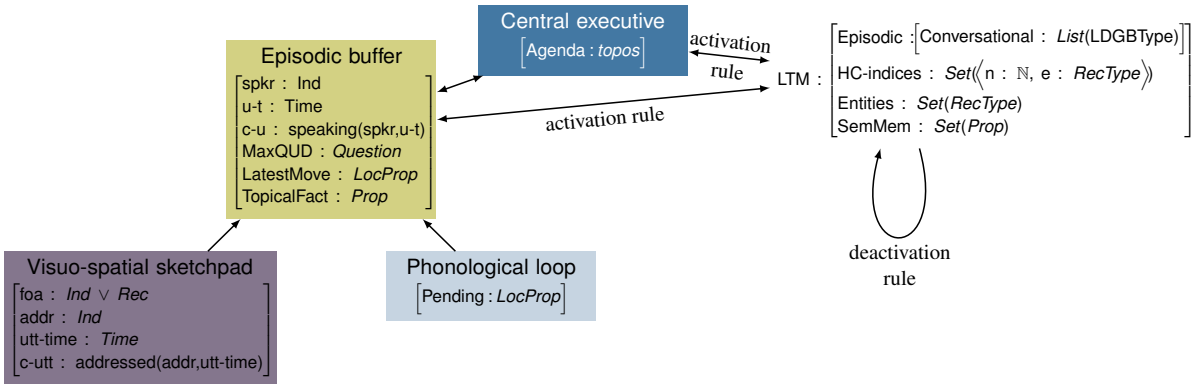


Figure 1: Fusing M-WM and DGB, and adding LTM.

i is it the case that $i \sqsubseteq \left[\mathrm{l}_i : \mathrm{Ind}\right]$ and assume r is a record type such that for some j $r = i \overset{\cdot}{\wedge} j$ ('merge'), then i is an *individual-oriented* subpart of r.

b. $i = \begin{bmatrix} \mathrm{x} & : \mathit{Ind} \\ \mathrm{C} & : \mathit{faceshape} \\ \mathrm{c1} & : \mathrm{C(x)} \\ \mathrm{c}_{\mathit{name}} & : \mathrm{Name(Emmo,x)} \end{bmatrix}$,

$r = \begin{bmatrix} \mathrm{x} & : \mathit{Ind} \\ \mathrm{C} & : \mathit{faceshape} \\ \mathrm{c1} & : \mathrm{C(x)} \\ \mathrm{c}_{\mathit{name}} & : \mathrm{Name(Emmo,x)} \\ \mathrm{y} & : \mathit{Ind} \\ \mathrm{c2} & : \mathrm{Hammer(y)} \\ \mathrm{t} & : \mathit{Time} \\ \mathrm{c3} & : \mathrm{Hold(x,y,t)} \end{bmatrix}$

We will assume that entities emerge in LTM as individual-oriented parts of episodes from episodic memory:

(9) **Entity memory update**: If $\langle n, r \rangle \in$ HC-Indices and i is an individual-oriented part of r, then Entities $:=$ Entities $\cup \{i\}$

The principle we sketch for the emergence of semantic memory involves a subcase of the FACTS update rule (5b) above. We assume that assertions communicating stative information update semantic memory. This is of course quite crude, but presumably a more refined typing of propositions can offer a reasonable starting point for such a procedure.

(10) **Semantic memory update**: If $p \in$ FACTS and p : *StativeProposition*, then SemMem $:=$ SemMem $\cup \{p\}$

We mention one additional principle, which we will not attempt to formalize in the current setup, but which is (partially) formalizable in the neural setup of section 4. It is intuitively correct for *inner dialogue*, and we think reasonably extensible to interactive dialogue:

(11) **Associative topics:** If q is a question whose similarity to MaxQUD $\geq \theta$, $\mathit{Ask}(q)$ is licensed as the LatestMove

3.2 Initial Account

We return to our initial example repeated here as (12):

(12) Dialogue 1: Neighbour: *This is Swann.* Me: *Nice to meet you.*

Given the tools we have, we can explain the following: the coherence of my response to the neighbour's introduction (on the basis of the **Interlocutor introduction rule**, (5a)); the update of entities with the individual Swann (as an update of entity memory, see (9)), the update of semantic memory with Swann's name (as an update of semantic memory, see (10)).

For the second dialogue repeated here as (13):

(13) Dialogue 2: (I see Chloé) Me: *How is um* (pause, frowns) *your sister*? Chloé: *Swann*? Me: *Yes.*

we can explain how the self-repair question introduced by a filled pause licences a frown (see (6)); we can explain the coherence of Chloé's confirmation request (see (5.c)). On the other hand, **we do not have a means of explaining my inability to recall Swann's name** (since it is in my semantic memory), **nor the restorative effect of Chloé's utterance on the availability of Swann's name. Nor do we have a means of explaining why I think of Swann when I see Chloé**; my asking

about Swann could be explained if we had a means of formalizing our rule of associative topics, as a question similar to asking how Chloé is. We suggest that dealing with these unresolved issues requires a brain-oriented semantics, to which we now turn.

4 Adding a Neural Level

4.1 The Semantic Pointer Architecture

We draw on the Semantic Pointer Architecture (SPA) approach to cognition (Eliasmith, 2013). The idea in a nutshell is the following: an input current is nonlinearly encoded within a population of neurons according to each neuron's tuning curve and spiking pattern. The encoded input can either be reconstructed by other populations of neurons by weighted linear decoding (the pair of encoding and decoding defines a neural *representation*), or transformed (by another weighted linear decoding). We employ vectors as a means for representing symbols, dubbing them *semantic pointers* (SPs), since we construe them as compressed representations that carry partial semantic content. Certain transformations can be defined on the class of SPs. One of the most important transformations is *circular convolution* (Plate, 1991), which *binds* two or more vectors into an output vector $\mathbf{v}$ without increasing dimensionality but ensuring also that the input vectors can be *unbound* or *decoded* from $\mathbf{v}$, albeit with some noise.[2,3]

If vectors $\mathbf{d}$ and $\mathbf{e}$ are bound into $\mathbf{p}$, $\mathbf{p} = \mathbf{d} \circledast \mathbf{e}$, "$\circledast$" being circular convolution, then $\mathbf{d}$ can be approximately recovered from $\mathbf{p}$ by binding $\mathbf{p}$ with the inverse of $\mathbf{e}$: $\mathbf{d} \approx \mathbf{p} \circledast \mathbf{e}'$ ($\mathbf{e}'$ being the inverse of $\mathbf{e}$). Encoding, decoding, and transforming are dynamic processes in time and are implemented using the software tool Nengo (Bekolay et al., 2014), which also allows for "biological compilation" in terms of neural simulations.[4]

The SPA has successfully been applied to a number of cognitive tasks, including the representation of concepts (Blouw et al., 2016), memory (Gosmann and Eliasmith, 2021),and emotion (Thagard et al., 2023), and underlies the world's largest functional brain model to date (Eliasmith et al., 2012).

4.2 SPA and Symbolic Representation

A key feature of the SPA is that it enables a systematic correspondence of symbolic and neural content in a way that meets Jackendoff's challenges for cognitive neuroscience (Jackendoff, 2002; Gayler, 2004). In recent work Larsson et al. (2023) show how to map TTR entities into SPA ones, offering a mapping that covers basic types, perceptual and cache-based judgements, singleton types, record types, meet types and merging of record types, ptypes, and subtyping.

4.3 Completing the Account

We employ the SPA to propose a simple model that completes our account of the simple name forgetting episode (1), and (12) and (13), respectively.[5]

Adding a neural level allows us to offer *rudimentary* accounts of desiderata (i) to (iv) from section 1, in particular a gradual emergence of forgetting. The current model is simplified as a brain model in a variety of aspects: no WM (so no short-term learning); consolidation is assumed to happen; there is no coupling between dialogue cognitive states; perfect perception/communication is assumed—no processing of vision or language is integrated into the account.

The model represents certain perceptual input (visual and linguistic) and resultant memory traces as semantic pointers. It models recollection of an entity's property P (e.g., x's name) as (i) finding the vector most similar to the current percept and (ii) unbinding the entity bound to P. If recollection is successful, (a) the entity found is updated with the information originating with the current percept and (b) a smile is triggered,[6] otherwise a frown is

[2] *Vector Symbolic Architectures* (VSA; Gayler, 2004) define symbolic operations on high-dimensional numerical vectors. See Schlegel et al. (2022) for a very useful survey of Vector Symbolic Architectures.

[3] *Circular convolution* $C = A \circledast B$ is defined as in (i) in a space of dimension D, whereas the inverse of a vector is defined as in (ii), and we use the notation B' for B^{-1}

(i) **Circular convolution**
$c_j = \sum_{k=0}^{D-1} b_k a_{j-k(modD)}$
for $j \in \{0, \ldots, D-1\}$

(ii) **Inverse for circular convolution**
$a_j^{-1} = a_{D-j(modD)}$
where $j \in \{0, \ldots, D-1\}$

In other words: $\langle a_0, a_1, \ldots, a_{D-1} \rangle^{-}1 = \langle a_0, a_{D-1}, \ldots, a_1 \rangle$

[4] https://www.nengo.ai/

[5] The code for the model is available here: https://github.com/aluecking/Swanns-Name. Note that you might obtain numbers that differ from those given in this paper due to the random initialization of vectors.

[6] In a more detailed model, the motor neurons responsible for the action sequence responsible for a smile would be

triggered.

Initially there is input about three girls, Swann ('girl_52'), Anais ('girl_53'), and Chloé ('girl_54'). Swann and Anais are differentiated in terms of their names and food preferences and have the same hair colour and are neighbours (of the observer), whereas Chloé has Swann's properties bundled with being her sister:[7]

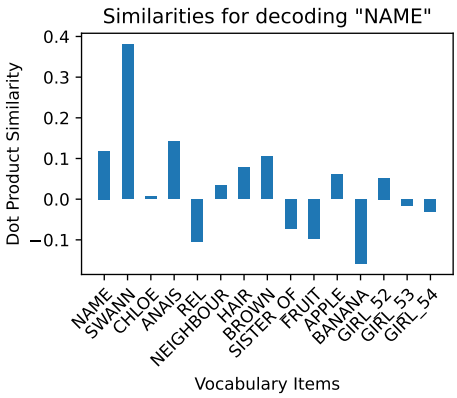
(14) a. GIRL_{52} = NAME ⊛ SWANN + REL ⊛ NEIGHBOUR + HAIR ⊛ BROWN + FRUIT ⊛ APPLE

b. GIRL_{53} = NAME ⊛ ANAIS + REL ⊛ NEIGHBOUR + HAIR ⊛ BROWN + FRUIT ⊛ BANANA

c. GIRL_{54} = GIRL_{52} + SISTER_OF + NAME ⊛ CHLOE

At this point, the state views Chloé and Swann as similar (their dot product is 0.59), and recalls Swann's name (the vector associated with the name SWANN is most similar to the decoded vector with a dot product of 0.38), as indicated in (15) for decoding NAME and in Fig. 2, where the most similar items when unbinding all its properties are shown:

(15) The name "Swann" is recalled:

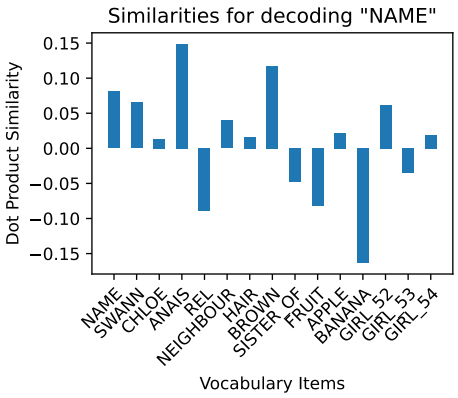


Subsequently there is input about Swann solely with respect to her hair and being a neighbour:

(16) REL ⊛ NEIGHBOUR + HAIR ⊛ BROWN + FRUIT ⊛ APPLE

This has the effect that the entity representing Swann has the properties associated with her hair and neighbourliness boosted. At this point, the state does not recall Swann's name (its similarity is below the threshold), as shown in (17) for NAME and in Fig. 3 for all properties.

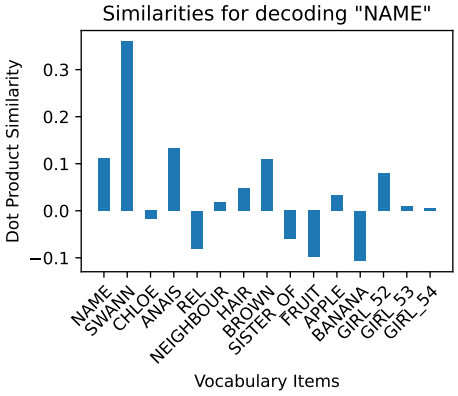
(17) The name "Anais" would be wrongly recalled, although very weakly (it is below the forgetting threshold of 0.3):



In terms of the sources of forgetting collected at the end of subsection 2.3 we can think of this as modeling forgetting by weight decay due to modification during activation.

Subsequently there is visual input about Chloé; Chloé and Swann remain similar, in other words Swann is associated (triggered as a possible topic) Finally, there is verbal input of Swann's name, which leads to it being recalled again as her name – see (18) and Fig. 4.

(18) The name "Swann" is regained:



5 Conclusions and Future Work

In this paper we have argued with reference to several concrete examples that dialogical semantics needs to be brain-oriented to account for a number of fundamental properties of cognition including forgetting and memory associativity. We have offered an initial synthesis of dialogue semantics where cognitive states are expressed in terms of external entities, though formulated with attention to the brain's memory structure, with a vector-based semantics that can be compiled into neurons and

triggered.

[7] All vectors are normalized, i.e., of unit length.

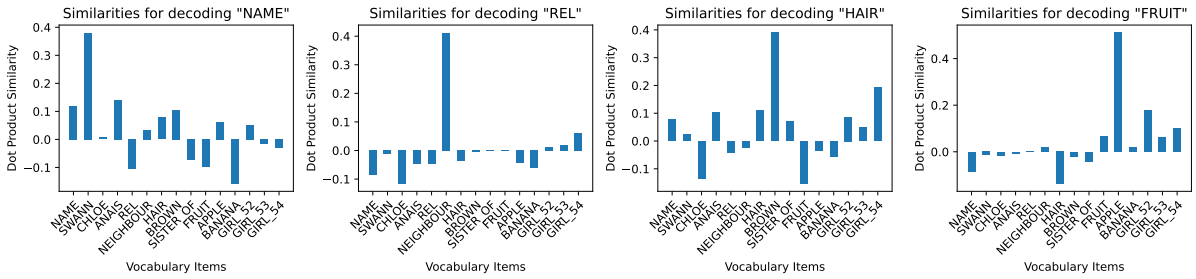


Figure 2: Unbinding the properties of the initial semantic pointer girl_52

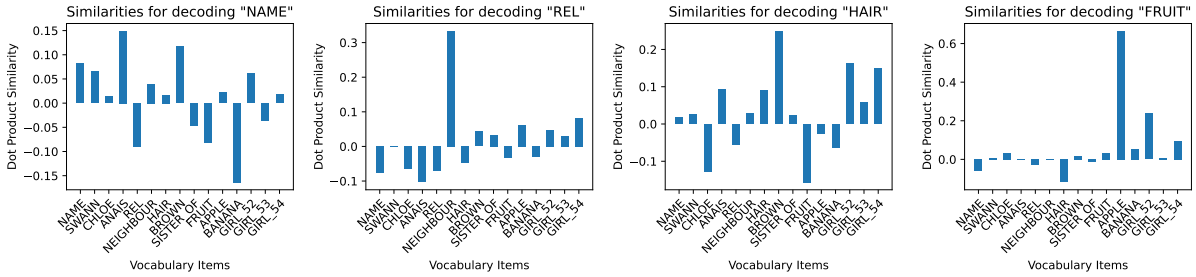


Figure 3: Unbinding the properties of girl_52 after updating REL and HAIR, but not NAME: the name 'Swann' counts as forgotten since it is not the most similar item any more and is below a similarity score of 0.3

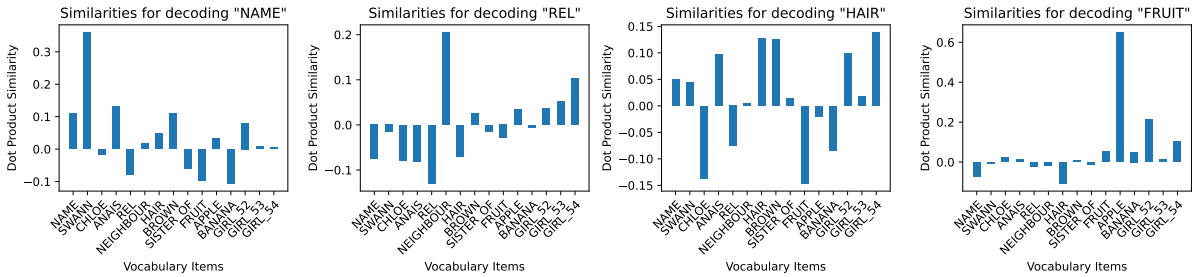


Figure 4: Unbinding the properties of girl_52 after updating NAME: the name is regained

neuron networks. The explanation we offer for the example we deal with in detail shows the need for a model that operates at various distinct levels, both the external and the neural. It is important to emphasize that such a model will clearly not be modular. For instance, our rule concerning associative topics makes reference to both a level of external content and to the neural level—more precisely the level where associations need to be computed, but the neural level is probably the more plausible level for this.

The neural model utilized here is very simplified, as we have pointed out, bypassing perception and working memory, in contrast to various existing work using the SPA architecture—see Borst et al. (2023) for a model demonstrating biological plausibility through the use of spiking neurons, and accounting for both human behavior and neuroimaging data across a whole task. In future work we hope to incorporate utterance processing and perception; an initial task being to provide neuralized versions of conversational rules.

Acknowledgments

Many thanks to Robin Cooper and Staffan Larsson for discussion and to three anonymous reviewers for TrentoLogue for very useful comments. We gratefully acknowledge support by the French *Investissements d'Avenir-Labex EFL* program (ANR-10-LABX-00) and by the *Deutsche Forschungsgemeinschaft* (DFG), grant number 502018965.

References

Alan Baddeley. 1988. Cognitive psychology and human memory. *Trends in neurosciences*, 11(4):176–181.

Alan Baddeley. 2012. Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63:1–29.

Giosuè Baggio. 2018. *Meaning in the brain*. MIT Press.

Christine Bastin, Gabriel Besson, Jessica Simon, Emma Delhaye, Marie Geurten, Sylvie Willems, and Eric Salmon. 2019. An integrative memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences*, pages 1–66.

William Bechtel. 2007. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Psychology Press.

Trevor Bekolay, James Bergstra, Eric Hunsberger, Travis DeWolf, Terrence Stewart, Daniel Rasmussen, Xuan Choo, Aaron Voelker, and Chris Eliasmith. 2014. Nengo: a Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, 7.

Peter Blouw, Eugene Solodkin, Paul Thagard, and Chris Eliasmith. 2016. Concepts as semantic pointers: A framework and computational model. *Cognitive Science*, 40(5):1128–1162.

Jelmer P Borst, Sean Aubin, and Terrence C Stewart. 2023. A whole-task brain model of associative recognition that accounts for human behavior and neuroimaging data. *PLOS Computational Biology*, 19(9):e1011427.

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

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2 edition, chapter 12, pages 375–407. Wiley-Blackwell, Oxford, UK.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114.

Chris Eliasmith. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford University Press, Oxford.

Chris Eliasmith and Carter Kolbeck. 2015. Marr's attacks: On reductionism and vagueness. *Topics in Cognitive Science*, 7(2):323–335.

Chris Eliasmith, Terrence C. Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, and Daniel Rasmussen. 2012. A large-scale model of the functioning brain. *Science*, 338(6111):1202–1205.

Jonathan R Epp, Rudy Silva Mera, Stefan Köhler, Sheena A Josselyn, and Paul W Frankland. 2016. Neurogenesis-mediated forgetting minimizes proactive interference. *Nature communications*, 7(1):10838.

Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College, London.

Ross W Gayler. 2004. Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. *arXiv preprint cs/0412059*.

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

Jonathan Ginzburg and Andy Lücking. 2020. On laughter and forgetting and reconversing: A neurologically-inspired model of conversational context. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (WeSSLLI)*, Brandeis University.

Jonathan Ginzburg and Andy Lücking. 2022. The integrated model of memory: a dialogical perspective. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue (DubDial)*, Dublin Technical University.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1):1–51.

Jan Gosmann and Chris Eliasmith. 2021. CUE: A unified spiking neuron model of short-term and long-term memory. *Psychological Review*, 128(1):104–124.

Daniel L. Greenberg and Mieke Verfaellie. 2010. Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological society*, 16(5):748–753.

Peter Hagoort. 2020. The meaning-making mechanism(s) behind the eyes and between the ears. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190301.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts, Amherst.

Ray Jackendoff. 2002. *Foundations of Language*. Oxford University Press, Oxford, UK.

Eric R Kandel, Yadin Dudai, and Mark R Mayford. 2014. The molecular and systems biology of memory. *Cell*, 157(1):163–186.

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

Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and Andy Lücking. 2023. TTR at the SPA: Relating type-theoretical semantics to neural semantic pointers. In *Proceedings of the Fourth Workshop Natural Logic meets Machine Learning*.

Donald G. MacKay, Laura W. Johnson, and Chris Hadley. 2013. Compensating for language deficits in amnesia ii: H.M.'s spared versus impaired encoding categories. *Brain Sciences*, 3(2):415–459.

John Macnamara and Gonzalo E. Reyes, editors. 1994. *The Logical Foundations of Cognition*. Number 4 in Vancouver Studies in Cognitive Science. Oxford University Press, New York.

Emar Maier. 2016. Attitudes and mental files in discourse representation theory. *Review of Philosophy and Psychology*, 7(2):473–490.

David Marr. 1982. *Vision*. Freeman, San Francisco.

Brenda Milner and Denise Klein. 2016. Loss of recent memory after bilateral hippocampal lesions: memory and memories—looking back and looking forward. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(3):230–230.

Dennis Norris. 2017. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992.

Tony Plate. 1991. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, IJCAI'91, pages 30–35.

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

Jamie Reilly, Jonathan E. Peelle, Amanda Garcia, and Sebastian J. Crutch. 2016. Linking somatic and symbolic representation in semantic memory: the dynamic multilevel reactivation framework. *Psychonomic Bulletin & Review*, 23:1002–1014.

Blake A. Richards and Paul W. Frankland. 2017. The persistence and transience of memory. *Neuron*, 94(6):1071–1084.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136. Reprinted in Semantics and Pragmatics, 2012.

Daniel Saumier and Howard Chertkow. 2002. Semantic memory. *Current Neurology and Neuroscience Reports*, 2:516–522.

Emanuel Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge.

Kenny Schlegel, Peer Neubert, and Peter Protzel. 2022. A comparison of vector symbolic architectures. *Artificial Intelligence Review*, 55(6):4523–4555.

William Beecher Scoville and Brenda Milner. 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):1121.

Melanie J Sekeres, Gordon Winocur, and Morris Moscovitch. 2018. The hippocampus and related neocortical structures in memory transformation. *Neuroscience letters*, 680:39–53.

Pieter A. M. Seuren. 2009. *Language from within: Vol. 1. Language in cognition*. Oxford University Press, Oxford.

Larry R. Squire and John T. Wixted. 2011. The cognitive neuroscience of human memory since H.M. *Annual Review of Neuroscience*, 34:259–288.

Greg J Stephens, Lauren J Silbert, and Uri Hasson. 2010. Speaker–listener neural coupling underlies successful communication. *Proceedings of the national academy of sciences*, 107(32):14425–14430.

Timothy J Teyler and Jerry W Rudy. 2007. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus*, 17(12):1158–1169.

Paul Thagard, Laurette Larocque, and Ivana Kajić. 2023. Emotional change: Neural mechanisms based on semantic pointers. *Emotion*, 23(1):182.

Endel Tulving. 1972. Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of memory*. Academic Press, New York.

Victoria I Weisz and Pablo F Argibay. 2012. Neurogenesis interferes with the retrieval of remote memories: forgetting in neurocomputational terms. *Cognition*, 125(1):13–25.

Laughter in Dialogues with Normal-Hearing and Hearing-Impaired Children: Do they all laugh alike?

Chiara Mazzocconi [1,2,3], Céline Hidalgo[2], Roxane Bertrand[1,3], Leonardo Lancia[1], Stéphane Roman[4], Daniele Schon[2,3]

[1] Aix Marseille University, CNRS, LPL, Aix-en-Provence, France
[2]Aix Marseille University, INSERM, INS, Inst Neurosci Syst, Marseille, France
[3]Aix Marseille University, ILCB, Aix-en-Provence, France
[4]La Timone Children's Hospital, ENT Unit, Marseille, France

Correspondence: chiara.mazzocconi@univ-amu.fr

Abstract

Despite the technological advancements, children with prostheses or cochlear implants, even when early implanted, show heterogeneous language skills and often struggle with pragmatic communication aspects. In our study, we focus on exploring laughter use and responsiveness to others' laughter in dialogue, comparing Normal-Hearing (N=13) and Hearing-Impaired (N=9) children while engaged in a series of conversational tasks with an adult experimenter. We observe significant differences between groups in the amount of conversational tasks complete in the allocated time and in terms of laughter frequency, speech-laughter and laughter mimicry occurrences. We discuss the observations on children and adult behaviour in relation to previous literature in adult-adult and child-caregiver interaction. Our results support the hypothesis that laughter use and responsiveness in dialogue might be related to pragmatic competences and informative about conversational quality.

1 Introduction

Despite the technological advancements, Hearing-Impaired children (HI) with prostheses or cochlear implants, even when early implanted, show heterogeneous language skills and often struggle with pragmatic communication aspects (Nicholas and Geers, 2006; Crowe and Dammeyer, 2021; Matthews and Kelly, 2022; Most et al., 2010). Difficulties in the pragmatics aspects of conversation negatively impact the quality of conversations, and are correlated with lower quality of life in school (Haukedal et al., 2022) and emotional problems (Holzinger and Fellinger, 2022). In the current work, we aim to explore an aspect of conversation which has received very little attention: the production and response to laughter during conversation in HI children. The interest in laughter arises from studies showing its crucial role in managing interactions, conveying meaning, establishing and maintaining relationships, being highly sophisticated from a pragmatic perspective (e.g. Glenn (2003); Mazzocconi et al. (2020); Dunbar (2022)), and informative about pragmatic abilities (Reddy et al., 2002; Mazzocconi and Ginzburg, 2023; Hoicka et al., 2017; Reddy, 2008). In Section 1.1, we review some literature about the pragmatic difficulties faced by HI children; in Section 1.2, we motivate our interest in laughter, highlighting its link to pragmatic competences and its role in their development and dialogue unfolding; in Section 1.3, we state the aim of our study while in Section 2 we present the corpus analysed and the methodology applied for annotation and analysis. In Section 3, we present our results and conclude by discussing them in relation to previous research in Section 4.

1.1 Pragmatics of dialogue in Hearing-Impaired (HI) children

Pragmatic abilities rely on a combination of linguistic skills, social-cognitive capacities, and executive functions (Matthews et al., 2018), including inhibition, cognitive flexibility, and working memory (Blain-Brière et al., 2014), as well as the capability to derive inferences from linguistic, behavioural and contextual cues (Goodman and Frank, 2016). Previous research has attempted to characterise the communicative difficulties faced by HI children using standardised batteries or by evaluating conversational dynamics (e.g. turn-taking, dialogue acts performed, explicitation of referents, contingency and topic-change etc.). Despite some inconsistencies in the results, likely due to small sample sizes, task differences, and

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

varying ages of implantation, most studies report significant differences in the pragmatic domain, even when phonological, syntactic and lexical skills are within the normal range for the child age (Crowe and Dammeyer, 2021; Matthews and Kelly, 2022; Most et al., 2010). Studies focusing on conversation have highlighted atypicalities in turn-taking, initiating topics, maintaining conversation, repairing and clarification requesting (Paatsch and Toe, 2014; Church et al., 2017; Most et al., 2010; Toe and Paatsch, 2013). Similar difficulties have also been found in narrative skills, in terms of coherence, and expliciting referents taking into account the eventual lack of common ground (Boons et al., 2013; Crosson and Geers, 2001; Toe and Paatsch, 2018). More generally Tuohimaa et al. (2023) reported inferential difficulties in a wide range of tasks, including theory of mind, verbal and visual information, and understanding conversational norms and emotions in context.

Most studies until now, especially those focused on evaluating the effects of using cochlear implants, have focused on the structural linguistic speech aspects of communication (Caselli et al., 2012; Church et al., 2017; Nicastri et al., 2014; Paatsch and Toe, 2014; Tye-Murray, 2003). More recently, scholars started to investigate other expressive channels contributing multimodally to the emergence of meaning and the unfolding of the dialogue (Perniss, 2018; Vigliocco et al., 2014; Holler and Levinson, 2019): such as facial expressions, gestures and prosody (Ambrose, 2016; Le Maner-Idrissi et al., 2020; Socher et al., 2019; Panzeri et al., 2021). In particular, Le Maner-Idrissi et al. (2020) observe lower performances in HI children with cochlear implants (age 5;3 – 13 years) in the ability to comprehend emotional speech on the basis of prosody as compared to NH children. Socher et al. (2019) observed specific difficulties in the non-verbal aspects of communication (including prosody, facial expressions recognition and attributing mental states and feelings to other people). A comprehensive assessment of linguistic and pragmatic abilities in Italian children with cochlear implants from a young age has confirmed several of the previously mentioned patterns (Parola et al., 2023): in general, HI children have lower performances than NH children, and difficulties are particularly marked in the paralinguistic scale (evaluating the comprehension and production of several non-verbal cues) and the contextual scale (evaluating the child's ability to use appropriate communicative behaviours with respect to different social contexts). The difficulties mentioned, can lead to misinterpretations and social awkwardness, negatively impacting social integration (Vissers and Hermans, 2018; Haukedal et al., 2022). Consequently, HI children might experience social isolation and have fewer opportunities for peer interactions (Hintermair, 2008, 2011), which are critical for developing social competence (Most, 2007), feeding therefore a vicious cycle (Bat-Chava and Deignan, 2001).

Some authors hypothesise that the difficulties HI children face in the pragmatic aspects of communication may be attributed to the increased cognitive load and fatigue required to process auditory information compared to their normal-hearing peers (Pisoni, 2000; Marsella et al., 2017; Rönnberg et al., 2010). The signal children receive indeed, in particular if fitted with cochlear implants, does not replicate the one received by NH peers, often resulting in less clear auditory input (Henry et al., 2021). To sum up, the literature reviewed highlights how the difficulties faced by HI children, go beyond mere speech, encompassing: atypical turn-taking patterns, conversational coherence, managing misunderstandings, pragmatic inferences, and emotion recognition via prosody or facial expression.

1.2 Laughter and Pragmatic skills

Laughter is a ubiquitous vocalisation in our interactions (Bryant and Bainbridge, 2022; Scott et al., 2014). It is often related to the appreciation of humour, but it is also a tool for managing conversation dynamics (e.g. turn-taking and topic-change) (Jefferson et al., 1977; Ludusan and Wagner, 2022; Bonin et al., 2015; Holt, 2010), providing feedback, showing agreement, expressing emotions, disambiguating interactants' intentions (e.g. smoothing, softening criticism) and speakers' meaning (e.g., marking irony, scare-quoting) (Glenn and Holt, 2013; Mazzocconi et al., 2020; Ginzburg et al., 2020; Glenn, 2003; Attardo et al., 2003; Hoicka and Gattis, 2008), crucial for defining group boundaries, building and maintaining relation-

ships (Jefferson, 1984; Morisseau et al., 2017; Davila-Ross and Palagi, 2022; Dunbar, 2022).

Several scholars have highlighted how laughter can be a means to track cognitive and pragmatic development in babies and older children (Piaget, 1945; McGhee, 1977; Mireault and Reddy, 2016). Even just laughing at something funny evolves drastically during development, being informative about the patterns learnt: concerning world knowledge, language and social and cultural conventions (Mireault and Reddy, 2016; Hoicka et al., 2017; Telli and Hoicka, 2022). Most scholars identify the presence of incongruity as one of the fundamental components of humour (Raskin, 1985; Attardo and Raskin, 1991; Yus, 2017; Maraev et al., 2021; Tannen, 1993; Mazzocconi et al., 2020). Therefore, appreciating humour in events is informative about the general inferential patterns exploited (Mazzocconi and Priego-Valverde, 2023; Ginzburg et al., 2020). Even just by observing others' laughter (and eventually sharing it) in different settings, children learn about cultural norms and context-specific communication practices, through social referencing (Semrud-Clikeman and Glass, 2010). Given the amount of inferential abilities and playing with shared and implicit information needed in humour (Flamson and Barrett, 2008; Cunningham, 2005), maybe not surprisingly studies show a correlation between humour comprehension and pragmatic skills (Aykan and Nalçacı, 2018; Bischetti et al., 2023).

Moreover, laughter use in relation to non-humourous events (e.g. laughter accompanying criticism, embarrassment, and asking a favour) emerges later in development, being correlated with the amount of shared attention on the object of the mothers' laughter and correlated to the acquisition of socio-cultural knowledge and pragmatic skills (Mazzocconi and Ginzburg, 2023). Even responsiveness to the interlocutor's laughter, increasing through the early years, has been suggested to be a marker of pragmatic development (Reddy et al., 2002; Mazzocconi and Ginzburg, 2022). Laughter seems therefore to be fundamental to socio-pragmatic development, while at the same time, in its use and responsiveness, being informative about it (Mireault and Reddy, 2016).

Laughter, far from being a reflex-like response, is therefore a more complex phenomenon than what is commonly thought. It can be used to disambiguate speech-acts, and can be crucial in interpreting speakers' intentions and meaning (Bryant, 2016). Since laughter can be informative about interactants' appraisals and attentional, cognitive, and emotional states (Mazzocconi et al., 2020), it is a valuable means for managing, commenting, and monitoring the conceptual alignment needed for conversation (Gandolfi et al., 2022). In particular, laughter is often related to the appraisal of some incongruities in the conversational or situational context, and the ability to interpret each other's laughter requires sharing (or at least inferring) general inferential patterns (Breitholtz, 2014), i.e., *topoi*, exploited by the interlocutor (Ginzburg et al., 2020). Laughter mimicry[1] can therefore be a precious signal for interactants, since it can effectively show meta-cognitive alignment on the evaluation of situations, propositions, or stances. Conversely, the lack of laughter mimicry in some situations can be a sign of misalignment in an evaluation or stance, or signal a lack of background in shared knowledge (Jefferson, 1979; Ginzburg and Mazzocconi, 2020). Similarly, unexpected laughter production can signal misalignment, prompting clarification requests, commentaries, or further discussions (Mazzocconi et al., 2018). Moreover, laughter mimicry is influenced by several "pragmatic" factors: context (Bryant, 2020), the interactional partner (Smoski and Bachorowski, 2003), the object of the laughter (Mazzocconi et al., 2020), e.g., it is not appropriate to reciprocate any type of laughter (Jefferson et al., 1977), and the developmental stage of the interactants (Nwokah et al., 1994; Mazzocconi and Ginzburg, 2022). The fact that laughter (mimicry) is tightly linked to pragmatic skills is also supported by studies showing atypical patterns, both in terms of occurrences and acoustic features, in neuro-different populations where pragmatic skills are characteristically divergent, such as for people in the Autistic Spectrum or with schizophrenic traits (Samson, 2013; Reddy et al., 2002; Jones, 2009;

[1]With the term *mimicry* we signify to the re-production of a behaviour shortly after a partner's one that is identical in certain dimensions, as used in Mayo and Gordon (2020) and El Haddad et al. (2019), and reviewed in Chartrand and Lakin (2013).

Polimeni and Reiss, 2006; Helt et al., 2019; Lavelle et al., 2018; Hudenko et al., 2009).

1.3 Current study

On the basis of the literature review presented, the aim of the current study was to investigate whether any difference would emerge in laughter use and laughter responsiveness in HI children as compared to NH children during conversation, being laughter use and responsiveness tightly linked to pragmatic abilities and being so important in the dialogue unfolding and conversation managing. We investigate an aspect of conversation which has never been addressed in the study of pragmatic communication difficulties in HI children. Based on the literature, we anticipated that HI children would face greater challenges in the pragmatically demanding conversational game and exhibit differences in laughter use compared to NH children. A confirmation of our hypotheses would corroborate the existence of a close relationship between laughter dialogic use and responsiveness and pragmatic competences, about which it can be informative.

2 Method

2.1 Corpus

Our corpus is constituted by 22 audio-recorded dyadic interactions of around 30 minutes (M= 31.51 ± 2.16) involving nine HI children and thirteen NH children engaging with an adult (female) during a referential (treasure-hunting) task, periodically alternated with role-reversal sub-tasks (e.g., child-led referential-tasks, child-storytelling).

2.2 Participants

Nine French-speaking children aged 5 to 9 years (3 girls, M = 75.2 months, sd = 14.1 months) with moderate (3), severe (3), profound deafness (3) were recruited via the Centre d'Action Médico-Sociale Précoce (CAMSP) and via the Institut Provençal de Suivi des Implantés Cochléaires (IPSIC) at the Salvator Hospital in Marseille. These children have a variety of devices, including bilateral conventional hearing aids (4), one (1) or two (3) cochlear implants, as well as a cochlear implant accompanied by a conventional prosthesis (1). They received hearing aids at different ages (M = 31.6 ±24.2 months). All of them have no additional disorders, were born from NH parents, communicate orally, and had language abilities in the norm for their age. The control group was constituted of 13 French-speaking normal-hearing children aged 5 to 9 (7 girls, M = 87.1 ± 13.6 months) with heterogeneous socio-demographic profiles equivalent to that of the experimental group. They had no known language, cognitive, neuro-developmental or sensory atypicalities or deficits.

2.3 Tasks and procedure

The child sit opposite the adult experimenter, in a quiet room. Audio from both participants is recorded with a unidirectional headset microphone connected to a ZOOM H4n digital recorder.

Main Map-tasks: A map is placed in the centre of the table so that the child and the adult can see it. On the map multiple items are drawn. These items have been selected according to their frequency of use in French according to the children age (New et al., 2001). The items included in the task were balanced between frequent items (known by the child), infrequent items (likely unknown to the child) and invented items (unknown to the child). The choice of challenging children with unfamiliar terms, was motivated by the aim of investigating the different strategies used to compensate for their lack of knowledge, as well as any conversational failures. The drawings on the map belong to nine semantic categories of items (e.g. bird, ship etc.), multiple exemplar of the same item are present, but differ in terms of physical (e.g. size and colour) or spatial (e.g. at the top, at the corner of the map) features. Participants are engaged in a treasure-hunt map task. The goal for the child is to collect enough hints to discover where the treasure is. These hints are gained by posing questions to the adult experimenter to disambiguate the target item of the category mentioned by the adult. An example instruction from the adult might be "The next hint is hidden behind the bird". As several items may correspond to this description, the child is expected to implement strategies to find the correct target item among the possible candidates within the category.

Sub-tasks The main task is periodically alternated by sub-tasks. These sub-tasks have been

included in order to help sustain attention, but especially to acquire data from different types of conversation where the roles are more balanced or even reversed as compared to the main task where the adult detains more information than the child and therefore holds a "leading role". The sub-tasks are: (1) "Guess who?": the child and the adult have the same set of cards, not shown to the partner. The child secretly chooses a card and the adult, asking questions, has to guess which of the available cards has been chosen by the child. In this task the roles are reversed as the child detains more information than the adult; (2) "Picture story": the child is given a series of three sequential pictures arranged randomly which s/he has to put back into chronological order to tell the story; (3) " Child Story telling": the adult elicits an unstructured narration asking the child about their holiday or about the plot of their favourite movie; (4) "Find the differences": the child has to find seven differences between two images, by communicating them verbally to the adult. The task is rather difficult for children who are spontaneously led to focus attention and ask the adult for help.

2.4 Laughter Annotation

All our annotations have been carried out using the software ELAN (Brugman et al., 2004). The coding was carried out by one annotator listening to each audio-file until a laugh occurred. The coder then marked the onset and offset of the laugh, distinguishing between laughter not overlapping or overlapping with speech (Laughter/Speech-laughter). Our criteria for laughter identification and annotation are in line with previous work, though adapted since we relied exclusively on the auditory modality (e.g., El Haddad et al. (2019); Mazzocconi and Ginzburg (2022)).[2] Our study focuses on: the occurrence of laughter (frequency), duration, position in relation to speech (laughter/speech-laughter), and to the partner's laughter (Non-/Mimicking). For the purposes of this paper, a *Mimicking* laugh (produced by interlocutor B) refers to any laugh that shortly follows the onset of a *Non-mimicking* laugh (produced by interlocutor A). The following describes our method for identifying Mimicking and Non-Mimicking laughs, where A_i and B_j are the i^{th} and j^{th} laughs produced by interlocutors A and B, respectively, T_{start} and T_{stop} are the start and stop times, respectively, and ΔT is set to 1 second. In order for laugh B_j to mimic laugh A_i, B_j must occur after the *start* time of A_i (1) with an onset before the *stop* time of A_i with a margin ΔT. To avoid duplication, B_i must stop before the start time of laugh A_{i+1} (2).

(1) $T_{start}(A_i) < T_{start}(B_j)$

(2) $T_{start}(B_j) < min\{T_{stop}(A_i) + \Delta T, T_{start}(A_{(i+1)})\}$

Inter-annotator agreement was assessed having a second coder for 20% of the conversations (covering 30% of the laughs annotated by the first annotator). For segmentation (onset-offset) we observed an average degree of organisation of 0.74 (Staccato algorithm, Lücking et al. (2011)).[3] The observed labelling agreement on matched annotations was 98% and Cohen's kappa was 0.9.

3 Results

3.1 Task-completion

In the given time (about 30 minutes), all NH children (N=13) completed all the tasks (mean time 1896.06 sec, sd 95.8 sec), while only 5 out of 9 HI children completed all the tasks (mean time 1892.55 sec, sd 177.60 sec). The other 4 HI children (44%) did not manage to complete the last two tasks in the allocated time.

3.2 Laughter frequency and duration

Over the full corpus, 830 laughs were identified and annotated: 669 in the 13 NH dyads (376 produced by children) and 161 in the 9 HI dyads (89 produced by children). Figure 1 represents the counts of laughter occurrences for each participant. Means and standard deviations for laughter occurrences, laughter duration and laughter frequency over 10 mins by Participant and Group are reported in Table 1. Laughter is overall significantly more frequent in the NH group than in the HI group ($W = 86$, $p < .001$). Nevertheless, the frequency of laughter is not significantly different between HI children and NH children ($W = 34$, $p = 0.11$), while for Adults in the NH group laughter frequency is significantly higher than for adults interacting with HI children ($W = 2$, $p < .001$).

[2] Annotation protocol at https://osf.io/mbv8z.

[3] This is a measure based on Thomann (2001, p.243). It ranges in the interval (-1, 1). A value of 0 corresponds to the agreement expected from random annotations.

Group	N	Participant	Laughter Count	Laughter Mean (sd)	Duration (sec) Mean (sd)	Freq/10min Mean (sd)	Speech-laugh %/Tot.
HI	9	Adult	72	8 (2.96)	0.97 (0.45)	2.58 (1.03)	12.50%
HI	9	Child	89	9.89 (10.7)	1.00 (0.76)	3.13 (3.39)	22.47%
NH	13	Adult	293	22.5 (6.60)	0.91 (0.47)	7.20 (2.34)	14.68%
NH	13	Child	376	31.3 (33.7)	0.89 (0.60)	9.28 (11.3)	40.96%

Table 1: Mean and standard deviation of laughter occurrences, duration and frequency over 10 mins according to Group (HI: Hearing Impaired; NH: Normal Hearing) and Participant (Adult; Child)

Group	Participant	Total L. Mean (sd)	Non-Mimicking L. Mean (sd)	Mimicking L. Mean (sd)	% Mimicking Mean (sd)
HI	Adult	8 (2.96)	7.44 (2.88)	1.25 (0.5)	6.58 (8.25)
HI	Child	9.89 (10.7)	9 (9.99)	2 (1.41)	7.76 (10.86)
NH	Adult	22.5 (6.60)	18.9 (5.01)	5.22 (4.94)	14.9 (16.11)
NH	Child	28.9 (33.4)	25.8 (30.9)	6 (5.08)	18.46 (17.31)

Table 2: Laughter Mimicry distribution and Transitional Probabilities (HI: Hearing Impaired; NH: Normal Hearing) and Participant (Adult; Child)

We run a linear mixed-effect model on laughter duration, having *Group* and *Participant* as fixed effects, and *Dyad* as random factor. Our analysis did not reveal any significant difference in terms of duration neither between Groups (β = -0.07, se = 0.10, df = 33.73, t = -0.74, p = 0.46), nor between Participants (β -0.11, se = 0.08, df = 815.97, t = -1.36, p = 0.17), nor in their interaction (β = 0.08, se = 0.09, df = 817.96, t = 0.90, p = 0.37).

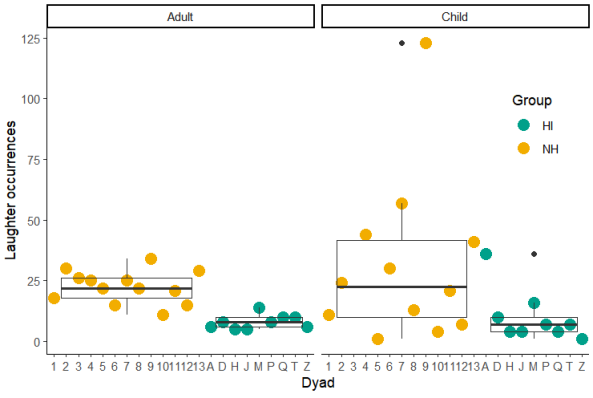


Figure 1: Laughter occurrences by Dyad, Group and Participant

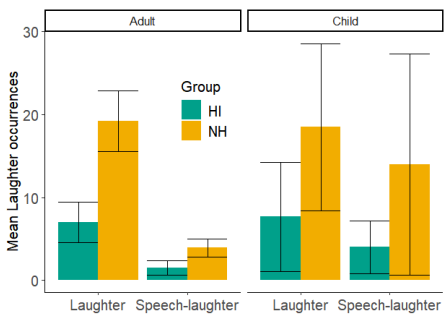


Figure 2: Mean laughter and speech-laughter occurrences by Group and Participant

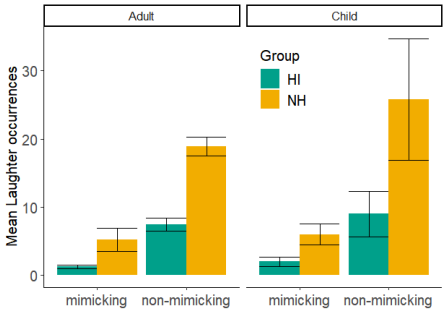


Figure 3: Mean Mimicking and Non-Mimicking laughter occurrences by Group and Participant

3.3 Speech-laughter

We observe a significant difference in the occurrence of speech-laughter between the two groups (χ^2 = 8.0, df = 1, p-value < .005). While no difference is observed in Adults according to Group (χ^2 = 0.08, df = 1, p-value = 0.78), we observe HI children to produce significantly less speech-laughter compared to NH children (χ^2 = 9.72, df = 1, p-value = 0.001).

3.4 Laughter Mimicry

In Table 2 we report means and standard deviations for the occurrences and percentages of Mimicking laughter by *Group* and *Participant*. Despite the high inter-individual variability (Figure 3), statistical testing shows that Mimicking laughter occurrences are overall rarer in the HI group than in the NH group (χ^2 = 7.83, p = .005). Significantly fewer occurrences of Laughter Mimicking are observed in the HI group both for adults (χ^2= 3.91, p <.05) and children (χ^2 = 3.94, p < .05).

4 Discussion

We investigated for the first time laughter occurrences and responsiveness to the partner's laughter in Hearing-Impaired (HI) children while engaged in different conversational tasks as compared to Normal-Hearing (NH) children. The aim of our study was to test HI and NH children performance in a pragmatically charged conversational game and whether they differ in their laughter behaviour in conversation. We overall observed difficulties for the HI children to complete the conversational tasks in the allocated time and different patterns of laughter behaviour across groups confirming our hypothesis. Interestingly, we observe also some significant differences in the adult behaviour depending on whether she was interacting with NH or HI children.

4.1 Laughter frequency

The first striking result is that laughter is significantly less frequent in the HI group, and especially so for the adult (Tab. 1). Interestingly, the frequency of laughter production observed in the adult interacting with a HI child is more similar to those observed in mother-infant interaction (e.g. Nwokah et al. (1994); Mazzocconi and Ginzburg (2022)) than those observed in adult interaction: friendly conversations 5.8 (± 2.5)/10 min (Vettin and Todt, 2004); speed-dating 21(± 9.28)/5 min (Fuchs and Rathcke, 2018); friendly loosely-controlled conversation French 45/10 min, Chinese 26/10 min; fully ecological and diverse contexts BNC 5/10 min (Mazzocconi et al., 2020). For children, the frequency does not result significantly different between groups, especially due to the high variability in the NH group and the considerable overlap in the distribution observed (Fig. 1). It is nevertheless interesting to remark that while values at the high extreme of the distribution are all NH children, those on the lower extreme of the distribution are predominantly HI children. The fact that frequencies are lower (and balanced) in the HI dyads as a whole, highlights the fact that the dyad works as an organic system where the interactants influence each other (Fusaroli et al., 2014).

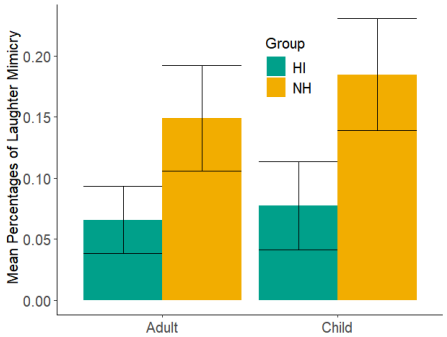


Figure 4: Mean percentages of laughter mimicry

4.2 Speech-laughter

We observe HI children to produce significantly less speech-laughter in comparison to NH children. This is interesting when considering that speech-laughter emerges rather late in development (for most children absent even at 36 months, (Mazzocconi and Ginzburg, 2022)) despite laughter emerging around the third month of life (Sroufe and Wunsch, 1972; Nwokah et al., 1994; Oller et al., 2021) and speech being present since the second year of life. Mazzocconi and Ginzburg (2022) proposed two possible explanations for the late and rare use of speech-laughter in infants: speech-laughter might require quite advanced language abilities, as a matter either of vocal control and coordination, or the development of laughter's pragmatic use to shape (and evaluate) verbal contributions. Interestingly, only the NH children display a percentage of speech-laughter production similar to those observed in adult-adult dyadic naturalistic interaction (e.g. French: 31%, Chinese: 47%, BNC: 30%, Mazzocconi et al. (2020), 50% Trouvain (2001); 60% O'Connell and Kowal (2005); 58% Devillers and Vidrascu (2007)). The adult therefore displays significantly lower percentages of speech-laughter when interacting with children participating in our study (regardless of the group) than in adult-adult conversation. This observation could be attributed to the semi-controlled nature of the interaction, where the adult experimenter engages in the same task with various children following a rather scripted flow. Moreover, she assumes the role of a speech and language therapist, which may lead her, particularly for HI children, to deliberately make her speech as controlled and clean as possible.

4.3 Laughter mimicry

We observe fewer occurrences of laughter mimicry in the HI dyads and significantly lower percentages of mimicking laughter in HI

children as compared to NH children. Different hypotheses can be put forward to explain these patterns. On the one hand, the lower occurrences of laughter mimicry might be a signal of lower conceptual alignment. Laughter is indeed an ambiguous signal highly context-dependent, while at the same time it is highly informative about speakers' mental states and general inferential patterns exploited (and can be a precious means to manage conversations and monitor (and signal) conceptual alignment (Gandolfi et al., 2022). Alternatively, the lack of laughter alignment might be explained considering that HI children face a higher cognitive load in engaging in speech processing in interaction (Marsella et al., 2017), which according to Pickering and Garrod (2013, 2004) would impact the predictions made by interlocutors and the motivation to affiliate or communicate. On the other, following Giles et al. (1991, 2023)'s communication action theory lower alignment might derive from lower motivation to affiliate. This might derive from the fact that the experimental conditions of our data collection might highly resemble those of a speech and language therapy session, an activity to which HI children are extremely highly exposed and which might even elicit a distancing effect. Finally, based on studies showing that initiating laughs (those that are reciprocated by the interactant with laughter mimicry) have peculiar acoustic characteristics as opposed to those that are not reciprocated (Truong and Trouvain (2012); Mazzocconi et al. (2024), it is possible that HI children are not able to pick up on these subtle modulations of the signal, due to the distorted quality of the sound perceived (Pisoni, 2000; Henry et al., 2021), and therefore do not interpret them as an invitation to join the laugh. It is worth noting that, contrarily to what is observed in NH children, for the adult (in both groups) and for HI children percentages of laughter mimicry are significantly lower than those observed in other adult-adult interactions (around 35% across languages and contexts (Mazzocconi et al., 2020; Vettin and Todt, 2004; Smoski, 2004)). A possible interpretation of this is that laughter occurs when mutual comprehension, and laughter interpretation therefore, are granted, and especially laughter mimicry can be used to show similar stances and appraisals. HI interactions are particularly delicate because mutual comprehension (based on alignment) cannot always be given for granted. In general, the lower percentages observed in the adult while interacting with children compared to those observed in adult-adult interaction might be related either to a misalignment in the appraisal of laughables, or to the fact that the adult is avoiding distractions for the child, attempting to help sustain attention in a cognitively demanding task.

5 Conclusion

Overall, we observe difficulties for the HI children group in completing the pragmatically demanding conversational tasks in the allocated time and significant differences between HI and NH children in conversational laughter use and responsiveness: frequency, speech-laughter and mimicking laughter. Considering the literature review highlighting communicative pragmatic difficulties in HI children, these results endorse the view that laughter behaviour might be linked to pragmatic competences and socio-cognitive development (Reddy, 2008; Mireault and Reddy, 2016; Mazzocconi and Ginzburg, 2022, 2023). Laughing while speaking and aligning to the interactional partner's laughter indeed, requires a complex mechanism involving not only understanding but also an evaluative attitude on own discourse or on the partner's discourse. To validate the hypothesis that laughter behaviour can be informative about pragmatic competences and conversational quality, further analyses will test correlations between laughter behaviour, pragmatic competence conversation measures (appropriate responsiveness to dialogue acts, quality of the strategies used to accomplish the tasks), convergence at different levels (pitch, intensity, syllabic rate), turn-taking timing dynamics and speaking time balance. Moreover, additional analyses looking at the laughter's arguments and pragmatic functions might better elucidate whether the differences observed are also related to underpinning differences in laughter pragmatic use to manage the dialogue unfolding, meaning, face-threatening acts and rapport.

6 Acknowledgments

This work was supported by the French National Agency for Research (ANR) under the reference ANR-16-CONV-0002—MUSCOORD Project. It was carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002) and has also benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). We are grateful to the three TrentoLogue reviewers for their constructive feedback.

References

Sophie E Ambrose. 2016. Gesture use in 14-month-old toddlers with hearing loss and their mothers' responses. *American journal of speech-language pathology*, 25(4):519–531.

Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3-4):293–348.

Simge Aykan and Erhan Nalçacı. 2018. Assessing theory of mind by humor: The humor comprehension and appreciation test (tom-hcat). *Frontiers in psychology*, 9:382586.

Yael Bat-Chava and Elizabeth Deignan. 2001. Peer relationships of children with cochlear implants. *Journal of Deaf Studies and Deaf Education*, 6(3):186–199.

Luca Bischetti, Irene Ceccato, Serena Lecce, Elena Cavallini, and Valentina Bambini. 2023. Pragmatics and theory of mind in older adults' humor comprehension. *Current Psychology*, 42(19):16191–16207.

Bénédicte Blain-Brière, Caroline Bouchard, and Nathalie Bigras. 2014. The role of executive functions in the pragmatic skills of children age 4–5. *Frontiers in psychology*, 5:81239.

Francesca Bonin, Nick Campbell, and Carl Vogel. 2015. The discourse value of social signals at topic change moments. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Tinne Boons, Leo De Raeve, Margreet Langereis, Louis Peeraer, Jan Wouters, and Astrid Van Wieringen. 2013. Narrative spoken language skills in severely hearing impaired school-aged children with cochlear implants. *Research in developmental disabilities*, 34(11):3833–3846.

Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, University of Gothenburg.

Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating multi-media/multi-modal resources with elan. In *LREC*, pages 2065–2068. Lisbon.

Gregory A Bryant. 2016. How do laughter and language interact. In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Gregory A Bryant. 2020. Evolution, structure, and functions of human laughter. In *The handbook of communication science and biology*, pages 63–77. Routledge.

Gregory A Bryant and Constance M Bainbridge. 2022. Laughter and culture. *Philosophical Transactions of the Royal Society B*, 377(1863):20210179.

Maria Cristina Caselli, Pasquale Rinaldi, Cristiana Varuzza, Anna Giuliani, and Sandro Burdo. 2012. Cochlear implant in the second year of life: Lexical and grammatical outcomes.

Tanya L Chartrand and Jessica L Lakin. 2013. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308.

Amelia Church, Louise Paatsch, and Dianne Toe. 2017. Some trouble with repair: Conversations between children with cochlear implants and hearing peers. *Discourse studies*, 19(1):49–68.

Jillian Crosson and Ann Geers. 2001. Analysis of narrative ability in children with cochlear implants. *Ear and Hearing*, 22(5):381–394.

Kathryn Crowe and Jesper Dammeyer. 2021. A review of the conversational pragmatic skills of children with cochlear implants. *The Journal of Deaf Studies and Deaf Education*, 26(2):171–186.

J Cunningham. 2005. Children's humor. *Children's play. SAGE publications*.

Marina Davila-Ross and Elisabetta Palagi. 2022. Laughter, play faces and mimicry in animals: evolution and social functions. *Philosophical Transactions of the Royal Society B*, 377(1863):20210177.

Laurence Devillers and Laurence Vidrascu. 2007. Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Interdisciplinary workshop on the phonetics of laughter*, page 37.

RIM Dunbar. 2022. Laughter and its role in the evolution of human social bonding. *Philosophical Transactions of the Royal Society B*, 377(1863):20210176.

Kevin El Haddad, Sandeep Nallan Chakravarthula, and James Kennedy. 2019. Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles. In *2019 International Conference on Multimodal Interaction*, pages 259–263.

Thomas Flamson and H Clark Barrett. 2008. The encryption theory of humor: A knowledge-based mechanism of honest signaling. *Journal of Evolutionary Psychology*, 6(4):261–281.

Susanne Fuchs and Tamara Rathcke. 2018. Laugh is in the air? In *Proceedings of Laughter Workshop 2018, Paris, France*, pages 21–24.

Riccardo Fusaroli, Joanna Raczaszek-Leonardi, and Kristian Tylén. 2014. Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157.

Greta Gandolfi, Martin J Pickering, and Simon Garrod. 2022. Mechanisms of alignment: shared control, social cognition and metacognition. *Philosophical Transactions of the Royal Society B*, 378(1870):20210362.

Howard Giles, Nikolas Coupland, and IUSTINE Coupland. 1991. Accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

Howard Giles, America L Edwards, and Joseph B Walther. 2023. Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, 99:101571.

Jonathan Ginzburg and Chiara Mazzocconi. 2020. Laughing about laughter: comparing conversational analysis, emotion psychology, and dialogical semantics. In *Laughter and Other Non-Verbal Vocalisations Workshop: Proceedings (2020)*.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1).

Phillip Glenn. 2003. *Laughter in interaction*, volume 18. Cambridge University Press.

Phillip Glenn and Elizabeth Holt. 2013. *Studies of laughter in interaction*. A&C Black.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Christiane Lingås Haukedal, Ona Bø Wie, Stefan K Schauber, Björn Lyxell, Elizabeth M Fitzpatrick, and Janne von Koss Torkildsen. 2022. Social communication and quality of life in children using hearing aids. *International journal of pediatric otorhinolaryngology*, 152:111000.

Molly S Helt, Deborah A Fein, and Jacob E Vargas. 2019. Emotional contagion in children with autism spectrum disorder varies with stimulus familiarity and task instructions. *Development and psychopathology*, pages 1–11.

Fergal Henry, Martin Glavin, and Edward Jones. 2021. Noise reduction in cochlear implant signal processing: A review and recent developments. *IEEE reviews in biomedical engineering*, 16:319–331.

Manfred Hintermair. 2008. Self-esteem and satisfaction with life of deaf and hard-of-hearing people—a resource-oriented approach to identity work. *Journal of deaf studies and deaf education*, 13(2):278–300.

Manfred Hintermair. 2011. Health-related quality of life and classroom participation of deaf and hard-of-hearing students in general schools. *Journal of deaf studies and deaf education*, 16(2):254–271.

Elena Hoicka, Jessica Butcher, Felicity Malla, and Paul L Harris. 2017. Humor and preschoolers' trust: Sensitivity to changing intentions. *Journal of experimental child psychology*, 154:113–130.

Elena Hoicka and Merideth Gattis. 2008. Do the wrong thing: How toddlers tell a joke from a mistake. *Cognitive Development*, 23(1):180–190.

Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Elizabeth Holt. 2010. The last laugh: Shared laughter and topic termination. *Journal of Pragmatics*, 42(6):1513–1525.

Daniel Holzinger and Johannes Fellinger. 2022. Conversation difficulties rather than language deficits are linked to emotional problems in school children with hearing loss. In *Frontiers in Education*, volume 7, page 944814. Frontiers Media SA.

William J Hudenko, Wendy Stone, and Jo-Anne Bachorowski. 2009. Laughter differs in children with autism: An acoustic analysis of laughs produced by children with and without the disorder. *Journal of autism and developmental disorders*, 39(10):1392–1400.

G. Jefferson. 1979. A technique for inviting laughter and its subsequent acceptance/declination. *Everyday language: Studies in ethnomethodology*, 79:96.

Gail Jefferson. 1984. On the organization of laughter in talk about troubles. *Structures of social action: Studies in conversation analysis*, 346:369.

Gail Jefferson, Harvey Sacks, and Emanuel Schegloff. 1977. Preliminary notes on the sequential organization of laughter. *Pragmatics Microfiche*, 1:A2–D9.

Errin Jones. 2009. *Humour and laughter in children with Autism Spectrum Disorders*. Ph.D. thesis, University of Ballarat.

Mary Lavelle, Christine Howes, Patrick GT Healey, and Rosemarie McCabe. 2018. Are we having a laugh? conversational laughter in schizophrenia. *DaP 2018*, page 35.

Gaïd Le Maner-Idrissi, Sandrine Le Sourn Bissaoui, Virginie Dardier, Maxime Codet, Nathalie Botte-Bonneton, Fanny Delahaye, Virginie Laval, Marc Aguert, Géraldine Tan-Bescond, and Benoit Godey. 2020. Emotional speech comprehension in deaf children with cochlear implant. *Psychology of Language and Communication*, 24(1):44–69.

Andy Lücking, Sebastian Ptock, and Kirsten Bergmann. 2011. Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann. In *International Gesture Workshop*, pages 129–138. Springer.

Bogdan Ludusan and Petra Wagner. 2022. Laughter entrainment in dyadic interactions: temporal distribution and form. *Speech Communication*, 136:42–52.

Vladislav Maraev, Ellen Breitholtz, Christine Howes, Staffan Larsson, and Robin Cooper. 2021. Something old, something new, something borrowed, something taboo: Interaction and creativity in humour. *Frontiers in Psychology*, 12:654615.

Pasquale Marsella, Alessandro Scorpecci, Giulia Cartocci, Sara Giannantonio, Anton Giulio Maglione, Isotta Venuti, Ambra Brizi, and Fabio Babiloni. 2017. Eeg activity as an objective measure of cognitive load during effortful listening: A study on pediatric subjects with bilateral, asymmetric sensorineural hearing loss. *International journal of pediatric otorhinolaryngology*, 99:1–7.

Danielle Matthews, Hannah Biney, and Kirsten Abbot-Smith. 2018. Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Language Learning and Development*, 14(3):186–223.

Danielle Matthews and Ciara Kelly. 2022. Pragmatic development in deaf and hard of hearing children: A review. *Deafness & Education International*, 24(4):296–313.

Oded Mayo and Ilanit Gordon. 2020. In and out of synchrony—behavioral and physiological dynamics of dyadic interpersonal coordination. *Psychophysiology*, 57(6):e13574.

Chiara Mazzocconi and Jonathan Ginzburg. 2022. A longitudinal characterization of typical laughter development in mother–child interaction from 12 to 36 months: Formal features and reciprocal responsiveness. *Journal of Nonverbal Behavior*, 46(4):327–362.

Chiara Mazzocconi and Jonathan Ginzburg. 2023. Growing up laughing: Laughables and pragmatic functions between 12 and 36 months. *Journal of Pragmatics*, 212:117–145.

Chiara Mazzocconi, Vladislav Maraev, and Jonathan Ginzburg. 2018. Laughter repair. In *Proceedings of SemDial 2018 (AixDial), The 22nd workshop on the Semantics and Pragmatics of Dialogue, Aix-en-Provence (France)*.

Chiara Mazzocconi, Benjamin O'Brien, Kubra Bodur, and Abdellah Fourtassi. 2024. Do children laugh like their parents? conversational laughter mimicry occurrence and acoustic alignment in middle-childhood. Under Review, Preprint.

Chiara Mazzocconi and Béatrice Priego-Valverde. 2023. Humour in early interaction: what it can tell us about the linguistic, pragmatic and cognitive development of the child. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*.

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

Paul E McGhee. 1977. A model of the origins and early development of incongruity-based humour. In *It's a funny thing, humour*, pages 27–36. Elsevier.

Gina C Mireault and Vasudevi Reddy. 2016. *Humor in infants: developmental and psychological perspectives*. Springer.

Tiffany Morisseau, Martial Mermillod, Cécile Eymond, Jean-Baptiste Van Der Henst, and Ira A Noveck. 2017. You can laugh at everything, but not with everyone: What jokes can tell us about group affiliations. *Interaction Studies*, 18(1):116–141.

Tova Most. 2007. Speech intelligibility, loneliness, and sense of coherence among deaf and hard-of-hearing children in individual inclusion and group inclusion. *Journal of Deaf Studies and Deaf Education*, 12(4):495–503.

Tova Most, Ella Shina-August, and Sara Meilijson. 2010. Pragmatic abilities of children with hearing loss using cochlear implants or hearing aids

compared to hearing children. *Journal of Deaf Studies and Deaf Education*, 15(4):422–437.

Boris New, Christophe Pallier, Ludovic Ferrand, and Rafael Matos. 2001. Une base de données lexicales du français contemporain sur internet: Lexique™//a lexical database for contemporary french: Lexique™. *L'année psychologique*, 101(3):447–462.

Maria Nicastri, Roberto Filipo, Giovanni Ruoppolo, Marika Viccaro, Hilal Dincer, Letizia Guerzoni, Domenico Cuda, Ersilia Bosco, Luca Prosperini, and Patrizia Mancini. 2014. Inferences and metaphoric comprehension in unilaterally implanted children with adequate formal oral language performance. *International Journal of Pediatric Otorhinolaryngology*, 78(5):821–827.

Johanna Grant Nicholas and Ann E Geers. 2006. Effects of early auditory experience on the spoken language of deaf children at 3 years of age. *Ear and hearing*, 27(3):286–298.

Evangeline E Nwokah, Hui-Chin Hsu, Olga Dobrowolska, and Alan Fogel. 1994. The development of laughter in mother-infant communication: Timing parameters and temporal sequences. *Infant Behavior and Development*, 17(1):23–35.

D Kimbrough Oller, Gordon Ramsay, Edina Bene, Helen L Long, and Ulrike Griebel. 2021. Protophones, the precursors to speech, dominate the human infant vocal landscape. *Philosophical Transactions of the Royal Society B*, 376(1836):20200255.

Daniel C O'Connell and Sabine Kowal. 2005. Laughter in bill clinton's my life (2004) interviews. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 15(2):275–299.

Louise E Paatsch and Dianne M Toe. 2014. A comparison of pragmatic abilities of children who are deaf or hard of hearing and their hearing peers. *Journal of deaf studies and deaf education*, 19(1):1–19.

Francesca Panzeri, Sara Cavicchiolo, Beatrice Giustolisi, Federica Di Berardino, Paola Francesca Ajmone, Paola Vizziello, Veronica Donnini, and Diego Zanetti. 2021. Irony comprehension in children with cochlear implants: The role of language competence, theory of mind, and prosody recognition. *Journal of Speech, Language, and Hearing Research*, 64(8):3212–3229.

Alberto Parola, Dize Hilviu, Sara Vivaldo, Andrea Marini, DI Diego, Patrizia Consolino, and Francesca Marina Bosco. 2023. Development of communicative-pragmatic abilities in children with early cochlear implants. *Journal of Child Language*, pages 1–17.

Pamela Perniss. 2018. Why we should study multimodal language. *Frontiers in psychology*, 9:342098.

Jean Piaget. 1945. *Play, dreams, and imitation in childhood*. New York: Norton.

Martin J Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212–225.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

David B Pisoni. 2000. Cognitive factors and cochlear implants: Some thoughts on perception, learning, and memory in speech perception. *Ear and hearing*, 21(1):70–78.

Joseph Polimeni and Jeffrey P Reiss. 2006. Humor perception deficits in schizophrenia. *Psychiatry Research*, 141(2):229–232.

V. Raskin. 1985. *Semantic mechanisms of humor*, volume 24. Springer.

Vasudevi Reddy. 2008. *How infants know minds*. Harvard University Press.

Vasudevi Reddy, Emma Williams, and Amy Vaughan. 2002. Sharing humour and laughter in autism and down's syndrome. *British journal of psychology*, 93(2):219–242.

Jerker Rönnberg, Mary Rudner, Thomas Lunner, and Adriana A Zekveld. 2010. When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health*, 12(49):263–269.

Andrea Christiane Samson. 2013. Humor (lessness) elucidated–sense of humor in individuals with autism spectrum disorders: Review and introduction. *Humor*, 26(3):393–409.

Sophie K Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. 2014. The social life of laughter. *Trends in cognitive sciences*, 18(12):618–620.

Margaret Semrud-Clikeman and Kimberly Glass. 2010. The relation of humor and child development: Social, adaptive, and emotional aspects. *Journal of Child Neurology*.

Moria Smoski. 2004. *The development of antiphonal laughter between friends and strangers*. Ph.D. thesis, Vanderbilt University.

Moria Smoski and Jo-Anne Bachorowski. 2003. Antiphonal laughter between friends and strangers. *Cognition & Emotion*, 17(2):327–340.

Michaela Socher, Björn Lyxell, Rachel Ellis, Malin Gärskog, Ingrid Hedström, and Malin Wass. 2019. Pragmatic language skills: A comparison of children with cochlear implants and children without hearing loss. *Frontiers in psychology*, 10:475839.

L Alan Sroufe and Jane Piccard Wunsch. 1972. The development of laughter in the first year of life. *Child Development*, pages 1326–1344.

Deborah Tannen. 1993. What's in a frame? surface evidence for underlying expectations. *Framing in discourse*, 14:56.

Burcu Soy Telli and Elena Hoicka. 2022. Humor and social cognition: Correlational and predictive relations in 3-to 47-month-olds. *Cognitive Development*, 64:101245.

Bruno Thomann. 2001. Observation and judgment in psychology: Assessing agreement among markings of behavioral events. *Behavior Research Methods, Instruments, & Computers*, 33:339–348.

Dianne Toe and Louise Paatsch. 2018. Communicative competence of oral deaf children while explaining game rules. *The Journal of Deaf Studies and Deaf Education*, 23(4):369–381.

Dianne M Toe and Louise E Paatsch. 2013. The conversational skills of school-aged children with cochlear implants. *Cochlear implants international*, 14(2):67–79.

Jürgen Trouvain. 2001. Phonetic aspects of "speech-laughs". In *Proceedings of the 2nd Conference on Orality and Gestuality*, pages 634–639.

Khiet P Truong and Jürgen Trouvain. 2012. On the acoustics of overlapping laughter in conversational speech. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Krista Tuohimaa, Soile Loukusa, Heikki Löppönen, Taina Välimaa, and Sari Kunnari. 2023. Development of social-pragmatic understanding in children with congenital hearing loss and typical hearing between the ages of 4 and 6 years. *Journal of Speech, Language, and Hearing Research*, 66(7):2503–2520.

Nancy Tye-Murray. 2003. Conversational fluency of children who use cochlear implants. *Ear and Hearing*, 24(1):82S–89S.

Julia Vettin and Dietmar Todt. 2004. Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution.

CTWM Vissers and Daan Hermans. 2018. Social-emotional problems in deaf and hard-of-hearing children from an executive and theory of mind perspective. *Evid Based Pract Deaf Educ*, 28:455–76.

Francisco Yus. 2017. Incongruity-resolution cases in jokes. *Lingua*, 197:103–122.

Laughter in the cradle: A taxonomy of infant laughables

Yingqin Hu[1,2], Capucine Brillet[1], Bosko Rajkovic[4], Gauhar Rustamova[1], Chiara Mazzocconi[3], Catherine Pelachaud[2], Jonathan Ginzburg[1],

[1] CNRS, Université Paris-Cité, Laboratoire Linguistique Formelle (LLF)
[2] Sorbonne Université, Institut des Systèmes Intelligents et de Robotiques (ISIR)
[3] Aix-Marseille Université, Institute of Language, Communication and the Brain (ILCB)
[4] Independent Scholar

Abstract

This paper proposes a taxonomy for the laughables (events giving rise to laughter) of the child in their first year of life. We hypothesize that a child's laughables (within the first year) may relate to the following factors: sensory stimulation, cognitive challenges, and social interaction. We use Piaget's theory as a starting point for explicating the cognitive basis of the laughables, taking into account much subsequent literature. To test our hypothesis, we ran two longitudinal corpus studies using the Rollins Corpus and the SAYCam Corpus. On the basis of the results obtained, we developed a taxonomy of laughables. We believe this to be the most detailed empirical study of laughables hitherto conducted in research on child laughter.

1 Introduction

Understanding a baby's laughter is a complex yet crucial aspect of developmental psychology that has been historically overlooked (Addyman and Addyman, 2013; Addyman, 2020). Infant laughter serves as a form of communication and bonding between parent and child, reflecting cognitive and emotional development (Sauter et al., 2018). However, babies lack the ability to verbally express their thoughts, making it difficult to understand the context of their laughter. Without a clear context, interpreting the meaning behind their laughter becomes challenging. Additionally, infants are at a stage where their cognitive and emotional development is evolving radically. This developmental process can impact on the causes underlying their laughter, adding to the complexity of interpretation (Mireault and Reddy, 2016; Mazzocconi and Ginzburg, 2023). Therefore, further exploration of the factors contributing to infant laughter is necessary to gain a deeper understanding of its significance and implications.

The structure of the paper is as follows: in section 2 we review previous work, summarize potential factors that may elicit laughter, and discuss the feasibility of using Piaget's schema theory to explain laughables surrounding cognition. In section 3, we present our research questions and the objectives of this paper. In section 4, we explain how we classify laughables using data from two longitudinal corpus studies. Then, in section 5, we introduce a more comprehensive taxonomy based on the results of these studies (Section 5.1) and discuss inter-annotator agreement (IAA) (Section 5.2). Finally, in section 6, we summarize our findings.

2 Related Work

2.1 Sensorimotor stage

According to Piaget's cognitive development theory (Piaget et al., 1952), babies in their first year remain in the sensorimotor stage, a period characterized by the development of basic motor skills through perception and interaction with their environment via physical sensations and body coordination. At this stage, children progress from simple reflexes in response to sensory stimuli to actively exploring their environment and the objects within it. Through repeated actions, they start to understand the notion of *cause-and-effect*, e.g., realizing that crying loudly will draw a caregiver's attention or that pressing a button will make a toy produce sounds. The stage culminates in the understanding of *object permanence*—objects continue to exist even when they are out of sight.

2.2 Sensory stimulation

Sroufe and Wunsch (1973) observed that infants begin to laugh around four months of age. Initially, their laughter is primarily triggered by tactile or auditory stimuli, such as light touches on sensitive areas or high-pitched voices. These triggers become less potent over time, while more visual and social

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

stimuli become increasingly successful in eliciting laughter in the first year of life. From 5 to 8 weeks, babies are most responsive to dynamic visual stimulation, such as a nodding head. Other studies also indicate that babies exhibit strong responses to rhythmic, high-pitched voices, and moving objects across the first year (Slater et al., 1985; Singh et al., 2002; Kitamura and Burnham, 2003). Therefore, from the perspective of sensory stimulation lacking communicative significance, the sensory stimuli for laughter change over time, to include physical, visual, and auditory stimuli.

2.3 Cognition

As rapid cognitive development ensues, the sources of laughter are not limited to sensory stimulation; they begin to include laughter based on cognitive understanding: in Shultz and Zigler (1970)'s study, a stationary clown emerges as a more effective stimulus than a dynamic clown (dynamic visual stimulation) for 3-month-old babies. Why do they laugh at a stationary clown? We relate this to a view of adult laughter developed by (Ginzburg et al., 2020). Two basic meanings are postulated for laughter, one involving the person laughing to express her enjoyment of the laughable l, the other expressing her perception of l as being *incongruent*. Building on work in humour theory (Raskin, 1985), incongruity can be explicated as a notion that relates a contextually salient entity l with a defeasible rule (a *topos* τ (Breitholtz, 2020)) which represents normal expectations in case there exists a contextually salient characterization of l that is incompatible with τ. In accordance with this theory, if we use incongruity to explain why three-month-old babies laugh when staring at a stationary clown, we assume that infants have certain expectations/cognition about human faces. The clown's face clashes with these expectations, causing the baby to laugh. Nonetheless, the question remains— why do babies derive pleasure from the incongruity?

2.3.1 Violation of a Schema

Piaget and other researchers posit that this pleasure is derived from a *cognitive challenge*, whereby a young child finds that they require effort to make sense of incongruent events using their existing knowledge, referred to as the *schema* in Piaget's theory (Piaget, 2013; Berlyne, 1960; Harter, 1974, 1978; Schultz, 1976; McGhee and Pistolesi, 1979). Piaget believed that individuals organize their knowledge into mental schemas, which help them to understand the world around them. These schemas include expectations about how objects, actions, and events should occur based on past experiences. Infants' expectations are formed by various schema types, including event schemas, self-schemas, object schemas, and role schemas. When they are born, they have innate schemas, such as grasping and sucking, to interpret and engage with their environment. As they grow, these schemas evolve and become more intricate. For instance, consider a child who encounters a dog for the first time. When shown a picture of a dog by their mother, the child forms a schema associating dogs with specific features like two ears, four legs, and a tail. Subsequently when a child sees a dog with only one ear instead of two, this conflicts with their schema of what a dog should look like.

Other researchers have argued that two necessary conditions must be met for children to appraise an event amusing when it violates their existing schemas. First, the child must be in an environment perceived as safe (Sroufe et al., 1974; Baillargeon et al., 1985; Mireault and Reddy, 2020). Second, the deviations/novelty should neither be too difficult nor too easy relative to the child's current knowledge. Instead, it should require an optimal amount of effort to understand, within their zone of proximal development (Vygotsky and Cole, 1978).

2.3.2 Exploration of New Schemas

Exploration itself can be a source of pleasure for babies. As early as the first year of life, children develop a strong sense of curiosity about their environment and themselves. Piaget argued that this "need" to explore novelty is an inherent part of a child's nervous system. For instance, when a baby encounters a new object and has not yet understood what it is or what it can do, they may engage in exploratory behaviors such as observing or patting the object (Piaget, 2013; Mc Reynolds, 1962; Hutt, 1966; Belsky et al., 1980; Bijou, 1980; Gibson, 1988; Rochat, 1989).

2.3.3 Conformity to a Schema

In addition, Piaget and other researchers have suggested that the pleasure babies derive is not limited to cognitive challenges or curiosity but also extends to a sense of recognition and mastery. McGhee and Pistolesi (1979) exemplify two situations in which babies experience a sense of recognition and mastery: social play and object play.

An example of social play is the game of peek-a-

boo, where a caregiver hides the baby's eyes and then reveals them while saying "peek-a-boo!". The first-time experience presents two novel events for the baby: the event schema (having vision blocked and then restored) and the concept of object permanence. After repeated play, the baby will eventually laugh when the caregiver removes their hands, as this action now conforms to the baby's existing event schema.

Object play often occurs when a baby visually examines and manipulates a novel object, such as a new toy. Unlike the pleasure derived from curiosity, this type of pleasure is elicited when the baby begins to understand the properties of the object and becomes less curious about it. For example, a baby may learn the function of an interactive toy or the concept of cause-and-effect by discovering that when they pat a toy pig, it responds with a pig sound. As with social play, the baby may laugh the moment the toy pig makes the sound, as it confirms their hypothesis.

2.3.4 Application of a Schema

At this stage, a baby is also actively involved in the emergence of pleasure (for themselves and others) by applying their schemas. Although schemas are not well-developed at birth, they gradually become refined and expanded through adaptation, which is a key process described in Piaget's schema theory. This adaptation can occur through either assimilation or accommodation. Assimilation occurs when the baby integrates new information into their existing schemas, while accommodation happens when new information alters or replaces their existing schemas. In this way, babies can incorporate novelty or incongruity into their current understanding. They might use their schemas to create joy in two different situations.

The first situation is social play. For example, after playing peek-a-boo multiple times with their mother, the baby becomes familiar with the event schema of peek-a-boo. As a result, when the mother covers the baby's eyes with her hands, the baby may start laughing in anticipation, having already predicted the mother's next action—removing her hands (an event that brings the baby true joy). Another possibility is that the baby uses the event schema of peek-a-boo to play a prank on the mother; for instance, by deliberately grabbing the mother's hands when she covers the baby's eyes (Trevarthen and Hubley, 1978; Reddy, 1991; Nomikou et al., 2017).

Another situation is object play. It has often been observed that a baby laughs when they see a toy they frequently play with, which excludes the possibility of curiosity about the toy, as it is already familiar to it. Piaget suggests that this laughter may be due to an affective response toward the object. He argues that there is as much construction in the affective domain as in the sensorimotor stage. This means that the construction of an object schema involves not only knowledge about the properties of the object but also emotional responses toward it. Thus, the toy evokes a sense of pleasure in the baby when they see it. Another hypothesis proposed by McGhee and Pistolesi (1979) is the function of make-believe play. For example, as described by Garvey (Garvey, 1990), instead of simply playing with a toy car, a baby might imagine the car in a race with themselves as the driver or pretend that the car is a spaceship. Humour would not be triggered by such play until attention shifts to the fact that the child is imagining the car doing something they know to be nonsensical, absurd, or impossible.

2.4 Social Interaction

Laughter can also occur in non-humorous forms, serving as a flexible social signal (McGhee and Pistolesi, 1979). In a similar fashion to how infants construct schemas for objects, Piaget argued that children also develop schemas of social interaction. The process of constructing this social schema can be considered a form of *social effort*, driven by an interest in others and *social reciprocity*, which involves spontaneous mutual engagement and the valuing of connections with others. Laughter may be a part of this social schema, helping to maintain attachment with caregivers. For instance, children may use laughter to elicite caregiver's positive caregiving gesture like patting or stroking (Ainsworth, 1967; Bowlby, 1982; Nelson, 2012). Laughter can also have a crucial role in learning how to direct others' attention and in establishing moments of shared attention in the child-caregiver dyad (Mazzocconi and Ginzburg, 2022; Parnell, 2023), which is a crucial building block in the neuropsychological development of a child, correlated with their later language and socio-communicative abilities (e.g. Lasch et al. (2023) Finally, laughter can be also a coping strategy to respond to a caregiver's laughter (El Haddad et al., 2019; Mazzocconi et al., 2023).

3 Research Questions

Based on our literature review, we hypothesize that a child's laughter may be triggered by three types of events: events involving sensory stimulation, cognitively demanding events (violation of a schema, exploration of new schemas, conformity to a schema, and application of a schema), and social interaction, where laughter serves as a flexible social signal.

This raises the following two questions: first, can these potential factors be used to exhaustively classify the laughable events in the first year of life? Second, can these factors be integrated in a precise way within Piaget's schema theory?

To address these questions, we conducted two longitudinal corpus studies analyzing the contexts of laughable events within the first year of life.

4 Method

4.1 The Corpus

To gather contextual data on baby laughter and to assess performance differences between laboratory and familiar environments, we conducted a longitudinal study using the Rollins Corpus (Rollins, 2003; Trautman and Rollins, 2006; Rollins and Trautman, 2011; Rollins and Greenwald, 2013) and the SAYCam Corpus (Sullivan et al., 2021).

4.1.1 Rollins Corpus

The Rollins corpus comprises a collection of longitudinal video recordings capturing the development of 61 infants from 3 months to 30 months of age and recorded in the laboratory.

Participating children were exclusively exposed to English as their primary language and minimal exposure to other languages (i.e., less than 7 hours per week).

The laboratory environment is child-friendly, equipped with two-way mirrors on both the front and back walls. During the recording sessions, parent-child pairs sat facing each other and engaged in spontaneous play using standardized age-appropriate toys (see Figure 1). Parents were encouraged to interact with their child naturally. Infants were initially seated in an infant seat with a tray for toy placement at 9 months, transitioning to seated floor play at 12 and 30 months. All sessions were recorded using split-screen video technology.

4.1.2 SAYCam Corpus

The SAYCam corpus comprises a collection of longitudinal video recordings of 3 infants aged from 6 months to 32 months, captured in various settings including their homes, cars, neighborhoods, and workplaces where the child spent time. The recording method involves the babies wearing a head-mounted camera (see Figure 2), allowing access to information from the child's perspective.

Figure 1: Observing the child from two perspectives in Rollins Corpus

All three families spoke English exclusively (see Table1). Alice and Asa are from a family that lived in the United States and Sam lived in Australia. Alice wore a headcam from 8 months to 31 months of age. Sam wore the headcam from 6 months to 30 months of age. He was diagnosed with autism spectrum disorder at age 3; as of this writing (at age 7), Sam is fully integrated into mainstream activities, has friends, and does not require any special support. Asa started wearing the headcam at 7 months. Due to the onset of the COVID-19 pandemic and the birth of a younger sibling, data collection for Asa ended at 24 months.

Table 1: Participant Information in SAYCam Corpus

Participant	Location	First recording (months)	Last recording (months)
Alice	USA	8	31
Asa	USA	7	24
Sam	Australia	6	30

Each family recorded approximately 2 hours per week, once at a fixed time and once at a randomly chosen time. All caregiver-infant activities were spontaneous and not designed.

4.2 Our Data

From the cohort listed in Table 2, in the Rollins corpus, we selected 15 children aged from 3 to 12 months. In the SAYCam corpus, we have 3 children aged from 8 to 12 months. Ultimately, we annotated 294 instances of baby laughter and 631 instances of caregiver laughter in the Rollins corpus, as well as 458 instances of baby laughter and

Figure 2: Participant (7 months old) wearing Veho camera with fish eye lens in SAYCam Corpus (Sullivan et al., 2021)

396 instances of caregiver laughter in the SAYCam corpus.

Table 2: Information about the Two Corpora

	Month	Caregiver	Child	Duration
Rollins	3	12	11	0:21:41
	6	99	44	2:50:49
	9	279	110	5:56:15
	12	241	129	5:20:06
	Total	631	294	14:28:51
SAYCam	8	67	96	5:24:49
	9	118	81	5:42:56
	10	72	106	5:35:14
	11	71	75	5:27:11
	12	68	100	6:16:22
	Total	396	458	28:26:32

4.3 Our Annotation

All our annotations were conducted using the software ELAN (Brugman and Russel, 2004). The coding was carried out by the first author and three other coders. The coders annotated both the laughter and the position of laughables within two corpora, providing natural language descriptions of the laughables. Laughter is defined as a segment starting when laughter-related auditory, facial, or bodily cues are observed, and ending with a perceived breath intake or, if absent when the facial or body movement ceases. If a breath intake occurs after a delay and the participant is still perceived as laughing, it is considered part of the laughter; otherwise, the segment concludes with the end of sound or movement. The laughable descriptions are then clustered using keywords they have in common (including their synonyms). These have been found in the videos being annotated and, hence, for now, the class of keywords used in the clustering is dependent on the data used.

4.4 Inter-annotator agreement

For the SAYCam corpus, we extracted 16% (74 instances) of laughs, and for the Rollins corpus, we extracted 23% (69 instances) of laughs. These laughs cross baby and age and the cross-annotation was performed by two other authors. The Inter-Annotator Agreement (IAA) is calculated under the same exact conditions, except that concerning laughter position.

5 Results

5.1 Laughable Taxonomy

Based on the studies mentioned in related work and Piaget's theory, we attempt to classify laughables in our dataset. We categorized laughable types into: *sensory stimulation, conformity to a schema, violation of a schema, application of a schema, exploration of new schema, and social interaction.*

Data presented in Table 3 show that most laughable events can be categorized within our schema-based taxonomy of laughables, with only 5 cases (3 in Rollins and 2 in SAYCam) not successfully explained. Secondly, half of the babies' laughter in both corpora is elicited by sensory stimulation or by encountering novel or incongruous events, which aligns well with the characteristics of the sensorimotor stage. Furthermore, there appears to be an influence from different environmental contexts. In laboratory settings, children respond more to sensory stimulation and schema violations, where caregivers play a prominent role, but less to self-directed laughable types, such as the application of a schema and social interaction. Babies in a naturalistic environment, however, exhibit a greater diversity and a more balanced distribution of laughable types.

Table 3: Distribution of Laughable Types in the Two Corpora

	Rollins		SAYCam	
Sensory Stimulation	95	32.31%	103	22.49%
Violation of a Schema	93	31.63%	64	13.97%
Conformity to a Schema	15	5.10%	28	6.11%
Exploration of New Schemas	47	15.99%	119	25.98%
Application of a Schema	14	4.76%	69	15.07%
Social Interaction	27	9.18%	73	15.94%
Other	3	1.02%	2	0.44%
Total	294	100.00%	458	100.00%

5.1.1 Sensory Stimulation

We categorized sensory stimulation into three main classes: *Physical, Auditory, and Visual.* In both

corpora, most sensory stimulation occurs in a combined form, like visuo-auditory. For example, a caregiver slowly approaches the baby while making a plosive sound ("booh!"). We only list the frequency of each type stimulus in Table 4.

The result suggests that physical stimuli, particularly tickling, are the most consistent triggers of laughter across both corpora. Auditory stimuli such as rhythmic and high-pitch sounds also play a significant role. Approaching (a person/object) is the most frequent visual stimulus in the SAYCam corpus but exhibits a lower frequency in the Rollins corpus.

Table 4: Distribution of Sensory Stimulation in the Two Corpora

		Rollins		SAYCam	
Physical	be held	0	0.00%	8	17.02%
	be kissed	1	1.85%	4	8.51%
	be lifted up	0	0.00%	6	12.77%
	be tickled	44	81.48%	13	27.66%
	be touched	8	14.81%	4	8.51%
	cannot keep balance	0	0.00%	10	21.28%
	good taste	0	0.00%	2	4.26%
	jump	1	1.85%	0	0.00%
	Total	54	100.00%	47	100.00%
Auditory	animal sound	2	5.26%	0	0.00%
	bumblebee sound	1	2.63%	0	0.00%
	clapping hands	3	7.89%	0	0.00%
	high pitch sound	4	10.53%	18	40.91%
	plosive sound	0	0.00%	2	4.55%
	rhythmic sound	20	52.63%	24	54.55%
	tickling sound	6	15.79%	0	0.00%
	whistling sound	2	5.26%	0	0.00%
	Total	38	100.00%	44	100.00%
Visual	approach	12	54.55%	15	41.67%
	be hided	1	4.55%	0	0.00%
	bumblebee sound	1	4.55%	0	0.00%
	clapping hands	4	18.18%	5	13.89%
	shaking hand	1	4.55%	1	2.78%
	shaking toy	3	13.64%	2	5.56%
	shining toy	0	0.00%	13	36.11%
	Total	22	100.00%	36	100.00%

5.1.2 Conformity to a Schema

We found 2 subcategories of conformity to a schema: *conformity to object schema* and *conformity to event schema*. Their distribution is shown in Table 5.

The primary difference between these categories is that conformity to an object schema occurs when a baby receives the expected reaction from an object after observation or repeated examination. For example, a baby pats a toy, and the toy starts singing. In contrast, the majority of cases for conformity to the event schema stem from the peek-a-boo game, wherein the caregiver obscures the child's vision with an object or their hands and then removes it while saying "peek-a-boo!". Consequently, through repeated exposure to this game, children are likely to develop an event schema for peek-a-boo, where the sequence involves having their vision obscured followed by its restoration. For example, if we consider the process of $vision\_obstructed \rightarrow vision\_restored$ as the laughable, then the start time of the laughable is the moment when the vision is obstructed, and the end time is the moment when the vision is restored. Therefore, the reaction time to this laughable is calculated as $laughter\_start\_time - laughable\_end\_time$. From the reaction time column in the Table 6, it can be observed that the baby's laughter and the caregiver's removal of hands are almost synchronous.

Table 5: Distribution of Conformity to a Schema in the Two Corpora

	Rollins		SAYCam	
Conformity to event schema	10	66.67%	20	71.43%
Conformity to object schemas	5	33.33%	8	28.57%
Total	15	100.00%	28	100.00%

Table 6: Reaction Time for Five Cases in Conformity to the Schema

Laughter		Laughable		Reaction Time (s)
Start Time	**End Time**	**Start Time**	**End Time**	
217.01	217.79	216.12	217.20	-0.19
1048.67	1049.39	1047.82	1048.67	0.00
1050.42	1050.90	1049.78	1050.42	0.00
1105.71	1106.19	1104.06	1105.71	0.00
633.48	634.46	632.02	633.47	0.00

5.1.3 Violation of a Schema

We observed 5 categories of violation of a schema. Their descriptions and examples are as follows and their distribution is shown in Table 7:

1. **Violation of facial schemas:**
   Description: Situations where the baby observes a caregiver's facial expressions deviate from the normal.
   Example: A surprised face, sticking out a tongue, opening the mouth wide open, a fierce face, a face showing discomfort, and a yawning face.

2. **Violation of object schemas:**
   Description: Situations where the intended use or characteristics of objects conflict with the established cognitive understanding.
   Example: A toy duck, typically stationary, is manipulated by the caregiver to speak with or kiss the baby.

3. **Violation of social role schemas:**
   Description: Situations where a caregiver engages in actions that do not align with their

typical role or identity.
Example: Mimicking the baby's actions or speech. When the baby screams, the caregiver also screams; when the baby opens their mouth wide, the caregiver does the same; when the baby says "bababa", the caregiver echoes "bababa".

4. **Violation of event schemas:**
Description: Occurs when the expected sequence of actions is disrupted, deviating from the established order. When the natural action sequence is known to be $Action\ A \rightarrow Action\ B$, but instead, it becomes $Action\ A \rightarrow Action\ C$.
Example: An example involves a caregiver playing a prank on the baby, such as when the baby reaches out to grab a ball, but the mother quickly picks it up and throws it away.

5. **Violation of behavior schemas:**
Description: Situations where the caregiver behaves in a manner inconsistent with the established schema of caregiver-baby interaction.
Example: A caregiver pretends not to see the baby and looks for the baby but the baby is just sitting in front of the caregiver.

Table 7: Distribution of Violation of a Schema in the Two Corpora

	Rollins		SAYCam	
Violation of behavior schemas	11	11.83%	0	0.00%
Violation of event schemas	0	0.00%	11	17.19%
Violation of facial schemas	11	11.83%	19	29.69%
Violation of object schemas	65	69.89%	22	34.38%
Violation of social role schemas	6	6.45%	12	18.75%
Total	93	100.00%	64	100.00%

5.1.4 Application of a Schema

We observed two categories of the application of a schema. Their descriptions are as follows and their distribution is shown in Table 8:

1. **Application of object schema:**
Description: Typically occurs when a child sees or receives their favorite toy.

2. **Application of event schema - Prediction:**
Description: Typically occurs when a child and caregiver have repeatedly engaged in a game with same sequence of actions $Action\ A \rightarrow Action\ B \rightarrow Action\ C \rightarrow Action\ D$, with $Action\ D$ being the truly laughable event. Once the child becomes familiar with this sequence, they tend to laugh even before $Action\ D$ occurs.

Table 8: Distribution of Application of a Schema in the Two Corpora

	Rollins		SAYCam	
Application of object schema	10	71.43%	13	18.84%
Application of event schema - Pranks	1	7.14%	26	37.68%
Application of event schema - Prediction	3	21.43%	30	43.48%
Total	14	100.00%	69	100.00%

Table 9: Distribution of Exploration of New Schema in the Two Corpora

	Rollins		SAYCam	
Explore Self Schema	1	2.13%	4	3.36%
Explore the environment	46	97.87%	115	96.64%
Total	47	100.00%	119	100.00%

3. **Application of event schema - Pranks:**
Description: Typically occurs when a child becomes familiar with the sequence of actions in a game. Assuming the sequence is $Action\ A \rightarrow Action\ B \rightarrow Action\ C \rightarrow Action\ D$, and the caregiver is the one performing these actions, the child will attempt to prevent the caregiver from performing $Action\ B$ once $Action\ A$ has been completed and $Action\ B$ is imminent.

5.1.5 Exploration of New Schemas

In Table 9, we categorized the exploration of new schemas into two types: exploring the environment and exploring self-schemas. In both datasets, infants are more engaged in exploring the environment by observing what happens after a novel event or action, such as patting or shaking objects to test the properties of unfamiliar objects. Exploration of self-schemas occurs when infants observe themselves in a mirror.

5.1.6 Social Interaction

We observed 6 categories of social interaction. Their descriptions are as follows and their distribution is shown in Table 10:

1. **Sharing:**
Description: When a baby obtains or discovers an object (denoted as A) and subsequently redirects their gaze from the object A to the caregiver, often accompanied by a gesture in-

Table 10: Distribution of Social Interaction in the Two Corpora

	Rollins		SAYCam	
Attempting to Capture Caregiver's Attention	0	0.00%	6	8.22%
Initiation of Engagement by the Caregiver	7	25.93%	19	26.03%
Invitation to Play with the Caregiver	0	0.00%	4	5.48%
Receiving Encouragement from the Caregiver	3	11.11%	5	6.85%
Receiving Friendliness from the Caregiver	14	51.85%	29	39.73%
Sharing	3	11.11%	10	13.70%
Total	27	100.00%	73	100.00%

dicating sharing, such as showing or offering the object A.

2. **Attempting to Capture Caregiver's Attention:**
   Description: When a baby notices that the caregiver's gaze is not directed towards them, they attempt to use laughter as a means to attract the caregiver's attention.

3. **Receiving Encouragement from the Caregiver:**
   Description: The caregiver typically provides encouragement through verbal utterances such as "yeah! <Baby's name>", "you did it!" accompanied by encouraging action like clapping hands.

4. **Receiving Friendliness from the Caregiver:**
   Description: The caregiver demonstrates friendliness by laughing or smiling at the baby, or by using greeting utterances such as "Hi, <Baby's name>."

5. **Invitation to Play with the Caregiver:**
   Description: Following a game with the caregiver, the baby give the game object to the caregiver, inviting them to engage in play once more.

6. **Initiation of Engagement by the Caregiver:**
   Description: While the baby is playing independently, the caregiver takes the initiative to ask or engage in the game with the baby.

5.2 Confusion on Laughable Annotation

The confusion matrix (Figure 3) shows that "sensory stimulation" and "violation of a schema" are the most divergent categories between the two annotators. For example, when a caregiver pronounces a plosive sound like "booh", it is usually accompanied by the mouth forming an exaggerated O-shape. Additionally, the category "exploration" is often confused with "conformity to a schema" or

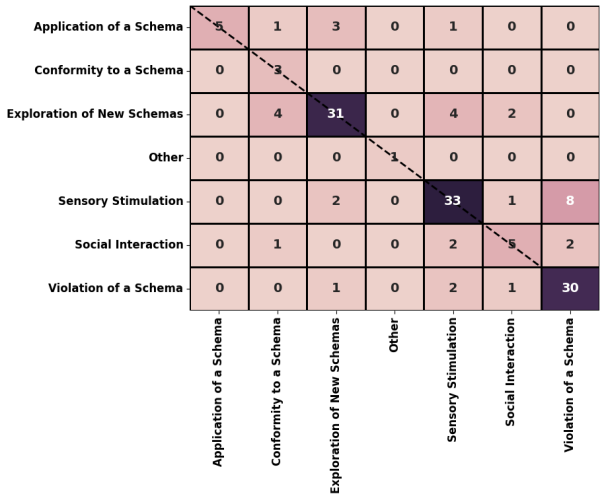


Figure 3: Confusion matrix for inter-annotator agreement (IAA) results across the two corpora, with a kappa IPF of 0.6783, a kappa max of 0.8897, and a raw agreement of 0.7552. The required minimum overlap percentage is 100%.

"sensory stimulation". For instance, when a child looks at a shiny toy, it can be interpreted in several ways: the child could be merely observing the shiny toy, understanding its function (e.g., patting the toy to make it shine), or laughing at the visual stimulation. Therefore, we argue that this divergence is unavoidable as it depends on plausible inter-subject differences in event classification.

6 Conclusions

This paper proposes a taxonomy of laughables for baby laughter, building on previous literature and evaluated on two corpora. The results demonstrate that a baby's laughables (events triggering laughter) in the first year align with our initial hypothesis, encompassing three main classes, namely *sensory stimulation, cognitive challenges, and social interaction.* Within the class of cognitive challenges we have a further, fine-grained partition into five sub-classes (*conformity to the schema, violation of a schema, application of a schema, exploration of new schema.*) inspired in part by the Piagetian notion of *schemas*. This ties in closely with a view of adult laughter meaning (Ginzburg et al., 2020) as expressing for the most part a laughable *l*'s being incongruent.

Acknowledgments

This work was supported by the French National Centre for Scientific Research (CNRS) under the reference 80PRIME2023—TELIN and by the French *Investissements d'Avenir-Labex EFL* program (ANR-10-LABX-00). We would like to thank the three TrentoLogue reviewers for their valuable comments on the first draft of this paper.

References

Caspar Addyman. 2020. *The laughing baby: the extraordinary science behind what makes babies happy*. Unbound Publishing.

Caspar Addyman and Ishbel Addyman. 2013. The science of baby laughter. *Comedy Studies*, 4(2):143–153.

Mary D Salter Ainsworth. 1967. Infancy in uganda: Infant care and the growth of love.

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. 1985. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208.

Jay Belsky, Mary Kay Goode, and Robert K Most. 1980. Maternal stimulation and infant exploratory competence: Cross-sectional, correlational, and experimental analyses. *Child development*, pages 1168–1178.

DE Berlyne. 1960. Conflict, arousal and curiosity.

Sidney W Bijou. 1980. Exploratory behavior in infants and animals: A behavior analysis. *The psychological record*, 30:483–495.

John Bowlby. 1982. Attachment and loss: retrospect and prospect. *American journal of Orthopsychiatry*, 52(4):664.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: the use of common sense reasoning in conversation*. Brill.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Kevin El Haddad, Sandeep Nallan Chakravarthula, and James Kennedy. 2019. Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles. In *2019 international conference on multimodal interaction*, pages 259–263.

Catherine Garvey. 1990. *Play*. Harvard University Press.

Eleanor J Gibson. 1988. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual review of psychology*, 39(1):1–42.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1):1–51.

Susan Harter. 1974. Pleasure derived by children from cognitive challenge and mastery. *Child development*, pages 661–669.

Susan Harter. 1978. Pleasure derived from challenge and the effects of receiving grades on children's difficulty level choices. *Child Development*, pages 788–799.

Corinne Hutt. 1966. Exploration and play in children.

Christine Kitamura and Denis Burnham. 2003. Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, 4(1):85–110.

Carolyn Lasch, Stephanie M Carlson, and Jed T Elison. 2023. Responding to joint attention as a developmental catalyst: Longitudinal associations with language and social responsiveness. *Infancy*, 28(2):339–366.

Chiara Mazzocconi, Kevin El Haddad, Benjamin O'Brien, Kübra Bodur, and Abdellah Fourtassi. 2023. Laughter mimicry in parent-child and parent-adult interaction. In *Multimodal Communication Symposium 2023*.

Chiara Mazzocconi and Jonathan Ginzburg. 2022. A longitudinal characterization of typical laughter development in mother–child interaction from 12 to 36 months: Formal features and reciprocal responsiveness. *Journal of Nonverbal Behavior*, 46(4):327–362.

Chiara Mazzocconi and Jonathan Ginzburg. 2023. Growing up laughing: Laughables and pragmatic functions between 12 and 36 months. *Journal of Pragmatics*, 212:117–145.

Paul Mc Reynolds. 1962. Exploratory behavior: A theoretical interpretation. *Psychological Reports*, 11(2):311–318.

Paul E McGhee and Edie Pistolesi. 1979. *Humor: Its origin and development*. WH Freeman San Francisco.

Gina C Mireault and Vasudevi Reddy. 2016. *Humor in infants: developmental and psychological perspectives*. Springer.

Gina C Mireault and Vasudevi Reddy. 2020. Making sense of infants' differential responses to incongruity. *Human development*, 64(2):55–63.

Judith Kay Nelson. 2012. *What made Freud laugh: An attachment perspective on laughter*. Routledge.

Iris Nomikou, Giuseppe Leonardi, Alicja Radkowska, Joanna Rączaszek-Leonardi, and Katharina J Rohlfing. 2017. Taking up an active role: emerging participation in early mother–infant interaction during peekaboo routines. *Frontiers in psychology*, 8:239785.

Vicki Parnell. 2023. " watch me, mom!" the development of infants' skills in eliciting others' attention.

Jean Piaget. 2013. *Play, dreams and imitation in childhood*. Routledge.

Jean Piaget, Margaret Cook, et al. 1952. *The origins of intelligence in children*, volume 8. International Universities Press New York.

V. Raskin. 1985. *Semantic mechanisms of humor*, volume 24. Springer.

Vasudevi Reddy. 1991. Playing with others' expectations: Teasing and mucking about in the first year.

Philippe Rochat. 1989. Object manipulation and exploration in 2-to 5-month-old infants. *Developmental Psychology*, 25(6):871.

Pamela R Rollins and Lisa C Greenwald. 2013. Affect attunement during mother-infant interaction: How specific intensities predict the stability of infants' coordinated joint attention skills. *Imagination, Cognition and Personality*, 32(4):339–366.

Pamela Rosenthal Rollins. 2003. Caregivers' contingent comments to 9-month-old infants: Relationships with later language. *Applied Psycholinguistics*, 24(2):221–234.

PR Rollins and CH Trautman. 2011. Caregiver input before joint attention: The role of multimodal input. In *International Congress for the Study of Child Language (IASCL), Baltimore, MD*.

Disa Sauter, Bronwen Evans, Dianne Venneker, and Mariska Kret. 2018. How do babies laugh? *The Journal of the Acoustical Society of America*, 144(3_Supplement):1840–1840.

Thomas R Schultz. 1976. A cognitive-developmental analysis of humor. *Humor and laughter: Theory, research and applications*, pages 12–13.

Thomas R Shultz and Edward Zigler. 1970. Emotional concomitants of visual mastery in infants: The effects of stimulus movement on smiling and vocalizing. *Journal of Experimental Child Psychology*, 10(3):390–402.

Leher Singh, James L Morgan, and Catherine T Best. 2002. Infants' listening preferences: Baby talk or happy talk? *Infancy*, 3(3):365–394.

Alan Slater, Victoria Morison, Carole Town, and David Rose. 1985. Movement perception and identity constancy in the new-born baby. *British Journal of Developmental Psychology*, 3(3):211–220.

L. Alan Sroufe and Jane Wunsch. 1973. The development of laughter in the first year of life. *Child development*, 43:1326–44.

LA Sroufe, E Waters, and L Matas. 1974. Contextual determinants of infant affective response. *The origins of fear*, 2:49–72.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. 2021. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5:20–29.

Carol Hamer Trautman and Pamela Rosenthal Rollins. 2006. Child-centered behaviors of caregivers with 12-month-old infants: Associations with passive joint engagement and later language. *Applied Psycholinguistics*, 27(3):447–463.

Colwyn Trevarthen and Penelope Hubley. 1978. Secondary intersubjectivity. *Action, gesture and symbol: The emergence of language*, pages 183–229.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

I hea- umm think that's what they say: A Dataset of Inferences from Natural Language Dialogues

Adam Ek[1], Bill Noble[1], Stergios Chatzikyriakidis[3], Robin Cooper[1], Simon Dobnik[1], Eleni Gregoromichelaki[1], Christine Howes[1], Staffan Larsson[2], Vladislav Maraev[1], Gregory Mills[4], and Gijs Wijnholds[5]

[1]University of Gothenburg first.last@gu.se; [2]first.last@ling.gu.se
[3]University of Crete stergios.chatzikyriakidis@uoc.gr
[4]Kingston University g.mills@kingston.ac.uk
[5]Leiden University g.j.wijnholds@liacs.leidenuniv.nl

Abstract

In this paper we describe a dataset for Natural Language Inference in the dialogue domain and present several baseline models that predict whether a given hypothesis can be inferred from the dialogue. We describe an approach for collecting hypotheses in the ENTAILMENT, CONTRADICTION and NEUTRAL categories, based on transcripts of natural spoken dialogue. We present the dataset and perform experiments using a flat-concatenating and a hierarchical neural network. We then compare these to baseline models that exploit lexical regularities at the utterance level. We also pre-train BERT with additional dialogue data and find that pre-training with additional data helps. Our experiments show that hierarchical models perform better when using a random split of the data, while flat-concatenation models perform better on Out-of-Domain data. Lastly, LLM prompting is performed on two models, Llama 2 and Zephyr, the former barely exceeding the baseline, while the latter showing an incremental increase in performance as context length increases.

1 Introduction

Natural Language Inference (NLI, or Textual Entailment, TE) is one of the core tasks for Natural Language Understanding (NLU) and central to NLU benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). The centrality and importance of NLI has been acknowledged early on by Cooper et al. (1996), arguing that NLI is the crux of Computational Semantics. Similarly, Bowman et al. (2015a) argue that understanding inference about entailment and contradiction, in effect the task of NLI, is an important aspect for constructing semantic representations, while on a more practical note, Nie et al. (2020) note that NLI is arguably the most canonical task in NLU.

Dialogue in particular and interactive reasoning more generally are an integral part of human language use. We, among others (Bender and Koller, 2020; Dobnik et al., 2021), believe that if we want to understand meaning in language we need to adapt language systems which attempt to understand language to language's central domain, namely spoken dialogue. However, there have only been a few attempts to combine dialogue and natural language inference. In this paper we outline our efforts to combine NLI and dialogue by reporting how we constructed a procedure for eliciting inference examples from dialogue data. Essentially, we take transcribed speech from naturally occurring dialogues and ask annotators to write hypotheses sentences with different inferential statuses based on the beliefs of the speakers. The final dataset contains examples like the following:[1]

A	so do you
B	I mean yeah but it wasn't that how many years ago was that ? eight years ?
A	oh when we graduated just six years ago wasn't it ?
B	yeah
A	two thousand and eight
HYPOTHESIS	they graduated last year
LABEL	**Contradiction**

Typically, information conveyed by a speaker is not limited to one turn only, but is spread out over several turns with other speakers asking clarification questions, expressing agreement and so on. Then the meaning of a turn can be considered as a joint construal of the interlocutors (Clark, 1996). A consequence of this for NLP systems is that meaning cannot be assigned to utterances independently of the dialogue history.

Thus, modelling dialogue involves both forming a representation of what has been said in the dialogue, and incorporating new utterances into this representation. By only looking at individual turns or utterances in a dialogue we are excluding the information conveyed by the *interaction* between the participants.

[1]The dataset is available at github.com/GU-CLASP/DNLI.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Another aspect of meaning in dialogue is that different speakers can have different interpretations given the same dialogue. Thus, when modeling dialogue and in particular multi-party dialogues, it becomes important to consider *whose* perspective we are modeling. The task of performing inference on dialogue examples presents models with a interesting set of challenges, not only does the model have to predict a label given a hypothesis, but also construct, or identify, a set of utterances that supports the hypothesis, both based on the semantic content expressed and pragmatic actions (speech acts etc.).

In sum, we present a dataset that contains natural language inference examples in the dialogue domain, named DNLI. The examples in our dataset differ crucially in at least three respects compared to existing NLI datasets that contain dialogue data: (1) a piece of dialogue can contain more than two participants (up to four), (2) a speaker may produce many utterances in one *turn* or core information may be spread out over several turns, and (3) the turns and utterances themselves might contain disfluencies like hesitations and also commonly found dialogue phenomena like repairs, split utterances and so on (Schegloff et al., 1977; Lerner, 1991; Purver et al., 2018).

2 Related work

The common ground-annotated dataset of Markowska et al. is the most similar work to ours to date. Dialogues from the CallHome dataset are annotated on the utterance level for (1) propositions that are introduced by the utterance and (2) the status of those propositions with respect to the common ground of the two speakers. By taking propositions that are considered common ground by both speakers at a given point in the dialogue, one could produce dialogue contexts and entailments along the lines of what is proposed in this paper. However, their dataset is much smaller (561 utterances), making it less suitable for machine learning. Moreover, our dataset also includes hypotheses labeled as *contradiction* and *neutral* with respect to the context, which is important for robustly training and assessing an inference model.

The MNLI dataset (Williams et al., 2018), which is a multi-genre NLI dataset, includes some examples that can be classified as dialogue—a little over a fifth of the examples are drawn from transcripts of telephone calls from the Switchboard corpus (Godfrey et al., 1992). However, none of the important characteristics of dialogue, which may influence (e.g. disfluencies, split utterances, repairs, interactivity, incrementality, or turn-taking) the beliefs of the speakers, are taken into account. Rather, the examples seem to function in exactly the same way as typical NLI cases: a single sentence from the source data is paired with a hypothesis generated by annotators. This is to be expected as the aim of MNLI was to include a multiplicity of text genres, rather than dealing with the intricacies of reasoning in dialogue settings specifically.

Welleck et al. (2019) presents an NLI dataset based on Persona-Chat (Zhang et al.). It is also referred to as Dialogue NLI, but has substantially different aims from the present work. The dataset consists of premise-hypothesis pairs, where the hypothesis is drawn from a set of *persona sentences* (facts about the speaker) and the premise is either a persona sentence or an utterance from the dialogue. The dataset seeks to improve the ability of chit-chat dialogue models to generate utterances consistent with the "persona" of the agent. In contrast, our dataset is interested in the ability to keep track of what is entailed by the dialogue itself, which requires reasoning over a dialogue context that includes multiple utterances. Moreover, our dataset is based on naturally-occurring transcribed face-to-face interaction, whereas Persona-Chat consists of text chat conversations between crowd workers play-acting as an assigned persona.

Khanuja et al. (2020) introduce a dataset for Natural Language Inference (NLI) from code-mixed Hindi-English conversations of Bollywood movies. It is comprised of 400 premises and 2240 hypotheses, annotated by Hindi-English bilinguals.The paper evaluates the dataset using an mBERT-based pipeline, revealing that existing multilingual models are not yet competent in handling code-mixed NLI tasks. Again, this is a different setup than ours, involving single premise-hypothesis examples, and does not require reasoning over a dialogue context that includes multiple utterances.

There exist a number of other dialogue and discourse datasets which might be helpful for natural language inference tasks. Many such dataset are summarised within the ParlAI (Miller et al., 2017) tool. In particular question-answering datasets may be relevant for the dialogue inference task, although they typically do not involve answering questions about the dialogue itself (e.g., Antol et al., (VQA)). While it is not dialogue, The bAbI (Weston et al., 2015) is another relevant question-answering dataset. Inputs consist of a sequence of statements representing an emergent context, followed by a question about the context. Paperno et al. (2016) put together the LAMBADA dataset, where context is comprised of a list of passages (including dialogical exchanges) and the task is to predict the last word of a target sentence which follows from the context.

3 Dialogue NLI

Typically, a Natural Language Inference example contains a *premise* statement and a *hypothesis* statement. Then, the task is to determine whether the hypothesis is *entailed* or not from the premise. That is, does the premise contain sufficient evidence to determine whether the hypothesis is true or not.

In our dataset we cast the premise as a continuous sequence of *utterances* from a dialogue. The hypothesis is a statement that one of the dialogue participants

would take to be true, false or neither true nor false. In particular we take the perspective of the speaker who most recently produced an utterance when evaluating the hypothesis. This is because the common-ground of different dialogue interlocutors may diverge without being acknowledged by the participants, but an outside observer could potentially observe this.[2]

To elicit statements about a particular speaker's point of view, we ask annotators to produce a statement which one of the speakers make a judgement about, based on what has been said in the dialogue. We follow Bowman et al. (2015a); Williams et al. (2018) and consider three types of inference judgements: TRUE (ENTAILMENT), FALSE (CONTRADICTION) or NEITHER (NEUTRAL), presented as follows to our annotators:

ENTAILMENT: A statement that the last speaker would take to be true at this point in the dialogue.

NEUTRAL: A statement for which there is no evidence that the last speaker would take it to be true or false at this point in the dialogue.

CONTRADICTION: A statement that the last speaker would take to not be true at this point in the dialogue.

Thus, each hypothesis is based on a particular point in time. As such, we can't know whether a participant would make the same judgement again if the dialogue continues, as new information can be expressed in the dialogue which may change what the participants believes. As an example, let us consider a dialogue whose hypothesis involves what the topic is being discussed. The hypotheses "they are talking about wine" will only be an entailment as long as they continue talking about wine. However, if the topic changes from "wine" to "saunas" as in the dialogue below, the previous hypothesis will no longer be an ENTAILMENT but a CONTRADICTION (since they are *not* talking about wine anymore).

D	so what was the conclusion with the wine thing should you pour it? is it
	...
HYPOTHESIS	they are talking about wine
LABEL	**Entailment**
A	I mean it does alter the taste
	...
C	I'd much prefer sitting in a sauna nice and dry and hot

To summarize, we consider a sequence of utterances $u_0^{S_i} ... u_n^{S_j}$ to be the premise, and a hypothesis h. The label of the hypothesis is based on the beliefs of speaker S_j (the one who uttered u_n) when u_n was uttered.

[2] A project that explores this is Ghosal et al. (2021), who considers dyadic dialogues and what common-sense inferences that one can draw from those.

4 Data collection

In this section we describe the corpora used to create the Dialogue Natural Language Inference dataset and the way annotations were elicited.

4.1 Dialogue Corpora

Our corpus of annotated dialogues draws from the BNC2014 (Love et al., 2017)[3] and the CHILDES (MacWhinney, 2000) corpora. The corpus contains 13,856 annotations distributed over 938 dialogues from the BNC data and 287 annotations on 17 dialogues from the CHILDES corpus.[4]

CHILDES is a collection of corpora of transcribed spontaneous conversations, mostly between children and their adult caregivers. We draw from dialogues in the Warren-Leubecker and Iii (1984) portion of the corpus, which is comprised of conversations between English speaking two- and five-year olds from suburban Atlanta and their parents. As CHILDES is a widely used resource in both the child language acquisition and computational modeling communities, a number of other annotation resources are available, including morphological and syntactic annotations (Buttery and Korhonen, 2005; Sagae et al., 2010; Villavicencio et al., 2012) and utterance-level semantic annotations (Bergey et al., 2021). The Warren-Leubecker and Iii (1984) portion of the corpus also includes intonation annotations.

BNC (Love et al., 2017) is a follow-up to the 1994 version of the BNC, comprised of conversations between adult native speakers of British English. A key component of the BNC dataset which makes it particularly interesting for NLI is that it is naturally occurring speech that has been annotated faithfully, such that repairs, disfluencies, and so on are included in the data. The dialogues in the dataset have 2-4 participants making it interesting for NLI as a model potentially has to learn 2-4 different belief representations if their beliefs diverge. The dialogues on the BNC dataset are also long and touch on many different topics (even within one dialogue). Thus, to successfully model these dialogues, a system must be able to handle that participants changing topics, or talk about different topics in the dialogues. Moreover, because the dialogues are naturally-occurring, a model must also learn to accurately model an open-ended range of topics (Chen and Gao, 2017; Shalyminov et al., 2020). This is in contrast to many other dialogue datasets such as MultiWoZ (Budzianowski et al., 2018) in which topics are constrained to a pre-determined set of task-oriented scenarios.

One major argument for including both adult-adult and adult-child dialogues is that in real-life dialogues, participants can not always assume symmetry in the

[3] Henceforth, BNC.

[4] An *annotation* refers to a hypothesis of one of the three labels (ENTAILMENT, CONTRADICTION, NEUTRAL) elicited from an annotator. See §4.3 for details.

linguistic resources of their interlocutors. Thus, there will be cases when a model is forced to interpret a speakers intention, even when it is not expressed in the most conventional or explicit way, just as a human would have to when speaking to a child.[5]

We believe the linguistic diversity of speakers is a key aspect of dialogue often neglected in dialogue research. The use-cases for dialogue systems often involve a *random person* talking to someone. This becomes very problematic when we consider the case of a child interacting with some QA system. If the system is developed with data collected from adult speakers only, it won't be able to take into account linguistic variations associated with children.

4.2 Dialogue formatting

The data from both the BNC and CHILDES corpus contain longer dialogues, with up to 15000 utterances in a dialogue (but about 900 on average). This presents a problem as we would like to give the annotators a dialogue that can easily be read and subsequently annotated. To make the annotation process feasible, we split each dialogue into n sub-dialogues, where each sub-dialogue contains around 50 utterances each. For each sub-dialogue we select $1-5$ utterances at random and insert an annotation flag, as a constraint to this we do not allow for consecutive annotation flags they must be separated by more than two utterances. Then for each annotation flag we elicit an annotation.

4.3 Annotations

In the annotation process we utilize both Amazon Mechanical Turk workers and Master students in the Language Technology program at the University of Gothenburg We noted that the task was difficult for AMT workers, which prompted us to manually go through all the AMT annotations and select the reasonable ones. The masters students were hired on an hourly basis and got paid 12 USD per hour. The AMT workers were paid around 3 USD per dialogue.

We created an online annotation tool that the workers used to do their annotationk. On the web page, the dialogue is presented incrementally, such that the annotators have the same information as the participants in the dialogue. So neither the annotators nor the dialogue participants are able to see future utterances.

We ask the annotators to take the perspective of the *last speaker* and freely write a hypothesis statement conforming to one of the three labels: *true*, *false* or *unknown*. Our approach to NLI is similar to other large scale datasets, such as (Bowman et al., 2015b; Williams et al., 2018; Khot et al., 2018), where the logical constraints have been relaxed.[6]

[5]Two speakers interacting NEVER share the 'same' language, so the model has to be able to deal with asymmetries. Child-adult conversations or dialogues between native and non-native speakers are just obvious examples of this (Clark, 1998).

[6]It is important to note that the notion of entailment is not uniform across all NLI datasets. An overview can be found in (Chatzikyriakidis et al., 2017; Bernardy and Chatzikyriakidis, 2019; Poliak, 2020).

B	see you in a year
A	so what do we do like what do I do if with the birthday card? can I send it to you? like will you have an address?
D	what birthday card?
B	yours
A	well you'll be away for your birthday
B	yeah
D	no don't bother
HYPOTHESIS	Speaker D don't want birthday cards
LABEL	**Entailment**

One core feature of this dataset is that for a model to accurately predict the label of a statement, the model must compose the information given over several turns and take into account a speaker's perspective which will be different depending on who the speaker is. This will be affected by the usage of pronouns and spatial perspective but also other facts about the speaker and their role in the conversation. In the example above, the model must infer that when speaker D says *"no don't bother"* they are referring to the giving of a birthday card which was proposed by A five turns earlier.

During the annotation process only simple instructions on how to refer to participants were given, resulting in a variety of strategies. We extracted these using a simple regular expression to get an idea how the annotators did this, as shown in Table 1. Primarily, we observe two ways of referring to a specific speaker, "Speaker X" or "Person X", additionally pronouns such as "they" were used often. We also note that "he" and "she" occur but much less frequently. It is often not clear from the dialogue alone which gender a speaker is (exemplified in the above dialogue). There are slightly more occurrences of male over female pronouns.

Referring expr.	Count
Speaker X	2963
Person X	2193
They	2152
He	240
She	229
Her	20
His	26

Table 1: Expressions used by annotators to refer to speakers in the dialogue.

5 Data analysis

In this section we describe some descriptive statistics of the dataset. The distribution of labels shown in Table 2, we see that the labels are distributed roughly evenly across the labels, slightly favoring Entailment.

Label	Count	Proportion
Entailment	4799	0.338
Contradiction	4677	0.329
Neutral	4723	0.333

Table 2: Distribution of labels in the dataset.

In total, we collected 14 179 hypotheses from the dialogues in our corpora and show the number of dialogues from each corpora in Table 3.

Source	Dialogues	Annotations
BNC	938	13 892
CHILDES	17	287

Table 3: Number of dialogues and annotations from BNC and CHILDES.

We collect data from 955 dialogues in total, where 938 of the dialogues are from BNC are 17 are from the CHILDES. In the BNC portion of the dataset there are 13 892 hypotheses annotated and in the CHILDES portion 287 hypotheses annotated.

One feature of our dataset is that some dialogues have more utterances than others. For example, the shortest dialogue contains 125 utterances, while the longest one contains 15 054 utterances. As shown in Appendix A, Figure 1, the number of utterances in the dialogues follow roughly a Zipfian distribution.

During the annotation process we randomly select sub-dialogues (see Section 4.2). A consequence of this is that longer dialogues tend to receive more annotations. We show the number of annotations available for each dialogue in Appendix A, Figure 2.

Additionally, we look at the number of tokens in both the premises and hypotheses, shown in Table 4. We see in Table 4 that the number of tokens per utterance can vary a lot with a standard deviation of 7.31 tokens and a mean of 6.05 tokens.

	Mean	STD
Tokens/Utterance	6.05	7.31
Tokens/Hypothesis	8.13	3.14

Table 4: Distribution of tokens in the dataset.

This is caused by for example utterances only containing back-channels and disfluencies. In general, this poses an interesting problem for models that also occur in real-life: namely to select the utterances that provide useful information to some belief of a speaker. For the number of tokens per hypothesis this contain less variation, but they tend to be longer than the premises.

5.1 Data splits

We provide a split of the data into a standard split, with the following data distribution: 80% training, 10% validation and 10% testing, ensuring that the label distribution is roughly uniform between the different data splits. Additionally, because the BNC corpora is larger and we elicited more annotations from this dataset we ensure that the validation and test splits contain more than two dialogues from the CHILDES corpora.

Another split we consider is an Out-of-Domain split (Zheng et al., 2020; Haddow and Koehn, 2012), where training and development data is randomly sampled from the BNC and the test data taken from CHILDES. This type of splitting allows us to estimate how much we can learn about dialogues regardless of domain (chit-chat versus more task-oriented dialogues). Another feature of this type of split is that we can evaluate how dialogues between adults transfer to dialogues between caregivers and children. As we have mentioned earlier, dialogues occur between different types of people and systems of dialogue need to handle this.

For reproducibility we perform all experiments in this paper with standard and Out-of-Domain split, but encourage future work to explore other data splits (see Gorman and Bedrick, 2019; Søgaard et al., 2020).

6 Experiments & Results

We perform experiments both on the standard split and Out-of-Domain split, and investigate the performance of two model architectures: flat-concatenation (Smith et al., 2020; Zhang et al., 2020; Li et al., 2021) and hierarchical (Serban et al., 2016; Tran et al., 2017).

In the flat-concatenation architecture the utterances preceding an annotation are concatenated together and fed to the model as one sequence. We apply max pooling over the sequence to get a dialogue representation. In the hierarchical architecture we consider two levels of representation: a token level representation where the tokens in each utterance are encoded (and as in the case of flat-concatenation, we use max pooling), and an utterances level representation where the representation from the token level are modeled. To get a dialogue representation D we use additive attention

$$D = \mathsf{softmax}(w^T \tanh(W_a^T k + W_b^T u))\, u$$

where u is the utterance representation and k the max-pooled hypothesis representations. We experimented with other ways of compiling this information (dot-product attention, self-attention, last hidden state, max/mean pooling) but found that additive attention yielded the best performance. The main idea is that tokens and utterances are distinct units of information, as such it could be beneficial to model these in a hierarchical fashion. An overview of the hierarchical architecture we employ is given in Figure 1.

For all architectures we model the interaction between the premise and hypothesis representations by concatenating u, h, $|u - h|$ and $u \odot h$ (element-wise multiplication) (Conneau et al., 2017), where u is the premise and h is the hypothesis. An overview of this procedure is given in Figure 2.

All experiments were conducted on a RTX Titan 12GB card with a batch size of 4 over 20 epochs with

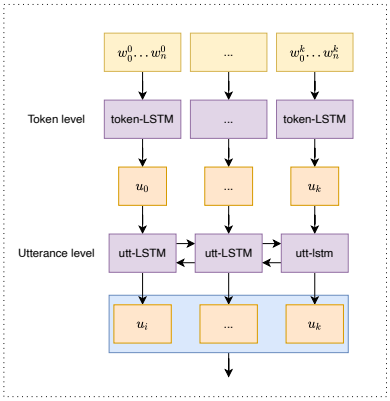


Figure 1: Overview of the dialogue encoder in the hierarchical architecture, where $w_0^k ... w_0^n$ are the tokens of utterance 0, and $u_0 ... u_k$ the utterance representation.

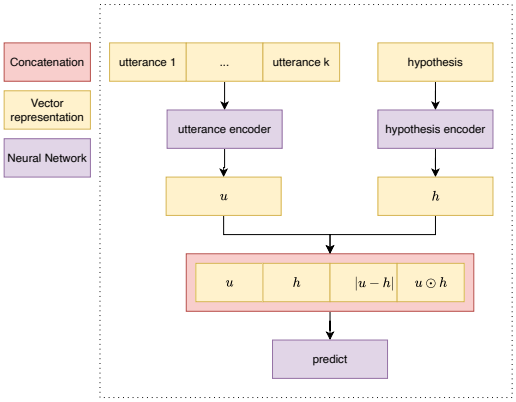


Figure 2: Base architecture of our dialogue NLI models. We encode the utterances and hypothesis separately, then before we predict a label we learn how the two representations interact by concatenating u, h, $|u - h|$ and $u \odot h$ (element-wise multiplication).

early stopping (two epochs of no improvement on the dev data). We use the Adam (Kingma and Ba, 2015) optimizer with default parameters, a Cosine Annealing learning-rate scheduler (Loshchilov and Hutter, 2017), with an initial learning rate of $1e{-}3$ and a minimal learning rate of $1e{-}7$, and weight decay of 0.01. For both architectures, we experiment using a transformer model (`bert-base-uncased`) [7] or a LSTM as the utterance and hypothesis encoders.

As annotators had limited access to context we explore how many utterances to use as the premise. Too much utterance context could introduce noise and too little could miss the utterances where relevant information is expressed. To investigate this question we perform our experiments using different numbers of utterances as context, namely contexts of size $1, 3, 5, 7, 9, 11, 13$ and 15. The performance of our models on the standard and Out-of-Domain split is given in Figure 3 and Figure 4 respectively.

[7] We also performed experiments with GPT2, but found no clear difference compared to BERT.

In the experiment on the standard data split we can observe than for context $1, 3$, and 5 the BERT based models (both flat-concatenation and hierarchical) perform slightly better than LSTM based approaches and in all other cases outperform the LSTM based approach. As we increase the context (and thus information available to the model), BERT-based approaches start to perform better than the LSTM approach, and the outlier here is for context 13, where the flat-concatenation BERT model performs roughly the same as hierarchical LSTMs. We see clearly that hierarchical BERT is most effective with 5 or more utterances as the premise, where we get a substantial increase in performance. The performance of the flat-concatenation model varies across contexts, only out performing the hierarchical BERT with contexts sizes of $1, 3, 9$ and 11. For the LSTM models there is a clear preference for the hierarchical architecture.

In the Out-of-Domain split we see a lower performance across setups. However, another pattern appears, namely that the flat-concatenation models perform better than the hierarchical models.

6.1 Baselines

We consider a number of baselines whose primary aim is to probe biases in the data and explore how far we can get without actually modelling real dialogues. We consider the majority class as one of these baselines, and the hypothesis-only baseline. In the hypothesis-only approach we simply try to predict the label based on the hypothesis and *not* the premise (i.e. the dialogue utterances) (Poliak et al., 2018). This baseline probes for biases in the hypothesis statements associated with different labels. For example, if the word "not" occurs in every contradiction, the model will likely learn to exploit that regularity in the hypothesis rather than modelling the relationship between premise and hypothesis. The performance of the baselines is shown in Table 5.

Model	Standard Split	Out-of-Domain
Majority Class	33.8	35.5
LSTM Hyp. only	51.3 ± 0.4	42.4 ± 0.2
BERT Hyp. only	58.9 ± 0.9	44.4 ± 0.4

Table 5: Baseline performance on the standard split and for training on BNC and testing on CHILDES (Out-of-Domain).

The majority class baseline reveals that the labels in both the standard split and Out-of-Domain data is more-or-less balanced, with the Out-of-Domain data showing a slightly higher bias to neutral hypotheses.

The hypothesis-only model does perform better than the majority class baseline, suggesting that there is some bias in the hypothesis statements for the models to exploit. A BERT-based approach to this baseline yields a higher accuracy of 58.9% versus 51.3% for the LSTM. Interestingly, the performance of the hypothesis only baseline is lower relative to the ma-

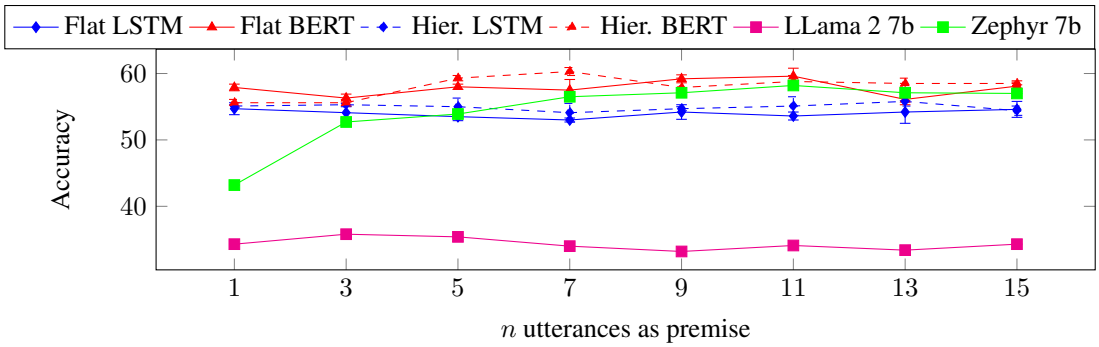


Figure 3: Mean accuracy and standard deviation over three runs on the standard split. We consider both a LSTM and a BERT-based approach. Additionally, we show the performance of the Llama 2 7b, which was prompted with three examples from the training set.

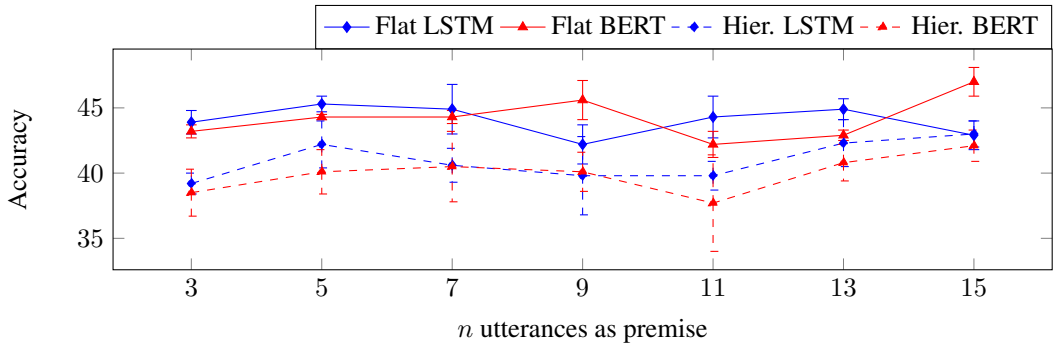


Figure 4: Mean accuracy and standard deviation over three runs on the Out-of-Domain test data. We consider both a LSTM and a BERT-based approach.

jority class for the Out-of-Domain data (8.6% for the LSTM vs 25.1% for the BERT hypothesis only). Thus, it seems that the hypotheses generated from CHILDES dialogues are less biased with respect to the label.

6.2 Dialogue pre-training for utterance encoders

While there is widespread evidence that large language models improve performance on Out-of-Domain tasks, models such as BERT, trained on text-only corpora, may have trouble representing features specific to spoken dialogue. Indeed, there is evidence that these models require fine-tuning to perform well on dialogue-specific tasks and that additional pre-training in the dialogue domain can be helpful (Noble and Maraev, 2021). For that reason, we experiment with a BERT utterance encoder that has been additionally pre-trained on in-domain data. In particular, we create a corpus from the dialogues in the spoken section of the BNC which were not included in the DNLI dataset. This amounts to 1,252 dialogues and 1,119,747 utterances (about 30% larger than the DNLI dataset). To assess the contribution of the original BERT pre-training, we train two BERT models: one with randomly-initialized parameters (BERT RandomInit), and one initialized with the standard pre-trained BERT-base parameters (i.e., the BERT model that is used in other experiments).

In each case, the model was trained with a masked language modelling objective (Devlin et al., 2018) over 100 epochs on the BNC pre-training corpus described above. Models were trained with a batch size of 64, though gradients were accumulated every 8 batches, making the effective batch size 512. We took the model from the epoch with lowest development loss.[8] These were 0.16 (epoch 98) and 0.53 (epoch 85) for pre-trained and from-scratch BERT models, respectively. These results suggest that BERT is able to leverage its text pre-training in the masked language modelling objective, but it remains to be seen whether the text pre-training is useful for the downstream inference task.

When applied to the DNLI dataset we see that the regular pre-training of BERT appears to be helpful, as downstream performance of BERT RandomInit consistently decreases, both for the standard and Out-of-Domain data splits. We also see that the pre-training of BERT using BNC improves the performance of the flat concatenation model while it decrease the performance of the hierarchical model, as shown in Table 6. We argue that the flat-concatenation model, operating on the token-level only, has a closer connection to the pretraining objective during finetuning and thus can more easily exploit the dialogue information provided by BNC, unlike the hierarchical model that operates on the utterance-level.

[8]The utterances from the validation split of the DNLI dataset were used for development.

	Standard Split		Out-of-Domain	
	Flat	Hierarchical	Flat	Hierarchical
BERT RandomInit	55.4(−2.1)	54.6(−5.7)	42.5(−1.8)	41.8(−3.1)
BERT+BNC	58.5(+1.0)	58.4(−1.9)	47.4(+3.1)	43.6(−1.3)

Table 6: Performance with BERT trained on only BNC (BERT RandomInit) or with BNC as additional pre-training (BERT+BNC), with 7 utterances as context on our DNLI dataset, the difference from standard BERT in terms of percentage points is shown within parenthesis.

6.3 LLM prompting

We also report the performance of two large language models (LLMs), Llama 2 7b (Touvron et al., 2023)[9] and Zephyr 7b (Tunstall et al., 2023)[10]. In both cases, the model was provided with a short prompt describing the task and three examples from the training set (see Appendix B for an example). The model's generation was constrained to one of the three target labels[11]. An example prompt in shown in the appendix. We observe a strong difference in results; where LLama 2 barely beats the majority class baseline, Zephyr displays a largely continual increase in performance as the context length increases. Overall it reaches performance slightly below the hierarchical BERT-based model.

7 Discussion

7.1 Annotations

During the annotation process we noted that several different strategies were used to refer to speakers, and sometimes pronouns or other referring expressions were used in the hypotheses. To know if a certain hypothesis is true or not requires anaphora resolution. Then, should this responsibility be placed upon the annotators or the models? We would like a model to disambiguate referring expression as it is a part of language use. But the hypotheses are written in a "meta-language", that describe beliefs of a speaker, and the question is: do we actually want the model to learn this meta-language or not? The goal of our dataset is to allow for dialogue understanding and how meaning is obtained, and then the task of disambiguating referring expressions is redundant. But we would also like systems to do this, so they can function in a real-world application, where meta-language does occur. In the dataset we put the burden of disambiguation on the models rather than the annotators.

7.2 Experiments

In our experiments, we observe that the hypothesis-only baseline is outperformed by a full model in LSTM-based approaches, but the converse occurs when using BERT.

Given that hypotheses are generally framed meta-linguistically, we are in fact already modeling two domains, the dialogue and the meta-language. So the question is, why doesn't dialogue context always help? As mentioned earlier, one issue may be that the model has to model two domains and also disambiguate referring expressions. Another issue is dialogue phenomena such as repairs, disfluencies and split-turns, can be interpreted as noise by BERT. To properly use these features of dialogue systems must recognize that they serve a *pragmatic function*, that a dialogue is a joint effort of two or more dialogue participants. If a model can not do this, utterances such as "umm" or repairs, that provide no clear semantic meaning will be modelled improperly. This may be one of the failings of the hierarchical model as each utterance, however small, gets a representation.

Among the large language models, we see that Zephyr greatly out-performs Llama 2 and that Zephyr is able to take advantage of longer context windows. This may be a result the fact that Zephyr was fine-tuned as a chat model using Direct Preference Optimization.

8 Conclusion

We have presented our data collection process for a dataset of natural language inference in dialogues, the first of its kind that uses both natural dialogues and hand-annotated hypotheses. We performed experiments using LSTM, BERT and prompted LLM baselines. The dataset is hard to model properly as shown by our experiments, where the best performance we obtained was about 2% better (Figure 3 with 7 utterances as context for hierarchical BERT) than a hypothesis-only baseline. It is unclear if the LSTM and BERT models are able to recognize pragmatic functions of dialogue phenomena such as disfluencies and back-channels, or make sense of split utterances or repairs. We believe this dataset fills a gap for both dialogue systems and natural language inference systems, presenting a challenging dataset in both research directions.

Future work includes collecting additional annotations, such as paraphrases or meaning-reversing modification of the current annotations. Additionally, we plan to continue annotating the BNC corpora to achieve full coverage. Another avenue in this direction is to explore how models deal with dialogue phenomena that serve a pragmatic function such as back-channels and disfluencies, and how to properly model these with neural networks.

[9] We use the AWQ-quantized version available from TheBloke/Llama-2-7B-AWQ on HuggingFace.

[10] Available from HuggingFaceH4/zephyr-7b-beta. We use the 4-bit bitsandbytes quantization configuration.

[11] We employ constrained decoding from the guidance library: https://github.com/guidance-ai/guidance

Acknowledgments

The research detailed in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Stergios Chatzikyriakidis gratefully acknowledges funding from the Special Account for Research Funding of the Technical University of Crete (grant number: 11218), as well as funding from the TALOS-AI4SSH ERA Chair in Artificial Intelligence for Humanities and Social Sciences grant (grant agreement: 101087269).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Claire Bergey, Zoe Marshall, Simon DeDeo, and Daniel Yurovsky. 2021. Learning communicative acts in children's conversations: A Hidden Topic Markov Model analysis of the CHILDES corpus.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *ICAART (2)*, pages 919–931.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

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.

Paula Buttery and Anna Korhonen. 2005. Large Scale Analysis of Verb Subcategorization differences between Child Directed Speech and Adult Speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of natural language inference data collection: The way forward? In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics, Workshop on Computing Natural Language Inference*, pages 1–6, Montpellier, France. Association for Computational Linguistics.

Yun-Nung Chen and Jianfeng Gao. 2017. Open-domain neural dialogue systems. In *Proceedings of the IJCNLP 2017, Tutorial Abstracts*, pages 6–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Herbert H Clark. 1998. Communal lexicons. In Kirsten Malmkjær and John Williams, editors, *Context in Language Learning and Language Understanding*, chapter 4, pages 63–87. Cambridge University Press, Cambridge.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680. Association for Computational Linguistics.

R. Cooper, D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman. 1996. Using the framework. Technical report LRE 62-051r, The FraCaS consortium. `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Simon Dobnik, Robin Cooper, Adam Ek, Bill Noble, Staffan Larsson, Nikolai Ilinykh, Vladislav Maraev, and Vidya Somashekarappa. 2021. We went to look for meaning and all we got were these lousy representations: aspects of meaning representation for computational semantics. *Preprint*, arXiv:2109.04949.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Cider: Commonsense inference for dialogue explanation and reasoning. *arXiv preprint arXiv:2106.00510*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International*

Conference on, volume 1, pages 517–520. IEEE Computer Society.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2786–2791.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on smt systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gene H. Lerner. 1991. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. *Preprint*, arXiv:2106.02227.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed edition. Lawrence Erlbaum, Mahwah, NJ.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. Finding Common Ground: Annotating and Predicting Common Ground in Spoken Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Bill Noble and Vladislav Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics*, pages 166–172, Groningen, the Netherlands (online). Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Adam Poliak. 2020. A survey on recognizing textual entailment as an nlp evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

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

Kenji Sagae, Eric Davis, Alon Lavie, Brian Macwhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729.

E.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.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Hybrid generative-retrieval transformers for dialogue domain adaptation. *arXiv preprint arXiv:2003.01680*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2020. We need to talk about random splits. *arXiv preprint arXiv:2005.00636*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arxiv:2302.13971.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, von Leandro Werra, Fourrier, Clémentine, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *Preprint*, arxiv:2310.16944.

Aline Villavicencio, Beracah Yankama, Rodrigo Wilkens, Marco Idiart, and Robert Berwick. 2012. An annotated English child language database. In *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*, pages 23–25, Avignon, France. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Amye Warren-Leubecker and John Neil Bohannon Iii. 1984. Intonation Patterns in Child-Directed Speech: Mother-Father Differences. *Child Development*, 55(4):1379.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

J. Weston, Antoine Bordes, S. Chopra, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks.

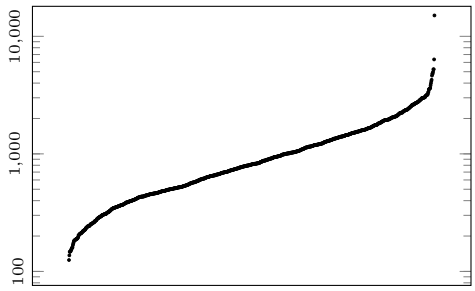
Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

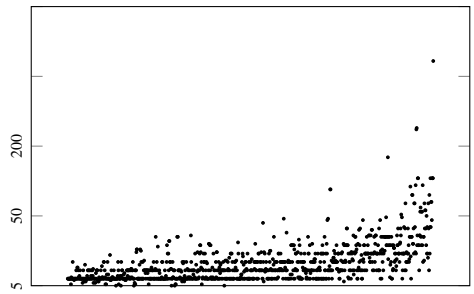
Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

A Additional descriptive statistics



Appendix Figure 1: Number of utterances in each dialogue sorted on a logarithmic scale. The y-axis indicate the number of utterances in each dialogue and the x-axis each dialogue.



Appendix Figure 2: Number of annotations in each dialogue are shown on the y-axis, and the dialogues sorted by number of utterances of the x-axis.

B LLM prompt

Given a dialogue excerpt and a Hypothesis, decide on the semantic relation between them, choosing between Entailment, Contradiction, and Neutral.

SPEAKER B: well his his his brother had cancer his brother died
SPEAKER C: did he?
SPEAKER A: and then his mum got ill
SPEAKER B: and then his mum got really ill he dropped he
SPEAKER C: mm
HYPOTHESIS: they are talking about fathers
RELATION: Contradiction
SPEAKER B: he was ex-
SPEAKER C: yeah
SPEAKER B: so it seemed to be a bit of a stopgap bit like when dad
SPEAKER A: yeah
SPEAKER C: yeah
HYPOTHESIS: they are not sure about dad
RELATION: Neutral
SPEAKER A: yeah
SPEAKER B: mm yeah yeah
SPEAKER C: but I I was paid for it and I got bonuses and everything and it was good money
SPEAKER A: yeah
SPEAKER C: in the early eighties early to mid-eighties
HYPOTHESIS: they are talking about eighties
RELATION: Entailment
SPEAKER C: yeah saucepan
SPEAKER D: yeah it should be a saucepan
SPEAKER C: small one heavy bottomed
SPEAKER A: and then like
SPEAKER D: low heat do it low heat
HYPOTHESIS: to make proper scrambled eggs, you must use a pot
RELATION: [FILL]

Figure 5: An example prompt for LLM prompting. We use three examples, one for each NLI label, [FILL] indicates the generation of the model which is constrained to the three possible labels.

Towards A Formal Semantics of Silence: An Analysis Based on the KoS Framework

Haseon Park
Saarland University
hpark@coli.uni-saarland

Abstract

Despite its nonnegligible communicative role in verbal communication, conversational silence has been outside the concern of formal and computational semantics because of the difficulty of analysis arising from its extreme multimodal context-dependency in detection and interpretation. However, I argue that, whereas the conventional formal semantic theories whose level of analysis is a sentence or truth-conditional/situational worlds do not provide the tool to grasp the denotation of silence, KoS, a multimodal conversation-level semantic framework can successfully handle it. In this paper, We focus on turn and inter-turn silence in Levinson's classification of silence (turn, inter-turn, intra-turn), further subclassify those two classes of silence into inability, refusal, acceptance, turn-passing, truthfulness, unwillingness, and awkwardness silence by their forms and meanings, and formally describe and analyze them by presenting the lexical entries and the conversation rules with the perspective and the notation of KoS. I believe that this analysis can facilitate further research of silence in theoretical, experimental, and computational manners by explicitly expressing the grammar and the semantic content of silence and also demonstrate the possibility of the semantic annotation of silence in dialogue corpora.

1 Introduction

Silence often conveys meaning in verbal communication. This type of silence is called 'conversational silence.' Consider (2) from Wang (2019). ('X +> Y' expresses that conversational implicatures trigger an inference from X to Y.)

(1) (A conversation between two passengers on the street)

Man: Excuse me Ma'am where is the No.67 bus stop?

Woman: [silence; having heard the man's question]

Man: [take a closer look at the woman] Sorry, miss, could you please tell me what the No.67 bus stop is?

Woman: Go straight ahead, and turn right at the next crossroad.

+> The woman did not want to answer because she was unhappy with being called 'miss.'

The woman's silence takes a 'turn' in conversation, performs an illocutionary act, and generates some conversational implicature.[1] Also, silence is often described as 'ambiguous' (Perniola, 2010; Ferguson, 2003; Glenn, 2004; Jaworski, 2011). These two observations imply that conversational silence is a semiotic object with semantic content that should be disambiguated and semantically parsed to be understood by the dialogue participants. By developing formal descriptions, we can systematically and explicitly study the meaning of silence, facilitating theoretical, experimental, and computational research and analysis of what silence means and how humans understand it. It will also allow semantic annotations to silence in conversation corpora.

However, to the best of our knowledge, no formal semantic analyses have been given to silence. This is because the conventional formal semantics, which is only concerned with proposition- or world-level semantic phenomena in a single modality (speech or text), has no or little space for silence. Silence is one of the most extreme cases of multimodal communication. Since it is (a subclass of) the absence of utterances or any signs, one cannot grasp it based on the connection between the

[1] There can be diverse approaches to the nature of silence' semantic contents and the inference derived from them. For example, some inferences required to interpret conversational silence can be analyzed as cases of conversational implicature, explicature, impliciture, or something else depending on the theory. However, I set the semantic contents of silence as simple as possible and all the relevant inferences as

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

signified and the overt signifiers cohering in any single modality, and all of the surrounding contexts are needed to detect its presence and interpret its meaning. Meanwhile, the KoS framework provides a way to define formal structures and conversational rules for multimodal communication by encoding any information about the environmental or social/discursive situations surrounding the participants. Thus, the natural next step is to use the framework to analyze the pragmatic mechanism surrounding silence.

In this paper, I classified and analyzed the form and the content of silence using KoS by formulating the lexical entries and the conversational rules that explain the semiosis of silence with some dialogue examples. Section 2 briefly reviews the previous studies and concepts on silence relevant to this paper, the TTR/KoS framework, which is the theoretical tool to analyze silence in this paper, and the other exemplary studies of multimodal and/or paralinguistic signs using the framework. The scope of analysis of this paper is formulated more in detail with the concepts reviewed in Section 3. Sections 4 and 5 classify and analyze the sub-classes of turn silence and inter-turn silence respectively.

2 Background

2.1 Previous studies on silence

Several scholars in the field of linguistics classified and analyzed silence as a linguistic object that has semantic and/or pragmatic functions. Levinson (1983) classified silence into three types based on the relation to conversation turns: intra-turn silence (*pause*), inter-turn silence (*gap* or *lapse*), and turn silence. Kurzon (1995) divided silence into two types (*intentional* and *unintentional*) and suggested the modal interpretation of silence, that silence can be interpreted as 'I cannot speak', 'I don't want to speak' or 'I must/may not speak.' Kurzon (2007) covers more diverse types of silence, giving four types of silence: conversational, thematic (avoiding talking about a specific topic while speaking), textual (silence when internally reading or reciting a specific text such as a prayer), and situational (silence required by sociocultural norms in specific spacetime). Ephratt (2007) distinguished *eloquent silence*, which has an active semantic content on its own, from *stillness* (e.g. just listening or in the library), *pause*, and *silencing* (prohibition to speak) and Ephratt (2008) analyzed the role of eloquent silence in terms of the six functions of language in Roman Jakobson's communication model. Wang (2019) followed the classification of Levinson (1983)(turn silence, inter-turn silence, and intra-turn silence) and focused on analyzing turn silence using Relevance Theory, describing the interpretation of silence as establishing its relevance in the conversation by three types: addition of a new contextual implication, strengthening of previously held assumptions, and elimination of false assumptions.

2.2 TTR/KoS framework

The KoS framework, which is a theoretical framework of conversation-oriented semantics was proposed first in Ginzburg (1994), in which the structures of a shared 'dialogue gameboard' and an 'information state table' are defined and the effects and the meanings of an utterance are analyzed as the updates of the dialogue gameboard and the information state table according to the conversational rules. Later, this approach was reformalized with Type Theory with Records(Cooper and Ginzburg, 2015), which facilitated a richer representation concerning every level of language from phonetic and syntactic to semantic and pragmatic levels. TTR's versatility and flexibility allowed it to extend KoS's ability to the realm of multimodal face-to-face communication(Lücking and Ginzburg, 2020), in which the diverse types of information such as gestures, facial expressions, the context from visual situations are exchanged together with the linguistic signs and take place sequentially or simultaneously. The advantages of this approach toward multimodal communication and paralinguistic signs are well exemplified in the analysis of laughter by Ginzburg et al. (2020).

On the other hand, the interpretation of (para)linguistic signs such as silence seems to heavily rely on their relevance in a dialogue and common sense reasoning. Relevance in terms of the KoS framework was explored by Ginzburg (2010), and the mechanism of common sense reasoning was deeply analyzed with the KoS/TTR framework by Breitholtz (2020).

2.3 Why use KoS to analyze silence?

The denotation of silence is extremely difficult, if not impossible, to formulate in the conventional theories of formal and computational semantics. The difficulties in the formal analysis of silence can be contemplated in three aspects: defining its

'phonetic' forms, its semantic contents, and the reasoning process behind the interpretation. Our view is that KoS (Ginzburg, 2008) is the framework that can provide a nice solution to the three problems.

Silence is difficult to define formally in terms of the 'phonetic' form. The forms of signs belonging to other classes of signs can be defined by the realization of specific patterns in certain modalities. For example, a class of speech utterances or laughs can be detected by certain patterns of the speaker's vocalization, and the act of nodding can be defined by the specific type of head movements. However, that is impossible for silence. One might be tempted to do that, but if they try to define the form of silence as the simple absence of speech for some time duration, they will get into two types of trouble immediately. First, nonverbal expressions such as head movements or gestures often replace the role of speech, and it seems inappropriate to say that silence is realized as an independent type of expression in those cases. Second, the conversational context is necessary to detect conversational silence. For example, silence has an expressive meaning when it is followed by a question to the silent person. In contrast, it seems inadequate to regard the unmarked stillness in non-conversational contexts (e.g. reading in a library) or the silence of the truthful listeners in the same light. The meaning implied by silence can be distinguished depending on the dialogical contexts such as the questions under discussion, the expected next move, or the common ground. Fortunately, the KoS framework provides a way to include the multimodal dimensions and the dialogue context in the definitions of various types of silence.

There are also numerous examples of silence whose semantic contents are difficult to analyze within the perspective of classical formal semantics in which units of analysis are propositions/sentences or worlds. Silence is interpreted as a sign of seriousness and truthfulness in some cultures while it means disengagement in the conversation or the negation of the following statement in other cultures. To formalize the semantic contents of these examples of silence, one should utilize a semantic theory on the conversation level. The KoS framework, which was born as a conversation-oriented semantic theory from the beginning, can handle them systematically by treating the contents and the effects of silence from the perspective of updating the 'dialogue gameboard.'

Lastly, the complex reasoning process that critically affects the interpretation of silence is another key obstacle to the formal understanding of silence. Consider (1) from Wang (2019). To interpret and respond correctly to the woman's silence, one has to derive the conversational implicature using background knowledge and multimodal situational information together. While the theoretical ideas in formal pragmatics such as Gricean theories or Relevance Theory in pragmatics give us a great insight into the principle of the reasoning behind conversational implicature, they do not provide the formal and explicit explanation of the reasoning mechanism at least at the level to which the KoS framework aims to attain. On the other hand, the KoS framework includes the development of the formalization of common sense reasoning during dialogues using concepts such as *enthymemes* and *topoi* (Breitholtz, 2020). Moreover, I expect that this theory of dialogue reasoning can be easily combined with the formal semantics of multimodal communication, which is also provided by the framework (Lücking and Ginzburg, 2020), and this connection is necessary to explain phenomena like (1), which is difficult to handle for most of the previous approaches in formal pragmatics. Although I will not cover the theoretical accounts for common sense reasoning during dialogues in this paper, the likelihood of productive future research strengthens our motivation to work in this direction.

3 The scope of analysis

From Kurzon (2007)'s four classes of silence, I will focus on conversational silence. Textual and situational silence have relatively minor importance in linguistic accounts of silence because they play no roles in verbal communication and they seem to be outside of the realm of dialogues generally. Even when they take part in the situational environments of conversations in some cases (for instance, a prayer interrupting a conversation and referred to by the participants afterward), the update mechanism of dialogue gameboard seems to be unclear and much harder to grasp at least in the current KoS framework. I also excluded thematic silence from our scope because it is not a prototypical case of silence where speech is not being produced. Although I anticipate that they may be analyzed in a similar manner to this paper in the future. they require another paper to be properly covered.

There are some problems to be resolved in set-

ting the definition and the scope of analysis for conversational analysis as well. There may be several different notions of silence that may produce confusion. For example, nodding, head-shaking, or pointing one's finger without speaking can constitute a sufficient answer to a question in many cases. The basic intuition here is that they should not be considered as the most basic cases of conversational silence despite the lack of speech because they convey the messages in non-phonetic *tiers* and therefore it is the nodding, the head-shaking, and the gestures, rather than silence, that convey the messages. I will call this notion of silence "conversational silence in the narrow sense," which requires no locutionary acts to be produced in any modalities. On the other hand, there is plenty of academic literature on silence in which silence is treated as something that can be combined with facial expressions or gestures. I will call their notion of silence "conversational silence in the broad sense," which requires only the vocal tiers to be empty. There seems to be a considerable range of phenomena where conversational silence and signals in non-verbal tiers interact to produce an additional meaning that cannot be explicated by silence or non-verbal signals alone and the broader notion of conversational silence is required. Nevertheless, I leave this for future research and will simply focus on the explication of the purer forms of silence by setting silence's 'phonetic' events as the absence of any markedly active 'phonetic' signals in all of the *tiers* in a dialogue here.

Among Levinson (1983)'s three classes of conversational silence(turn, inter-turn, intra-turn), I focus on turn and inter-turn silence. Pause (intra-turn silence) is excluded from our study because it is affected by diverse variables, including processing difficulties (Goldman-Eisler, 1958), breathing (Werner, 2023), and prosodic planning (Krivokapić, 2007), which make it an incomparably harder subject for formal linguistic description.

4 Turn silence

Turn silence is a type of conversational silence realized when a participant produces no overt signals while being expected to say something in their given turn. In conversational analysis, the expectations that create the condition of turn silence are explained in terms of a 'turn-taking' system and adjacency pairs. For example, a question from the current speaker is supposed to be followed by the answer from the current listener, a statement or a request by its acceptance (or rejection), a greeting by the counter-greeting, a calling by a response, or a complaint by the excuse or remedy. These sequentially and functionally related pairs of speech acts are called *adjacency pairs*. The turn exchange that takes place is processed by a conversational rule called 2-PTEP (Ginzburg, 2008), which is described in Figure 1. [2]

$$
\begin{bmatrix}
\text{preconds} & : & \begin{bmatrix}
\text{spkr} & : \text{Ind} \\
\text{addr} & : \text{Ind} \\
\text{P1} & : \text{IllocRel} \\
\text{LatestMove} & \\
= \text{P1(spkr,addr)} & : \text{IllocProp} \\
\text{qud} & : \text{poset(Ques)} \\
\text{facts} & : \text{set(Prop)}
\end{bmatrix} \\
\text{effects} & : & \begin{bmatrix}
\text{spkr} = \text{preconds.addr} \\
\text{addr} = \text{preconds.spkr} \\
\text{P2} \quad : \text{IllocRel} \\
c_1 \quad : \text{AdjPair(P1,P2)} \\
\text{Moves} \\
= [\text{P2(spkr,addr)}| \\
\quad \text{preconds.Moves}] \\
\quad : \text{list(IllocProc)} \\
\text{qud} = \text{preconds.qud} \\
\quad : \text{poset(ques)} \\
\text{facts} = \text{preconds.facts} \\
\quad : \text{set(Prop)}
\end{bmatrix}
\end{bmatrix}
$$

Figure 1: 2-PTEP

Propositional or illocutionary contents are given to turn silence. I argue that this is an adequate treatment considering several examples. (1) and (2) demonstrate that turn silence can generate conversational implicature, and conversational implicature is evidence of the existence of propositional contents (or *what is said*) because a conversational implicature occurs based on the meaning of *what is said* or the literal semantic contents (Harnish et al., 1976, pp. 339–341; Levinson, 1983, p. 113; Wilson and Sperber, 1981, p. 160). (3) and (4) provide

[2] There might arise a question of whether it is appropriate to assign the role of *spkr* to the silent actor when *spkr*'s phonation is empty or not, considering *spkr* in the narrower sense is defined as the one who produces verbal signals on the speech tier(Lücking and Ginzburg, 2020, p. 10). However, *spkr* without that narrower restriction is just an object typed as Ind and seems to be correctly handled by the pre-existing conversational rules. Therefore, I avoided uneconomical decisions here

examples where the propositional contents of turn silence are accepted and refuted respectively. (5) shows that silence can give rise to intended meaning clarification requests.

(2) A: Come on! I know that you are more generous than this.

B: [silence]

A: I'm just kidding.

+> B did not want to respond because A said something inappropriate.

(3) A: Am I disturbing you?

B: [silence]

A: Okay.

(4) A: Is it you who broke the coffee machine?

B: [silence]

A: No, I won't let you reimburse me for that. Just tell me frankly.

(5) A: Did you watch the last episode of [TV series]? It was better than I expected!

B: [silence]

A: Why are you saying nothing? Were you disappointed as a fan of the series?

I adopted a simplified version of Kurzon (1995)'s modal interpretation of silence, accepting the silence of "I can't speak" and the silence of "I do not want to speak," and discarding others. Acceptance silence was newly added as a distinct class of turn silence, which is outside of Kurzon (1995)'s modal interpretation and is strongly motivated by the dialogue structure of the KoS framework. Turn-passing silence that occurs in group conversations is discussed after the other types of turn silence.

4.1 Inability silence

A participant who is supposed to speak in a given turn sometimes fails to speak due to being unable to speak something at the very moment. This type of silence, which is caused by the participant's inability to speak, is called *inability silence*. Because it is not an intentional sign, it is inappropriate to explain its signification by a lexical item for it. Unlike refusal silence, there is no locutionary act because the silent person has no choice and no *moves* are added. Instead, the silent person's inability to express themself is discovered or inferred by other participants from the contexts. On the other hand,

$$
\begin{bmatrix}
\text{tcs} & : \begin{bmatrix} \text{dgb} & : \text{DGBType} \\ \text{private} & : \text{Private} \end{bmatrix} \\
\text{B} & = \text{dbg.spkr} : \text{Ind} \\
\text{B.preconds} & : \begin{bmatrix}
\text{spkr} & : \text{Ind} \\
\text{addr} & : \text{Ind} \\
\text{P1} & : \text{IllocRel} \\
\text{P2} & : \text{IllocRel} \\
\text{LatestMove} & : \text{P1(spkr, addr)} \\
c_1 & : \text{AdjPair(P1,P2)} \\
c_2 & : \begin{bmatrix} \text{addr.silence.length} \\ > \theta^{l}_{unable} \end{bmatrix}
\end{bmatrix} \\
\text{B.effects} & : \begin{bmatrix}
o & : \text{Ind} \\
c_3 & : \text{About}(o, \\
\multicolumn{2}{l}{\text{AbleToRespond(addr)?)}} \\
\multicolumn{2}{l}{\text{VisSit.InAttention}} \\
\multicolumn{2}{l}{= o : \text{Ind}}
\end{bmatrix}
\end{bmatrix}
$$

Figure 2: CheckInability - The conversational rule to deal with inability silence

we normally tend to check whether there are abnormal symptoms indicating a person's inability to answer when the person is supposed to answer but showing no responses. This reaction pattern can be expressed as a conversational rule, which can be roughly expressed as Figure 2.

θ^{l}_{unable} is the time threshold for the detection of inability silence, and if the time interval of silence is not sufficiently long, the conversational rule is not applied. In this paper, all types of silence have their own time thresholds, written as $\theta_{silence-type}$. $\theta^{u}_{silence-type}$ and $\theta^{l}_{silence-type}$ are the lower and the upper bound for the duration of silence, respectively.

4.2 Refusal silence

Refusal silence is a type of silence in which the silent agent expresses that they do not want to answer or express anything. Unlike inability silence, it generates illocutionary acts, updating the Moves in the dialogue gameboard. The lexical entry for refusal silence can be expressed as Figure 3.

Finding the reason why the silent person does not want to express anything is an important part of communication involving refusal silence. When it is failed, clarification requests on the intention can rise as (6). This is done by common sense reasoning, which is formulated using *enthymemes* and *topoi* in the KoS framework. The detailed

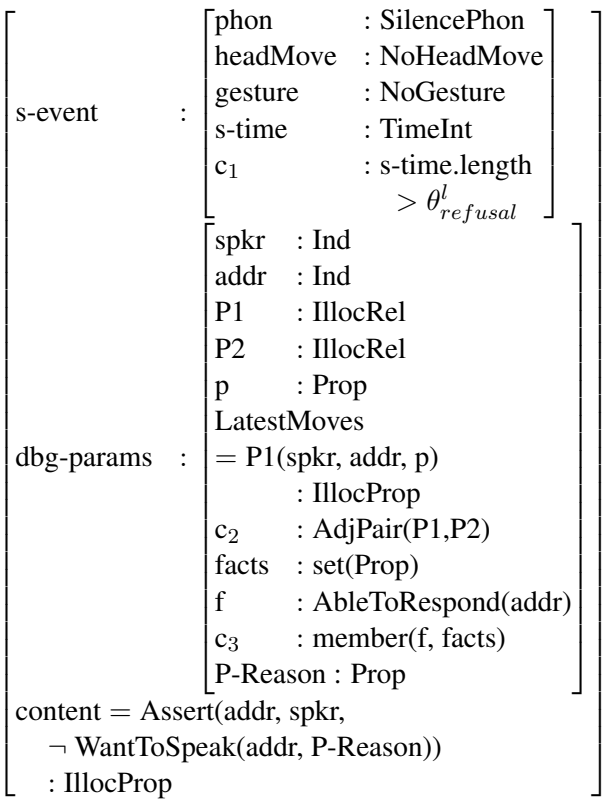


Figure 3: The lexical entry for refusal silence

mechanism is beyond the scope of this paper, but let's analyze a situation as an example.

(6) A: Can you help me with my homework? It is super hard for me!

B: [silence]

A: Why aren't you answering? Are you still angry at me because I ate the last piece of the cookies?

B: Yes.

A: Come on. I'll buy new ones for you.

4.3 Acceptance silence

Acceptance silence is motivated by the following examples where the questions of p? are resolved by silence.

(7) A: Am I disturbing you?

B: [silence]

A: [silence]

(8) A: I will open the window because it's hot here.

B: [silence]

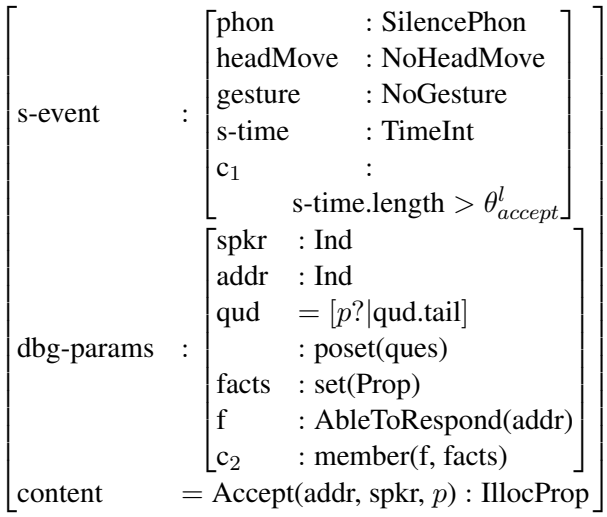


Figure 4: The lexical entry for acceptance silence

A: [A opens the window]

In the KoS framework, an act of assertion 'Assert(spkr, addr, p)' adds p? to the qud and this question should be resolved by either an accepting p or other p?-specific remarks, including the rebuttal of p. Therefore, if we assume that B's silence in (7) is an asserting move (refusal silence) in the dialogue, A's silence should be an accepting move towards B's silence. This gives us the motivation to distinguish these two types of silence: refusal silence and *acceptance silence*. B's silence in (8) shows that acceptance silence can be realized not only after another silence but also after an overt speech. The lexical entry for acceptance silence can be expressed as Figure 4.

5 Inter-turn silence (Gap)

Inter-turn silence is a subclass of silence that does not constitute an independent turn but takes place between other turns. Unlike turn silence, the semantic analysis of inter-turn silence cannot be analyzed as an independent speech act or move in a dialogue. Here, I suggest three examples of inter-turn silence: truthfulness silence, unwillingness silence, and awkward silence. This should not be considered to be a complete list of the sub-classes of inter-turn silence because the number of the classes may increase due to some possible discoveries in the future.

$$\begin{bmatrix} \text{s-event} & : & \begin{bmatrix} \text{phon} & : \text{SilencePhon} \\ \text{headMove} & : \text{NoHeadMove} \\ \text{gesture} & : \text{NoGesture} \\ \text{s-time} & : \text{TimeInt} \\ c_1 & : \theta^u_{truthful} \\ & \quad > \text{s-time.length} \\ & \quad > \theta^l_{truthful} \end{bmatrix} \\ \text{dbg-params} & : & \begin{bmatrix} \text{spkr} & : \text{Ind} \\ \text{addr} & : \text{Ind} \\ \text{facts} & : \text{set(Prop)} \\ \text{f} & : \\ & \quad \text{AbleToRespond(addr)} \\ c_2 & : \text{member(f, facts)} \\ \text{NextMove} & : \text{IllocProp} \\ c_3 & : \\ & \quad \text{NextMove.spkr} = \text{addr} \end{bmatrix} \\ \text{content} & = \text{Truthful(addr, NextMove)} : \text{Prop} \end{bmatrix}$$

Figure 5: The lexical entry for truthfulness silence

$$\begin{bmatrix} \text{s-event} & : & \begin{bmatrix} \text{phon} & : \text{SilencePhon} \\ \text{headMove} & : \text{NoHeadMove} \\ \text{gesture} & : \text{NoGesture} \\ \text{s-time} & : \text{TimeInt} \\ c_1 & : \theta^u_{unwilling} \\ & \quad > \text{s-time.length} \\ & \quad > \theta^l_{unwilling} \end{bmatrix} \\ \text{dbg-params} & : & \begin{bmatrix} \text{spkr} & : \text{Ind} \\ \text{addr} & : \text{Ind} \\ \text{facts} & : \text{set(Prop)} \\ \text{f} & : \\ & \quad \text{AbleToRespond(addr)} \\ c_2 & : \text{member(f, facts)} \\ \text{NextMove} & : \text{IllocProp} \\ c_3 & : \text{NextMove.spkr} = \text{addr} \\ c_4 & : \text{NextMove.IllocRel} \\ & \quad = \text{Accept} \end{bmatrix} \\ \text{content} & = \text{Unwilling(addr, NextMove)} : \text{Prop} \end{bmatrix}$$

Figure 6: The lexical entry for unwillingness gap

5.1 Truthfulness silence

Literature on the silence culture has reported that silence conveys truthfulness in some cultures such as Japan(Lebra, 1987; Saville-Troike, 1985) and Jordan(İbrahim, 2013) while it is not the case in the English-speaking world and the Latin American culture, and this difference often produces intercultural miscommunication(Nitta, 1987; Nakane, 2007; Brannen, 1997). These examples of miscommunication give us the motivation to formulate this type of silence as a sign lexically encoded depending on the culture. For example, we can explain the American people's misunderstanding of Japanese people's silence by the lack of truthfulness silence in their lexicon, which results in interpreting them as an awkward silence or refusal silence. The following table is the lexical entry for the silence indicating truthfulness that takes place between turns. Unlike Refusal silence, its truthfulness is about the future move that will be performed by the currently silent person. For this reason, it is inherently inter-turn silence preparing for the following turn, and if the silent ends their turn only with silence (e.g. not answering the question at all), one of the lexical conditions (c_3) cannot be satisfied, and thus the silence is not interpreted as a sign of truthfulness anymore.

There should be some conversational rules for connecting truthfulness to the next move, and I expect the rules can be made in a similar manner to the conversational rules for finding the affiliates of co-speech gestures as suggested by Alahverdzhieva (2013) and Lücking and Ginzburg (2020). However, I will not cover a detailed analysis of them in this paper.

5.2 Unwillingness gap

Depending on the culture and the context, the silence before answers can indicate reluctance, unwillingness, half-heartedness, or even the negation of the subsequent answers. According to Wang (2019), "in Philippines when an electric appliance such as TV or water heater does not work and the owner calls an electrician, the electrician who keeps silent for a while on the phone and then promises to come will not come at all and the owner will just call another one." This is not true in many other cultures. The cultural difference surrounding this type of silence can be analyzed either by setting a distinct lexical item of silence for 'unwillingness gaps,' or by assigning a different set of topoi and enthymemes depending on the culture. Figure 6 is the lexical entry for unwillingness gap.

5.3 Awkward silence

Even when there are no questions under discussion and no responses anticipated, silence can create an uncomfortable mood if the conversation does not

$$\begin{bmatrix} \text{s-event} & : & \begin{bmatrix} \text{phon} & : \text{SilencePhon} \\ \text{headMove} & : \text{NoHeadMove} \\ \text{gesture} & : \text{NoGesture} \\ \text{s-time} & : \text{TimeInt} \\ c_1 & : \text{s-time.length} > \theta^{l}_{awkward} \end{bmatrix} \\ \text{dbg-params} & : & \begin{bmatrix} \text{participants} : \text{set(Ind)} \\ \text{qud} = \emptyset : \text{poset(Ques)} \\ \text{facts} : \text{set(Prop)} \\ \text{f} : \forall a \in \text{participants}.\ \text{AbleToSpeak(a)} \\ c_3 : \text{member(f, facts)} \end{bmatrix} \\ \text{content} = \text{Awkward(participants, } \delta) : \text{Prop} \end{bmatrix}$$

Figure 7: The lexical entry for awkward silence

$$\begin{bmatrix} \text{preconds} & : & \begin{bmatrix} \text{LatestMove} : \\ \text{Awkward(participants, } \delta) \end{bmatrix} \\ \text{effects} & : & \begin{bmatrix} \text{NegativePleasantnessIncr}(\delta, \epsilon) \\ \text{.effect} \\ \text{Mood.Power.arousal} = 0 \end{bmatrix} \end{bmatrix}$$

Figure 8: AwkwardnessIncr - The conversational rule for the increase of awkwardness

continue smoothly. I call this type of silence as *awkward silence*, which is lexically encoded as in Figure 7. The intensity of awkwardness, written as δ, may be different by culture. Once an awkward silence is added to Moves, a conversational rule expressed in Figure 8, AwkwardnessIncr is applied. As a result, the pve (positive-value excitement) is reduced toward 0 and the nve (negative-value excitement) is increased toward δ. The AwkardnessIncr utilizes NegativePleasantnessIncr formulated in Mazzocconi (2019) as in Figure 9.

6 Conclusion

In this paper, I analyzed the forms and meanings of various types of silence with the KoS framework. In the history of semantics, most of the formal analyses have been showing their weakness later and refuted or counterargued by other researchers. I do not expect our analyses suggested in this paper to be faultless or complete. However, I believe that building formal analyses strict enough that can be rebutted and improved is an essential part of scientific research and its progress. I anticipate

$$\begin{bmatrix} \text{preconds} & : & \begin{bmatrix} \text{LatestMove.cont} \\ : \text{IllocProp} \end{bmatrix} \\ \text{effects} & : & \begin{bmatrix} \text{Mood.pleasant.affect.nve} \\ = \epsilon(\text{preconds.Mood.} \\ \quad \text{pleasant.affect.nve}) \\ \quad +(1-\epsilon)\delta \\ : \text{Real} \\ \text{Mood.pleasant.affect.pve} \\ = \epsilon(\text{preconds.Mood.} \\ \quad \text{pleasant.affect.pve}) \\ : \text{Real} \end{bmatrix} \end{bmatrix}$$

Figure 9: NegativePleasantnessIncr - The conversational rule for NegativePleasantnessIncr

that this direction of research can contribute to the scientific understanding of silence and intercultural differences in communication and the development of general-purpose dialogue systems that parse and understand human dialogues in the future.

References

Katya Alahverdzhieva. 2013. Alignment of speech and co-speech gesture in a constraint-based grammar. The University of Edinburgh.

Christalyn Brannen. 1997. *Going to Japan on Business: Protocol, Strategies, and Language for the Corporate Traveler*. Stone Bridge Press.

Ellen Breitholtz. 2020. *Enthymemes and topoi in dialogue: the use of common sense reasoning in conversation*. Brill.

Robin Cooper and Jonathan Ginzburg. 2015. TTR for natural language semantics. *Handbook of contemporary semantic theory*, 2:375–407.

Michal Ephratt. 2007. On silence–introduction. In *Silences–Silence in Culture and in Interpersonal Relations*, pages 7–25. Resling.

Michal Ephratt. 2008. The functions of silence. *Journal of pragmatics*, 40(11):1909–1938.

Kennan Ferguson. 2003. Silence: A politics. *Contemporary Political Theory*, 2:49–65.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the 1st international workshop on computational semantics, Tilburg: ITK, Tilburg University*.

Jonathan Ginzburg. 2008. Semantics for conversation. *Studies in Computational Linguistics, CSLI Publications, Stanford*.

Jonathan Ginzburg. 2010. Relevance for dialogue. In *SemDial: Workshop on the Semantics and Pragmatics of Dialogue (PozDial)*, pages 121–129.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1).

Cheryl Glenn. 2004. *Unspoken: A rhetoric of silence*. SIU Press.

Frieda Goldman-Eisler. 1958. The predictability of words in context and the length of pauses in speech. *Language and speech*, 1(3):226–231.

Robert Harnish et al. 1976. Logical form and implicature. *An integrated theory of linguistic ability*, pages 313–92.

ABU-SHİHAB İbrahim. 2013. Forms and functions of communicative silence in Jordan. *Eskiyeni*, (27):147–153.

Adam Jaworski. 2011. *Silence: interdisciplinary perspectives*, volume 10. Walter de Gruyter.

Jelena Krivokapić. 2007. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of phonetics*, 35(2):162–179.

Dennis Kurzon. 1995. The right of silence: A sociopragmatic model of interpretation. *Journal of pragmatics*, 23(1):55–69.

Dennis Kurzon. 2007. Towards a typology of silence. *Journal of pragmatics*, 39(10):1673–1688.

Takie Sugiyama Lebra. 1987. The cultural significance of silence in Japanese communication. Walter de Gruyter, Berlin/New York Berlin, New York.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge university press.

Andy Lücking and Jonathan Ginzburg. 2020. Towards the score of communication. *a. A*, 1:T1.

Chiara Mazzocconi. 2019. *Laughter in interaction: semantics, pragmatics, and child development*. Ph.D. thesis, Université Paris Cité.

Ikuko Nakane. 2007. Silence in intercultural communication. *Perceptions and performance. John Benjamin's Publishing Company*.

Fumiteru Nitta. 1987. "A Flower For You": Patterns of interaction between Japanese tourists and Hare Krishna. In *Culture and Communication: Methodology, Behavior, Artifacts, and Institutions: Selected Proceedings from the Fifth International Conference on Culture and Communication, Temple University, 1983*, volume 3, page 83. Praeger.

Mario Perniola. 2010. Silence, the utmost in ambiguity. *CLCWeb: Comparative Literature and Culture*, 12(4):2.

Muriel Saville-Troike. 1985. The place of silence in an integrated theory of communication. *Perspectives on silence*, pages 3–18.

Chunrong Wang. 2019. A relevance-theoretic approach to turn silence. In *4th International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2019)*, pages 1078–1084. Atlantis Press.

Raphael Johannes Werner. 2023. The phonetics of speech breathing: pauses, physiology, acoustics, and perception. Saarländische Universitäts-und Landesbibliothek.

Deirdre Wilson and Dan Sperber. 1981. On Grice's theory of conversation. In *Conversation and discourse*, pages 155–178. Routledge.

Perspectives on Language Model and Human Handling of Written Disfluency and Nonliteral Meaning

Aida Tarighat
UKRI CDT in NLP
School of Informatics
University of Edinburgh
tarighat.aida@gmail.com

Patrick Sturt
Department of Psychology
School of PPLS
University of Edinburgh
patrick.sturt@ed.ac.uk

Martin Corley
Department of Psychology
School of PPLS
University of Edinburgh
martin.corley@ed.ac.uk

Abstract

When written, disfluencies are intentional. Despite frequently being considered irrelevant noise and consequently excluded from transcriptions and training data of spoken language, disfluencies are now more commonly present in online writing. While humans can process the meanings conveyed by written disfluencies, language models struggle to understand them, mainly due to being trained on filtered data. We test BERTweet's capability to make human-like predictions in fluent and disfluent cases. We find that the model performs better than expected when handling fluent sentences; however, its performance significantly worsens when the context includes a written *um*. We believe that this decline in performance is related to sarcasm. We present two, not wholly successful, reading experiments to test our theory. We suggest that incorporating disfluencies into training data could improve model performance. We invite further comment.

1 Introduction

With the advent of easy electronic communication and social media, written language has taken on a more conversational and speech-like quality (e.g., Eisenstein et al., 2014). One aspect of this change is the use of written disfluencies, such as *um*. There is disagreement on whether these tokens are produced deliberately in speech (Clark and Tree, 2002; Corley and Stewart, 2008); however, in written language, they *must* be intentionally produced. This opens up the question of what their *meaning* might be, and whether language models (LMs) and large language models (LLMs) might fail to capture that meaning, and any distinction between spoken and written disfluency.

Although, to date, LMs/LLMs have tended to treat disfluency as noise, there has been growing interest in incorporating both spoken and written disfluencies into models to enhance their performance in applications such as real-time dialogue systems (e.g., Passali et al., 2022), autonomous vehicles (e.g., Large et al., 2017), question answering systems (e.g., Gupta et al., 2021), and stuttering detection (e.g., Al-Banna et al., 2022). However, the main focus of the recent comprehension and detection studies has been on retrieving the literal meaning with regard to the 'disruption' caused by the disfluency. This approach misses the fact that disfluencies could be of potential significance in interpreting nonliteral meanings. Whereas natural language processing (NLP) studies have looked into nonliteral language understanding by focusing on idiom, metaphors, and sarcasm (e.g., D'Arcey et al., 2019; Desai et al., 2021; Hu et al., 2022; Sporleder and Li, 2009), disfluencies remain understudied.

1.1 Our Way of Approaching Disfluencies

We previously studied the use of written disfluencies (*um*, *uh*, *hmm*, *erm*, and *er*) on Twitter and found that humans rated tweets containing *um* and *hmm* as slightly more, although not significantly more, sarcastic when fillers were in tweets compared to when the fillers were excised from the same tweets. Humans also considered the tweets containing fillers to be less formal (Tarighat et al., 2022). Therefore, we aimed to investigate the potential role of the written filler *um* in signaling nonliteral meanings using a set of materials to be tested in both LMs and behavioral experiments.

Although written disfluency has not been experimentally investigated to date, a number of studies have focused on the comprehension of spoken hesitations. Fillers such as *um* and *uh* speed up the processing of the word which follows them (Corley and Hartsuiker, 2003; Fox Tree, 2001), and help with the integration of unexpected words into their discourse (Corley et al., 2007). They bias expectations toward new rather than given information (Arnold et al., 2003). Importantly, spoken fillers influence listeners' pragmatic interpre-

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Meaning-Fluency	Item: word-by-word self-paced reading experiment
literal-fluent	Well, blue whales are an endangered species; so I'd say hunting them is a really **bad** move.
sarcastic-fluent	Well, blue whales are an endangered species; so I'd say hunting them is a really **wise** move.
literal-disfluent	Well, blue whales are an endangered species; so I'd say hunting them is a really um **bad** move.
sarcastic-disfluent	Well, blue whales are an endangered species; so I'd say hunting them is a really um **wise** move.
Meaning-Fluency	**Item: masked language modeling task – cloze test – eye-tracking reading experiment**
literal-fluent	Sitting through an hour of sermon would make most children **feral** on any day. You can ask them.
nonliteral-fluent	Sitting through an hour of sermon would make most children **merry** on any day. You can ask them.
literal-disfluent	Sitting through an hour of sermon would make most children, um, **feral** on any day. You can ask them.
nonliteral-disfluent	Sitting through an hour of sermon would make most children, um, **merry** on any day. You can ask them.

Table 1: Examples of the 4 versions of an item used in the four experiments. The target counterparts are in bold: BAD - WISE and FERAL - MERRY. In the second edition of materials, commas were used to enclose *um*. In the SPR experiment, ETR experiment, and cloze test, each participant saw only one version of an item. Target words were not bold in the experiments. In the MLM task and cloze test, the word denoting the literal/nonliteral meaning was masked and replaced by a blank space.

tations, guiding them toward particular meanings (Loy et al., 2017, 2019). For example, Loy et al. (2019) showed that, in a situation where interpreting *some* in its definitional sense as encompassing *all* would cause speakers to lose face ("I ate some cookies"), listeners were more likely to make that interpretation following a disfluency.

Our hypothesis that *written* disfluency might be used to make sarcasm easier to comprehend is related to the Graded Salience hypothesis (Giora, 2003; Giora and Fein, 1999). This hypothesis suggests that humans have difficulty understanding nonliteral meaning because salient (default) meanings have cognitive priority in language comprehension, and accessing an alternative (such as an ironic or sarcastic interpretation) is cognitively effortful. In line with this suggestion, Filik et al. (2014) found N400-like effects and disruptions in eye movements when participants encountered unfamiliar ironies. We hypothesize that the use of *um* in a sarcastic context (in speaking or in writing) signals an interruption of the salient context, making it easier for listeners or readers to access the intended, nonliteral, meaning.

As computers are increasingly being used to communicate with humans, it is important that the nuances of meaning are shared between them, on the surface level at least. Although an LM does not 'understand' disfluency, if it makes different assumptions about how *um* affects the words that are likely to be produced, then it will not communicate effectively. This matters when meaning is nuanced, because achieving human-like performance in LMs increases their ability to better reflect human cognitive processes, and address the complexities of language understanding and generation.

The present study is an investigation inspired by these considerations. We wanted to know how well LMs could handle written disfluencies, whether written disfluencies could signal nonliteral meaning, and whether they could influence the ways in which readers interpret what they are reading.

Our investigation has two parts. First, we compare the performances of an LM trained on informal speech-like data and of humans in predicting nonliteral meanings in the presence of written disfluencies for a set of carefully crafted sentences. Second, we study human behavior in controlled reading experiments using the same set of sentences. Here, we present results from a masked language modeling (MLM) task with BERTweet and a cloze test, conducted to compare meaningful word prediction between the LM and humans. We also report on a self-paced reading (SPR) experiment and an eye-tracking reading (ETR) experiment designed to investigate readers' handling of written disfluency.

2 Materials

We made the materials in two rounds. There were 32 items in the first round, 24 of which we used in the SPR experiment. There were 70 items in the second round, 48 of which we used in the MLM, cloze, and ETR experiments. Table 1 shows examples of the items used in the four experiments reported in this paper.

We made 32 grammatically correct speech-like sentences, each with its literal and sarcastic variations (*If you have a butler and a nanny, your life must be* EASY (LITERAL)/HARD (SARCASTIC) *to bear.*). We then recruited 12 L1-English speakers to rate the sentences for sarcastic tone (*How sarcastic do you think the author of this sentence was being?*) on a 7-point Likert scale (*not sarcastic at all*

Item	BERTweet top word	Cloze top word (count)
1. Keep speaking nonsense and people will think you are <mask> /, um, <mask> at some point. I'm telling you.	stupid - fluent stupid - disfluent	stupid (28) - fluent stupid (31) - disfluent
2. Having to listen to people's munching noise when I am trying to eat makes me <mask>/, um, <mask> about my life. It really does.	think - fluent think - disfluent	annoyed (12) - fluent think (17) - disfluent
3. A guy in the audience kept clearing his throat throughout the whole lecture. It was a truly <mask>/, um, <mask> distraction for all of us obviously.	unnecessary - fluent painful - disfluent	annoying (24) - fluent annoying (23) - disfluent

Table 2: Fluent and disfluent example items used in the MLM and cloze tasks followed by BERTweet's and cloze top word for each fluency version. The number next to the cloze word is the count of it in 80 responses. The critical tokens denoting the literal/nonliteral meanings were removed in the two tasks: 1. STUPID (LITERAL)/BRAINY (NONLITERAL); 2. ANNOYED (LITERAL)/PLEASED (NONLITERAL); 3. DISGUSTING (LITERAL)/DELIGHTFUL (NONLITERAL). In the cloze test, each participant saw only one version of an item. The critical token appeared as a blank space to be filled in with a word.

- *definitely sarcastic*).[1] Each participant was shown only one version of each sentence. We also asked them to provide feedback on interpretability and readability of sentences. We used a sarcastic-literal mean score difference of above 2.7 as a cutoff point. We kept 24 sentences and used them in the SPR experiment (Section 5).

We made changes to the items and increased their number before using them in the MLM task, cloze test (Section 3), and ETR experiment (Section 6). The literal and nonliteral words in the two versions of each item had the same numbers of characters (MERRY/FERAL). We also counterbalanced the literal and nonliteral readings of each word across items (MERRY (LITERAL)/FERAL (NONLITERAL) and FERAL (LITERAL)/MERRY (NONLITERAL)). In the revised materials, we used commas to enclose *um*, to help with readability and increase the salience of the disfluency. Lastly, we added more words after each target word, often in the form of a short second sentence, to minimize gaze regressions out of the target interest area in the ETR experiment. To rate the newly made and edited 60 items for potential sarcastic tone on a 7-point Likert scale online,[2] we recruited 36 neurotypical L1-British-English speakers between the ages of 18 and 34 with no reported reading disorders. We only kept the counterbalanced items with a nonliteral-literal mean score difference above 2. For the items with good scores in only one reading, we repeated the procedure with 10 items rated by 20 other participants. Overall, we kept 48 counterbalanced items for the ETR experiment. There were 4 variations of each of the 48 experimental items based on meaning and fluency (Table 1).

[1]Informatics Research Ethics Process, RT number 789617.

[2]Informatics Research Ethics Process, RT number 789617.

3 BERTweet Masked Language Modeling and Cloze Test

We compared the LM and human predictions of meaningful words and how they might be influenced by written disfluencies. We expected that, given a context, the LM would perform better in predicting words in utterances which did not contain *um*.

3.1 MLM task

We first ran an MLM task on BERTweet (Nguyen et al., 2020). The tokens denoting the literal and nonliteral meanings were excised. We had 48 fluent items without *um* and 48 disfluent items with *um*, totaling 96 items. The critical tokens assigned to signify literal/nonliteral meanings in the items were masked.

We chose BERTweet due to the presence of fillers such as *um* in its training data and the higher structural similarity between the tweets and the speech-like materials we created for our experiments. We obtained BERTweet's top 10 predictions for 96 materials, using the first eligible predicted word in each list in further analyses (details below in 3.3).

3.2 Cloze test

We then conducted a cloze test using the same materials to study the humans' predictions and the possible effect of written disfluencies on their predictions. There were 48 fluent items without *um* and 48 disfluent items with *um*, totaling 96 items. The critical tokens were replaced by a blank space.

For the cloze test, we recruited 160 neurotypical L1-English participants between the ages of 18 and 34 with no reading disorders.[3] We asked them to fill in the blanks using the first word (only a single word without a space or a hyphen) that came to

[3]Informatics Research Ethics Process, RT number 789617.

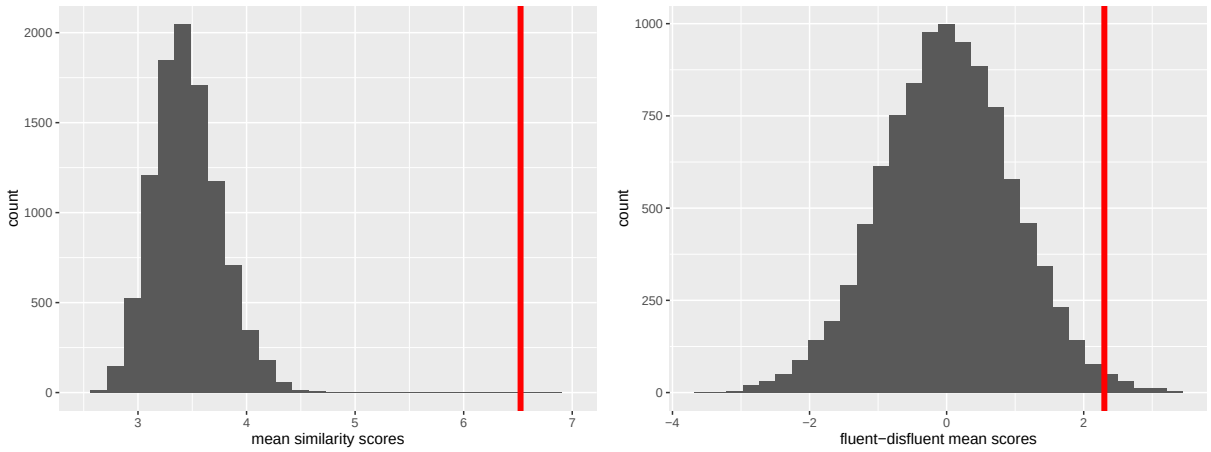


Figure 1: Left: simulated mean similarity scores, with the red vertical line indicating the mean similarity score of 6.52. Right: simulated fluent-disfluent mean scores, with the red vertical line indicating the observed fluent-disfluent difference in similarity score of 2.30. We ran 10,000 permutations of the scores to recalculate the means. BERTweet's continuations were better matches to human continuations following fluent items compared to disfluent items.

mind (autocompletion and autocorrection options were disabled on the participants' devices). Each participant saw only one version (fluent or disfluent) of an item. The participants were remunerated £3.70 for completing 48 items which took about 10 minutes on average.

3.3 Analysis

The first step was obtaining the most frequent response for each item in the cloze test to compare with the MLM predictions. However, for six items, there were ties where equal numbers of participants provided two words equally often in either the fluent or disfluent condition. To resolve the ties, we selected the word that was not used in the other condition for the relevant item. If both words were not used in the other condition, the selection was made at random. For the MLM data, we ensured that for each item, we had the most popular predicted word while adhering to the following criteria: no punctuation; no symbols (e.g., @); and no stop words such as "a", "as", "and", "be", or "not".

Next, we standardized spellings of the completions to American, and calculated Latent Semantic Analysis (LSA) cosine similarities by making pairwise comparisons using word2vec (Google News, 300 dimensions: University of Colorado). We obtained a similarity score between cloze completions and BERTweet predictions for each item by multiplying the number of identical cloze completions by the BERTweet confidence scores and then by the LSA cosine similarity between words. For example, in one item (*Well, blue whales are an endangered species. So, I'd say hunting them is a really <mask> choice environmentally speaking.*), the most popular cloze completion BAD was chosen by 30 participants, while the highest ranked LM completion was GOOD, which had a confidence rating of 0.333. The word2vec similarity score between BAD and GOOD was 0.719. Therefore, the overall score was $30 \times 0.333 \times 0.719 = 7.183$.

Calculated in this way, the mean similarity score between BERTweet and human cloze completions was 6.52. To assess BERTweet's performance against chance, we permuted the LSA and BERTweet scores 10,000 times, recalculating the mean similarity for each permutation. The visualization of these scores (Figure 1) shows that BERTweet predicted what humans would write significantly better than a baseline of random guessing ($p < .0001$). Table 2 shows fluent and disfluent example items used in the MLM and cloze tasks along with the BERTweet's and cloze top word for each fluency version.

Importantly, we also assessed the effect of fluency, by calculating the difference between mean similarity scores for fluent and disfluent items. The observed difference in similarity scores (fluent-disfluent similarity score = 2.30) was compared to the distribution of mean differences derived from 10,000 permutations of the data (Figure 1). BERTweet's continuations were better matches to human continuations following fluent items compared to disfluent items ($p = 0.0096$).

4 Behavioral Experiments

One possible explanation for BERTweet's significantly worse performance in the presence of *um* could be the role of the filler in implying nonliteral, namely sarcastic, meaning. We conducted two reading experiments to test whether written disfluencies could signal nonliteral meaning. We predicted that (1) words compatible with a sarcastic reading of a sentence (*hunting blue whales is a really* WISE *move*) should be easier to read when preceded by *um* (*really* UM WISE *move*) than when not preceded by *um*, and (2) words compatible with a literal reading of a sentence (*hunting blue whales is a really* BAD *move*) might be harder to read when preceded by *um* (*really* UM BAD *move*) than when not preceded by *um*. This leads to the prediction of an interaction between fluency and meaning, with longer reading times and/or more regressions for fluent literal items than disfluent sarcastic ones. To summarize, the disfluency *um* could signal a shift toward a nonliteral or sarcastic interpretation. Example materials for both experiments are in Table 1.

5 Experiment 1: Self-paced Reading

We implemented the online word-by-word SPR experiment (Mitchell and Green, 1978) as a moving-window reading task, using jsPsych.[4] Each item had 4 variations based on (a) meaning (whether the critical word was literal or sarcastic in context), and (b) fluency (whether the target word was preceded by *um* or not): literal-fluent, literal-disfluent, sarcastic-fluent, and sarcastic-disfluent (Table 1). In each item, we were interested in reading times for a target word and the following word (for spillover). The target word was a word selected to be [in]consistent with a sarcastic/literal interpretation. We predicted an interaction between fluency and meaning; i.e., disfluency would signal a nonliteral or sarcastic meaning whereas fluency would signal a literal meaning.

5.1 Participants and procedure

We recruited 101 L1-English, UK-based, and non-dyslexic participants through Prolific.[5] Participants were remunerated £1.75 for reading 26 items: 2 practice items and 24 experimental items. There were 4 variations of the 24 experimental items based on meaning and fluency (Table 1). Each participant read only one variation of each experimental item, pressing the space bar to reveal each new word of the sentence. Items were selected such that participants read 6 items in each experimental condition. There were 8 attention checks. Experiment settings ensured that the target word was never the last word of the sentence and was followed by at least one word. The experiment took about 10 minutes to complete.

	target		target+next	
Meaning-Fluency	**Mean**	**SD**	**Mean**	**SD**
literal-fluent	381.02	184.16	809.95	351.46
sarcastic-fluent	388.49	204.51	869.17	544.28
literal-disfluent	470.19	367.25	940.54	539.39
sarcastic-disfluent	465.15	344.96	985.18	618.05

Table 3: SPR experiment: mean and standard deviation of reading times in milliseconds for target and target+next regions.

5.2 Data preparation

We analyzed the reading time data from 99 participants. We removed 2 participants because they got fewer than 6 of 8 attention-check questions correct. Moreover, 1 item was miscoded in the experiment, resulting in 28 missing trials (1.18% of the data).

5.3 Analysis

We compared the log-transformed reading times of the target word and of the target word plus the next word (for spillover). Mean and standard deviation of reading times in milliseconds for target and target+next regions are in Table 3. Contrary to our hypotheses, words compatible with the sarcastic interpretation of the sentences were not faster to read when preceded by *um*. Maximally-fitting linear mixed-effects models only showed an effect of fluency, indicating that fluent sentences were faster to read in both literal and sarcastic versions (target, β = -0.13, SE = 0.02, $p < .001$; target+next, β = -0.10, SE = 0.01, $p < .001$). We found no interaction between meaning and fluency (target, β = 0.03, SE = 0.03, p = .27; target+next, β = 0.02, SE = 0.02, p = .31).

6 Experiment 2: Eye-tracking Reading

Whereas the SPR experiment failed to show that written disfluency indexes nonliteral meaning (at least, in the form of sarcasm), it did show that readers were sensitive to written *um*. One possibility is that the artificial segmentation needed for self-paced reading disrupted the rhythm with

[4] https://github.com/UiL-OTS-labs/jspsych-spr-mw

[5] https://prolific.co/

which readers might have read the experimental sentences, reducing any interruption effect that the traditionally spoken element *um* might have had in writing. For that reason, the ETR experiment was a replication of the SPR experiment using an eye-tracking methodology in which natural reading prosody was not disrupted. Our hypotheses were the same: (1) words compatible with a nonliteral reading of a sentence (*hunting blue whales is a really* WISE *move*) should be easier to read when preceded by *um* (*really* UM WISE *move*), and (2) words compatible with a literal reading of a sentence (*hunting blue whales is a really* BAD *move*) might be harder to read when preceded by *um* (*really* UM BAD *move*), and that this would predict longer reading times and/or more regressions for fluent (relative to disfluent) literal items, and vice versa for nonliteral items. Once again, this predicts an interaction between fluency and meaning.

We used Experiment Builder[6] version 2.4.1 to set up the experiment for presentation on an EyeLink 1000 Plus tracker for in-person data collection.

6.1 Participants and procedure

We recruited 60 neurotypical L1-English participants between the ages of 18 and 34 with normal/surgically-corrected-to-normal vision and no reported reading disorders.[7] Participants were remunerated £10 for reading 152 items: 2 practice items, 48 experimental items, and 102 filler items. Each participant read only one variation of each experimental item, selected such that they read 12 items in each experimental condition. There were 32 attention checks, 16 for experimental items and 16 for filler items. Experiment settings ensured that the target word was always followed by at least two words before a line break, and that the target word never fell at the beginning of a line and was always preceded by at least two words. The experiment took about 35 minutes to complete, and participants were given breaks after items 50 and 100.

6.2 Data preparation

We used Data Viewer[8] to prepare and summarize the eye-tracking data, and did the statistical modeling in R. Since all participants had answered 80% (26) or more of the attention checks correctly, their data was included in the analyses. Data preparation included removing the filler trials, merging nearby fixations, removing fixations less than 80 milliseconds, aligning the fixations vertically within the preassigned interest area bounds, and monitoring the number of horizontally misaligned trials for each participant for removal. If more than 20% (10) of the experimental trials for a participant needed to be removed due to severe horizontal misalignment, that participant's data was excluded from analysis. This left us with 59 participants.

	target		target+next	
Meaning-Fluency	**Mean**	**SD**	**Mean**	**SD**
regression path time				
literal-fluent	297.41	188.51	612.09	376.32
nonliteral-fluent	313.23	200.00	684.62	471.09
literal-disfluent	299.50	211.19	629.34	408.57
nonliteral-disfluent	318.34	217.30	695.69	443.70
first pass time				
literal-fluent	235.92	109.13	472.02	186.38
nonliteral-fluent	237.28	107.59	490.46	202.84
literal-disfluent	254.70	121.28	500.14	215.25
nonliteral-disfluent	268.91	126.48	538.85	217.96
total dwell time				
literal-fluent	297.40	193.84	612.73	343.77
nonliteral-fluent	329.37	201.87	681.62	386.43
literal-disfluent	323.36	195.31	652.62	358.07
nonliteral-disfluent	365.96	226.34	730.76	393.75

Table 4: ETR experiment: mean and standard deviation of regression path time, first pass time, and total dwell time in milliseconds for the target interest area (target) and the summation of target and next interest areas (target+next).

6.3 Analysis

We focused on the target and target+next interest areas and compared the log-transformed reading times for 3 measures: (1) *regression path time* (go-past time) which is the summed fixation duration from when the current interest area is first fixated until the eyes enter a later interest area; (2) *first pass time* which is the sum of the duration of all fixations before the interest area is exited for the first time; and (3) *total dwell time* which is the summed duration of all fixations on the current interest area. Table 4 shows the mean and standard deviation of the 3 measures in milliseconds for the target and target+next interest areas. We also compared the proportions of *first pass regressions out* for the target and next regions; i.e., whether regression(s) were made from the current interest area to the earlier interest area prior to leaving the interest area in a forward direction (Table 5).

Consistent with our prediction of an interaction between fluency and meaning, we expected the

[6] https://www.sr-research.com/experiment-builder/

[7] PPLS Research Ethics Committee, reference number 392-2223/1.

[8] https://www.sr-research.com/data-viewer/

presence of the word *um* to signal a nonliteral or sarcastic meaning, while fluency would signal a literal meaning. However, the results of the ETR experiment did not fully support this hypothesis, as was the case for the SPR experiment. Analyses of regression path time, first pass time, and total dwell time revealed significant effects of both fluency and meaning across the interest areas. Specifically, words signaling literal meanings were consistently read faster than those signaling nonliteral meanings, indicating an overall effect of meaning on reading behavior. Additionally, fluency also influenced reading speed, with target words generally being read faster in the fluent sentences than disfluent ones.

The maximally-fitting linear mixed-effects models of regression path time showed an effect of meaning for the target interest area (β = 0.06, SE = 0.02, p = .01), and target+next interest areas (β = 0.10, SE = 0.03, p < .001) indicating that literal words were faster to read than nonliteral ones. We found no interaction between meaning and fluency (target, β = -0.03, SE = 0.04, p = .35; target+next, β = -0.01, SE = 0.04, p = .70)

The maximally-fitting linear mixed-effects models of first pass time showed an effect of fluency in the target interest area (β = -0.10, SE = 0.02, p < .001) indicating that fluent sentences were read faster than disfluent ones. For the target+next interest areas, the models showed an effect of fluency (β = -0.07, SE = 0.01, p < .001) and one of meaning (β = 0.06, SE = 0.02, p = .004) indicating that fluent sentences were read faster than disfluent ones and that literal meanings were read faster than nonliteral ones. However, there was no interaction between meaning and fluency in target+next interest areas (target, β = -0.05, SE = 0.03, p = .06; target+next, β = -0.04, SE = 0.03, p = .11).

As for total dwell time, the maximally-fitting linear mixed-effects models showed the effects of meaning (β = 0.10, SE = 0.03, p < .001) and of fluency (β = -0.10, SE = 0.02, p < .001) for the target interest area indicating that literal meanings were read faster than nonliteral ones and that fluent items were read faster than disfluent ones. However, there was no interaction between meaning and fluency in the target interest area. The models also showed the effect of meaning (β = 0.11, SE = 0.03, p < .001) and of fluency (β = -0.07, SE = 0.02, p < .001) in the target+next interest areas indicating that literal meanings were faster to read than nonliteral ones and fluent items were read faster than disfluent ones. However, there was no interaction between meaning and fluency in the target+next interest areas (target, β = -0.03, SE = 0.04, p = .40; target+next, β = -0.01, SE = 0.03, p = .74).

	target	**next**
Meaning-Fluency	**Mean**	**Mean**
literal-fluent	0.18	0.16
nonliteral-fluent	0.19	0.21
literal-disfluent	0.08	0.16
nonliteral-disfluent	0.10	0.18

Table 5: ETR experiment: proportions of first pass regression out, i.e., the regressions that were made from the target and next interest areas to the earlier interest area prior to leaving the interest area in a forward direction.

Lastly, for the proportions of first pass regressions out, the maximally-fitting logistic mixed-effects models only showed an effect of fluency for the target interest area (β = 1.01, SE = 0.16, p < .001) indicating that regressions were more likely to be made following a fixation on the target word when the items were fluent. No other effects were reliable, for the target word or the word which followed, and there was no interaction between meaning and fluency (target, β = -0.11, SE = 0.25, p = .66; next, β = 0.25, SE = 0.22, p = .26).

The results suggest that the effects of fluency and meaning on reading behavior were independent of each other, contrary to our initial prediction of an interaction. However, it is important to note that fluency and meaning each had distinct effects on reading behavior, underscoring the complexity of their influence on comprehension.

7 Discussion

We investigated the handling of written disfluencies, which could indicate nonliteral meanings like sarcasm, by an LM and humans. We found that although BERTweet made human-like predictions, its performance was significantly worse when the disfluency *um* was present. Additionally, in our reading experiments, we found that readers were faster to read fluent sentences without *um* and sentences compatible with literal meanings rather than nonliteral or sarcastic ones. We found no interaction between fluency and meaning in the sense that disfluency did not signal a nonliteral or sarcastic meaning and fluency did not signal a literal meaning.

Our results suggest that BERTweet's difficulty

with written disfluencies may be due to training on filtered data that excludes disfluencies. The decline in performance, especially in contexts involving sarcasm, highlights the model's limitations in understanding the subtleties of human communication. Previous research has often dismissed disfluencies as irrelevant noise. However, our findings align with more recent studies that recognize the communicative value of disfluencies in online writing. The observed challenges in BERTweet's performance are consistent with other studies that highlight the limitations of LMs in NLP.

8 Limitations

Our experiments to date have investigated a specific disfluency in a specific language and context. Our results may have been influenced by the specific design and sample size. Whereas we have established that written disfluencies are worth investigating, with LMs as well as humans sensitive to their presence, this study is just a starting point. To gain a more complete picture, attention should be paid to the naturalness of the stimuli used, and work should be generalized to other languages and disfluencies.

9 Future Steps

Future studies should explore more sophisticated methods for integrating disfluencies into LM training. Our next step would involve manipulating the filler placement and removing the commas on the LM to monitor any changes in model behavior. The model could produce different output if disfluency occurred earlier in the sentence and not immediately preceding the masked token, and it would treat *um,* as a very different token from *um*. A later approach could be for us to further pre-train BERTweet using a data set of tweets containing fillers from our previous study, since its performance could potentially be improved. Then, another masked-token prediction task could follow to evaluate the model's improved ability to handle disfluencies.

Another major aspect of future research would be testing disfluencies in an LLM (e.g., Llama) to check differences and potential improvements in performance, which could be the result of the set parameters and/or training data. Since LLMs are different from BERT-type models and are increasingly preferred, it would be important to know if and how they would produce better outputs.

We would also need to compare our findings with other psycholinguistic and computational experiments that focus on licensing nonliteral interpretation. This comparison could identify strengths and weaknesses in current approaches and guide future improvements in human experiments as well as model training and evaluation, especially for developing purpose-built models and data sets for specific tasks. For instance, we know that not all humans understand disfluency in the same way (Li et al., 2022; McKenna et al., 2015), or that nonliteral and sarcastic interpretation is influenced by social and cultural factors (Katz et al., 2004). Therefore, a simple model-training approach might not work when considering how computers should interact with humans.

10 Conclusion

Our findings highlight the challenges LMs face in handling disfluencies and probably also in interpreting nonliteral meanings conveyed by disfluencies. Incorporating such elements into training data could improve model performance. Future research should explore more sophisticated methods for integrating disfluencies and other nonliteral indicators into LMs. Additionally, investigating the nuances of sarcasm detection in written text remains a promising area for further study. Well-designed behavioral experiments can capture fine-grained variations in comprehension by focusing on specific psycholinguistic features. Such evidence would be beneficial in evaluating the behaviors of models trained on large, usually written, language corpora. With more information, we can determine how and to what extent to reintroduce disfluencies into data sets.

Acknowledgments

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics, and School of Philosophy, Psychology & Language Sciences. Aida thanks Amr Keleg for his technical advice on the initial steps of the MLM task on BERTweet. Aida also thanks Oli Liu and Coleman Haley for the discussions on the validity of LM/LLM processing techniques.

References

Abedal-Kareem Al-Banna, Eran Edirisinghe, Hui Fang, and Wael Hadi. 2022. Stuttering disfluency detection using machine learning approaches. *Journal of Information & Knowledge Management*, page 2250020.

Jennifer E Arnold, Maria Fagnano, and Michael K Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of psycholinguistic research*, 32:25–36.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Martin Corley and Robert J Hartsuiker. 2003. Hesitation in speech can... um... help a listener understand. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25.

Martin Corley, Lucy J MacGregor, and David I Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.

Martin Corley and Oliver W Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2021. Nice perfume. How long did you marinate in it? Multimodal sarcasm explanation. *arXiv preprint arXiv:2112.04873*.

J Trevor D'Arcey, Shereen Oraby, and Jean E Fox Tree. 2019. Wait signals predict sarcasm in online debates. *Dialogue & Discourse*, 10(2):56–78.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Ruth Filik, Hartmut Leuthold, Katie Wallington, and Jemma Page. 2014. Testing theories of irony processing using eye-tracking and erps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):811.

Jean E Fox Tree. 2001. Listeners' uses ofum anduh in speech comprehension. *Memory & cognition*, 29(2):320–326.

Rachel Giora. 2003. *On our mind: Salience, context, and figurative language*. Oxford University Press.

Rachel Giora and Ofer Fein. 1999. Irony comprehension: The graded salience hypothesis. *Humor: International Journal of Humor Research*, 12:425–436.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-qa: A benchmark dataset for understanding disfluencies in question answering. *arXiv preprint arXiv:2106.04016*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Albert N Katz, Dawn G Blasko, and Victoria A Kazmerski. 2004. Saying what you don't mean: Social influences on sarcastic language processing. *Current Directions in Psychological Science*, 13(5):186–189.

David R Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics*, 63:53–61.

Wei Li, Hannah Rohde, and Martin Corley. 2022. Veritable untruths: Autistic traits and the processing of deception. *Journal of Autism and Developmental Disorders*, 52(11):4921–4930.

Jia E Loy, Hannah Rohde, and Martin Corley. 2017. Effects of disfluency in online interpretation of deception. *Cognitive Science*, 41:1434–1456.

Jia E Loy, Hannah Rohde, and Martin Corley. 2019. Real-time social reasoning: the effect of disfluency on the meaning of some. *Journal of Cultural Cognitive Science*, 3(2):159–173.

Peter E McKenna, Alexandra Glass, Gnanathusharan Rajendran, and Martin Corley. 2015. Strange words: Autistic traits and the processing of non-literal language. *Journal of autism and developmental disorders*, 45:3606–3612.

Don C Mitchell and David W Green. 1978. The effects of context and content on immediate processing in reading. *The Quarterly Journal of Experimental Psychology*, 30(4):609–636.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

University of Colorado. Word embedding analysis tools. Accessed: 2024-05-27.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. Lard: Large-scale artificial disfluency generation. *arXiv preprint arXiv:2201.05041*.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762.

Fatemeh S Tarighat, Walid Magdy, and Martin Corley. 2022. Understanding fillers may facilitate automatic

sarcasm comprehension: A structural analysis of twitter data and a participant study. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 215–217.

Disfluencies in conversation: a comparison of utterances with and without metaphors

Amy Han Qiu[1], Vanessa Vanzan[1], Chara Soupiona[2], and Christine Howes[1]

[1]Department of Philosophy, Linguistics and Theory of Science,
Faculty of Humanities, University of Gothenburg
[2]Department of Philology, Division of Linguistics, University of Crete

Abstract

Disfluencies are pervasive in conversations and commonly regarded as indicative of cognitive difficulties. However, they have rarely been examined in utterances with metaphors, which are considered to be more cognitively challenging than those without any metaphors. In this paper, we investigate the occurrence of filled pauses and self-repairs in conversational turns with and without metaphors, across various word counts. Results showed that metaphor presence and word count contributed significantly to the probabilities of filled pauses and self-repair. Notably, there was a significant interaction between metaphor presence and word count, highlighting the combined cognitive demands elicited by using metaphors and producing longer utterances as key factors influencing disfluencies in spontaneous conversations.

1 Introduction

Disfluencies are often characterised as disruptions or breaks in the flow of communication, such as hesitations, pauses, filled pauses, and self-repairs such as repetitions or reformulations. These "interruptions" occur commonly in everyday interactions and impact how language is conveyed and interpreted. Self-repairs and filled pauses reflect incremental processing, with real-time adjustments made word-by-word, as the speaker progresses through the utterance. Following the incremental view, disfluencies are natural byproducts of the dynamic processes involved in generating speech. Disfluencies may occur in different cases, for example cognitive difficulties (Levelt, 1983; Bortfeld et al., 2001; Clark and Tree, 2002), heightened attention of the ongoing communication (Cienki, 2020), and interactive issues (Goodwin, 1979).

Disfluencies have been extensively explored in various linguistic dimensions, for example word-related features like word class, utterance features like utterance types and sentence lengths, and conversational dynamics like speaker exchange (Shriberg, 1996). However, they were rarely examined for their relationships with the use of metaphors, which involves talking and potentially thinking about something in terms of something else (Semino, 2008).

Processing and producing metaphors are typically assumed to demand extra cognitive resources due to the inherent complexity in cross-domain mappings. The mapping of features between two distinct domains, as well as the wording, may lead to heightened cognitive pressure (Lakoff and Johnson, 1980; Steen, 2023), which makes metaphor use an intriguing yet unexplored avenue for exploring the conversational dynamics of disfluencies.

Below is an example of disfluencies in metaphor use, cited from a conversation about an ethical dilemma of sacrificing one person to save more people (metaphorical parts in bold type and disfluency markers italicised):

(1) "*Bu- bu- but* are you *s- saying* that *um uh* we need to **value** the sort of the **worth** of each person?"

In this example, the importance of a person is interpreted in terms of the financial worth of a property. The utterance is characterised by disfluencies, indicated by repetitions ("bu- bu-"), filled pauses ("um uh"), and self-repair ("s-saying"). They interrupt the flow of speech and may indicate uncertainty, interactive issues or even difficulty in articulating the intended message smoothly.

Despite the well-acknowledged link between cognitive pressure and disfluencies (Levelt, 1983; Bortfeld et al., 2001; Clark and Tree, 2002), whether the cognitive complexity associated with metaphor use contributes to the probabilities of disfluencies in an utterance remains an interesting research question.

Based on the transcripts of 19 triadic conversations, this study compares the probability of

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

filled pauses and self-repairs in utterances with and without metaphors. Disfluencies in metaphor use, which is a linguistic phenomenon characterized by inherent cognitive complexities, could provide insights into the interaction between metaphor, word count, and different types of disfluencies.

1.1 Filled pauses and self-repairs

Filled pauses and self-repairs are two of the most studied forms of disfluencies (Clark and Tree, 2002). Below is an example:

(2) *Mmm*, is there *any any* other line of thought, that we can think?

Filled pauses like *"mmm"* in example (2) can serve as markers of language processing, indicating moments of word retrieval, linguistic uncertainty and speech planning (Clark and Tree, 2002).

Self-repair in metaphorical dialogues reveals how speakers manage errors and refine their language in real-time communication. Unlike filled pauses, self-repairs like repetitions (*"any any"*) specifically involve the speaker interrupting their ongoing speech to correct or revise what they have just said. Self-repairs play a crucial role in maintaining shared understanding and mutual interpretation in effective communication (Clark, 1996). Additionally, self-repairs contribute to the negotiation of meaning between speakers, as they indicate the awareness of one speaker of the other's comprehension needs and their willingness to clarify or elaborate on their message.

Some studies showed that disfluencies can occur due to heightened cognitive pressure. Previous research has examined disfluencies in different utterance types (Oviatt, 1995; Shriberg, 1996; Lickley, 2001). Longer and more syntactically complex turns were found to have a higher frequency of repetition disfluencies, and giving instructions or expressing uncertainty when answering questions was associated with a greater use of filled pauses. Similar patterns have been found at the beginning of utterances, where cognitive pressure is assumed to be high due to speech planning. Some empirical studies found that disfluencies can serve as a compensatory cognitive strategy, aiding the speaker to manage the cognitive load in conversation (Brennan and Schober, 2001; Bailey and Ferreira, 2007; Howes et al., 2017).

Disfluencies could also arise for interactive reasons, assisting the interlocutors in adjusting their communicative comprehension strategies. For example, self-repair can reflect the speaker's intention to maintain their turn and to regulate the flow of conversation (Goodwin, 1979). When an interlocutor is puzzled or needs more time for speech planning, filled pauses can facilitate smoother communication by affording a longer time for accommodating these cognitive challenges.

There are also studies that linked disfluencies with the speaker's awareness of the ongoing communication. According to Cienki (2020), the occurrence of disfluencies can be a key signal of the speaker's awareness of the impact of their language use on the hearer, reflecting their "metacommunicative awareness (MCA)"[1] in conversations. The more effortful the speech is, the stronger the MCA can be (Cienki, 2020).

1.2 Disfluencies in metaphor use

An interesting phenomenon often overlooked in disfluency research is the use of metaphors. According to Kaal (2012), 2.9% to 10.1% of lexical units in conversations are metaphor related. Below is an example of conversational turns with metaphorical lexical units:

(3) He is quite **far away from** a breakthrough.

In example (3) the lexical units "**far**" "**away**" and "**from**" are metaphorically deployed to signify physical distance from achieving success, which may introduce a layer of cognitive complexity to the interpretation process. By contrast, utterances like "He is unlikely to achieve a breakthrough". is a direct account of the low probability of achieving success, without linking to a more basic meaning.

Utterances containing metaphors are typically regarded as more cognitively challenging than those without any metaphors (Gibbs, 1994). According to Lakoff and Johnson (1980); Steen (2023), processing metaphor presumably requires extra cognitive resources due to the need for inferential work and the mapping of complex relationships between domains. This argument finds support in a recent study, which showed that when word count is controlled for, speakers invest a longer time articulating an idea with metaphors compared to those without metaphors (Qiu et al., 2024, in progress). Neuroscientific research has also shown

[1] Other possible MCA signals include gestures, prosodic markers like stress, marked intonation, and use of pauses, verbal cues like modification, diversification, extension, literalisation, etc, (refer to Cienki, 2020 for more details).

that metaphor production, compared with the use of literal language, involves more intense cognitive work, and an increase of brain activation grows with the increase of creativity in the metaphors (Benedek et al., 2014).

As noted earlier, heightened cognitive pressure elicited by complicated linguistic tasks and producing long utterances may be major contributors to increased disfluencies in spontaneous speech (Levelt, 1983; Bortfeld et al., 2001; Clark and Tree, 2002). Focusing specifically on filled pauses and self-repairs, this paper explores whether the inherent cognitive complexity associated with metaphor use adds to the probability of disfluencies in conversation turns.

2 Research questions

This paper aims to address the following research questions:

1. What is the difference between utterances with and without metaphors in the probabilities of containing filled pauses?

2. What is the difference between utterances with and without metaphors in the probability of self-repairs?

3. Does metaphor presence interact with word count in terms of the occurrence of filled pauses or self-repairs?

According to previous research on the inferential processes involved in metaphor use (Steen, 2023; Benedek et al., 2014) and research on disfluencies (Lavelle et al., 2012; Bortfeld et al., 2001; Oviatt, 1995; Shriberg, 1996), metaphor processing imposes greater cognitive demands and should thereby lead to more disfluencies, while longer sentences similarly increase cognitive load and disfluency rates (Bortfeld et al., 2001). We therefore hypothesise that:

1. Utterances containing metaphors are more likely to contain filled pauses compared to utterances not containing metaphors

2. Utterances containing metaphors are more likely to contain self-repairs compared to utterances not containing metaphors

3. When metaphors are present, longer turns will be associated with an increased frequency of self-repairs and filled pauses compared to turns without metaphors.

3 Methods

3.1 Data

The data consists of 19 face-to-face triadic conversations between 57 participants who were unfamiliar with each other. The conversations were video and audio recorded, lasting from 5 to 10 minutes each. The data were collected earlier as the control condition in an experiment comparing conversations among healthy people to those involving a patient with schizophrenia. The participants were instructed to discuss the balloon task, an ethical dilemma in which one of the four hot air balloon passengers should sacrifice themselves by jumping out to their certain death in order to save the other three. The data collection procedure and other details are reported in Lavelle et al. (2012) and Howes and Lavelle (2023).

This study examines conversation utterances at the level of turns, which refers to all sub-utterances, segmented by filled pauses, unfilled pauses, laughters, etc., produced by one speaker before the next speaker starts to talk. Turns may vary in length; some turns may comprise multiple sub-utterances and are thus longer than others.

The 19 conversations consisted of $3,785$ turns, among which 849 turns contained only laughter, cough, unclear utterances, or backchannels (e.g., "yeah", "ummm", "okay"). As including these turns may inflate the number of utterances without metaphors, they were filtered out from further analysis. $2,631$ turns were preserved, which include a total of $24,476$ words. The mean of total word count per conversations is 1288.21 (95%CI: 1060.85 - 1515.58). The mean word count of each turn is 9.28 (95%CI: 8.92 - 9.69).

3.2 Disfluencies Identification

Filled pauses were identified manually based on a find-and-replace operation on inconsistently spelt cases (see Howes et al. 2017 for more details).

Self-repairs were identified with STIR (STrongly Incremental Repair detection), an automatic incremental self-repair detection system (Hough and Purver, 2014). STIR was trained and initially tested on the Switchboard corpus of telephone conversations (Godfrey et al., 1992). The system has a high accuracy rate and high correlations with human coders in detecting self-repair rates (Howes et al., 2014).

Although the numbers of disfluencies detected

in each turn are available, convergence issues were found when running the statistical models. Therefore, filled pauses and self-repair were annotated as binary variables based on whether a disfluency marker of the relevant type was present in the turn. These annotations were taken directly from Howes et al. (2017).

3.3 Metaphor Identification

Metaphorically used lexical units were annotated manually following the Metaphor Identification Procedure VU (MIPVU; Steen et al., 2010). The criteria for identifying metaphoricity is whether the word has a more basic meaning that is "more concrete, body-related, more precise, or historically older" (Steen et al., 2010), and whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.

For the present study, each lexical unit was annotated based on its basic meaning and contextual meaning provided by three dictionaries, i.e., the Longman Dictionary, the Oxford English Dictionary, and WordNet. Annotations provided by the VUAMC[2] (Steen et al., 2010), the largest available corpus hand-annotated for metaphorical language use, were used as references to enhance inter-reliability. The presence of metaphors was annotated at the level of turns as a binary variable. Turns that contained at least one metaphorically used lexical unit were annotated as metaphorical, and those without any metaphorically used lexical units were annotated as non-metaphorical.

To test the inter-rater reliability, two annotators worked independently on 10% of randomly selected data. The annotations reached 97.1% agreement (Cohen's kappa = 0.88). More details about metaphor identification and inter-rater reliability checks are reported in Qiu et al. (2024, in progress).

Following this approach, 690 turns were identified as containing metaphorically used lexical units, and 1,941 as not containing any metaphors.

3.4 Statistical Methods

We ran a series of binomial Generalized Linear Mixed Models (GLMMs). A random intercept assigned for conversation groups was included in the models to account for potential correlation among observations within the same group. In this study, we compared models that took word count as an interaction term and those with word count as the co-variate. For cases where the interaction effect was significant, we compared the effect of word count on the two levels of metaphor presence with further stratified analyses.[3]

$P < 0.05$ was set as the threshold of statistical significance for all models. The analyses were run with the lmer function from the lme4 package of R.

4 Results and Discussion

4.1 Descriptive statistics

Among the 2,631 turns, 282 contained filled pauses and 551 contained self-repair. The overall rate of filled pause presence is 10.72%, and that of self-repair presence 20.94%. In utterances with and without metaphors, the proportion of filled pauses presence are 19.42% and 7.62%, and the proportion of self-repair presence are 30.43% and 17.57%. Both disfluency markers occurred more frequently in utterances with metaphors. Descriptive statistics are summarized in Table 1.

Metaphor	**Filled pauses**		**Self-repair**		**Total**
Presence	Yes	No	Yes	No	
Yes	134	556	210	480	690
No	148	1793	341	1600	1941
Total	282	2349	551	2080	2631

Table 1: Filled pauses and self-repair presence in the dataset

The probability of disfluency increases with word count. The Biserial correlation between word count and the presence of filled pauses is 0.3 (p<.01) and that between word count and the presence of self-repairs is 0.38 (p<.01).

Presence of filled pauses

The modelling results are summarised in Table 2. Both metaphor presence and word count have a significant main effect. In particular, when word count is held constant at 9.30, utterances with metaphors are more likely to contain filled pauses compared to utterances without metaphors. According to

[2]http://www.vismet.org/metcor/documentation/home.html (last accessed May 26, 2024).

[3]Including a random slope for word count and adding Participant ID as nested in Group caused singularity, which makes the options unfeasible.

the co-variate model, the predicted probabilities of filled pauses in the two utterance types are 12% ($95\%CI : 9\% - 16\%$) and 8% ($95\%CI 6\% - 10\%$), respectively.

The interaction effect between metaphor presence and word count on the presence of filled pauses is also significant. To compare the impact of word count on the two levels of metaphor presence, stratified analyses were performed (see Table 3). In both cases, word count has a significant effect on self-repair, with a more pronounced impact observed on utterances without metaphors. As word count increases by one unit, the probability of filled pauses increases more in utterances without metaphors (by approximately 0.075 units) than in utterances with metaphors (by about 0.048 units).

The predicted probabilities of filled pauses in the two levels of metaphor presence are plotted in Figure 1.

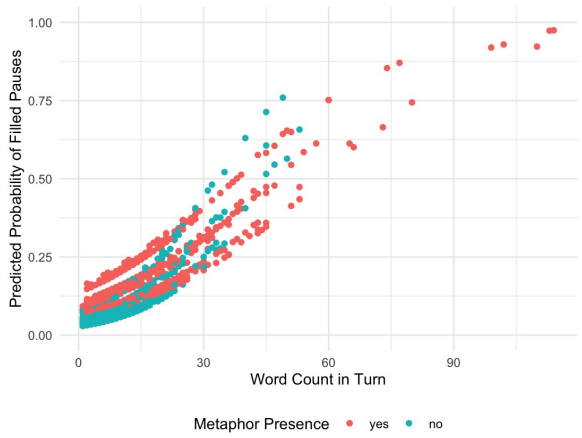


Figure 1: Predicted probabilities of filled pauses

When word count is below 30, utterances with metaphors are more likely to contain filled pauses compared to those without any metaphors. When word count is between 30 and 50, the probability of filled pauses in utterances without metaphors surpasses that in utterances with metaphors. When the word count goes beyond 50, the probability in utterances with metaphors continues to increased at a lower rate; however, no utterances without metaphors with comparable lengths were found in this range.

Presence of self-repair

The modelling results on the presence of self-repair are summarized in Table 4. Both metaphor presence and word count have significant fixed effects. When word count is held constant at 9.30, utterances with metaphors are significantly more likely to contain self-repair compared to those without metaphors. According to the co-variate model, the predicted probabilities of self-repair are 11% ($95\%CI : 9\% - 15\%$) and 8% ($95\%CI, 6\% - 10\%$).

The interaction effect between word count and metaphor presence on the presence of self-repair is also significant. Results of stratified analyses are summarized in Table 5. We can see that word count has a significant effect on self-repair in both cases, and the impact is more pronounced on utterances without metaphors. As word count increases by one unit, the probability of self-repair in utterances without metaphors increases more (by approximately 0.133 units) than in utterances with metaphors (by about 0.075 units).

The predicted probabilities of self-repair in the two levels of metaphor presence are plotted in Figure 2.

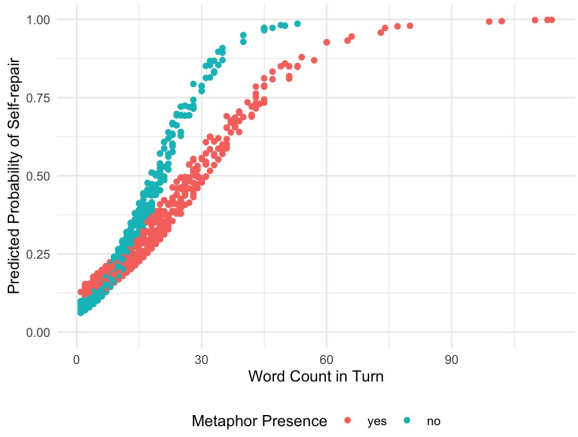


Figure 2: Predicted probabilities of self-repair

When the word count is below 10, utterances with metaphors generally have a higher probability of containing self-repairs than utterances without metaphors. When the word count exceeds 10, utterances without metaphors have a higher probability of containing self-repair.

Discussion

The significant main effect of metaphor presence on the probabilities of filled pauses and self-repair highlights its contributions to the occurrence of disfluencies. When word count is held constant, utterances with metaphors are associated with a heightened likelihood of filled pauses, which manifests as increased hesitation and interruptions in the speech flow. Utterances with metaphors also

Fixed Effects	Estimate	SE	z-value	p-value
(Intercept)	-2.391	0.205	-11.639	$< .001$
metaphor presence	-0.791	0.218	-3.625	0.01
word count	0.047	0.007	6.356	$< .001$
metaphor presence * word count	0.029	0.012	2.318	$< .001$

Table 2: Fixed effects of the interaction model on filled pauses

Metaphor presence	Fixed Effects	Estimate	SE	z-value	p-value
N	(Intercept)	-3.164	0.163	-19.362	$< .001$
	word count	0.075	0.010	7.344	$< .001$
Y	(Intercept)	-2.387	0.209	-11.40	$< .001$
	word count	0.049	0.008	6.44	$< .001$

Table 3: Stratified analysis of filled pauses in metaphor presence

have higher rates of self-repair, which plays out in the form of repetitions, substitutions, and deletions.

The occurrence of disfluency markers can provide insights into cognitive processing involved in metaphor production. Our results support the view that utterances with metaphors, compared to those without metaphors, may pose heightened cognitive demands on the speaker's end (Steen, 2023). As mentioned earlier, some previous studies take disfluencies as indicative of cognitive burdens or communication problems (Levelt, 1983; Colman and Healey, 2011), and some see disfluencies as a communicative solution to manage the cognitive pressure (Brennan and Schober, 2001; Bailey and Ferreira, 2007; Howes et al., 2017). Based on our results, it is plausible the increased cognitive demands associated with metaphor use requires more cognitive resources, potentially resulting in higher disfluencies rates.

Disfluencies may also be related to the speaker's consciousness over language use in conversation. Cienki (2020) proposed the concept of metacommunicative awareness (MCA) to account for the speaker's degree of awareness of the form and/or content of their language use. Disfluency markers are recognised as key signals of MCA. When the signals are present, compared to cases with less effortful or no signals, the speaker is more likely to be aware of their ways of self-expression (Cienki, 2020). Following this line of thought, utterances with metaphors, given the increased disfluency rates, may be produced with higher degrees of MCA compared to those without metaphors. The heightened occurrence of disfluency markers, as exemplified above, may reflect the speaker's active engagement in shaping and refining their linguistic choices to effectively convey the complicated ideas.

Another interesting observation is that the disfluency markers are not necessarily attached to the metaphorical parts of the utterance. Rather, they may occur before or after, and sometimes quite far away from the metaphorical parts; example (1) presented earlier is illustrative. This suggests that the cognitive pressure may have arisen before the utterance is articulated, and may influenced the entire production process.

The findings offer clues regarding the interactive relationship between metaphor presence and word count in terms of the occurrence of disfluencies. Interestingly, the patterns differ across the two types of disfluencies. From Figure 1 we see that utterances containing metaphors have higher rates of filled pauses than those without metaphors. However, utterances without metaphors increase more sharply in filled pause rates than those with metaphors, especially when word count goes above 30. The presence of self-repair, as shown by Figure 2, exhibits a different pattern. Despite the fact that utterances without metaphors have higher

Fixed Effects	Estimate	SE	z-value	p-value
(Intercept)	-2.095	0.178	-11.787	$< .001$
metaphor presence	-0.512	0.200	-2.563	0.01
word count	0.073	0.008	8.788	$< .001$
metaphor presence *word count	0.059	0.013	4.527	$< .001$

Table 4: Fixed effects of the interaction model on self-repair

Metaphor presence	Fixed Effects	Estimate	SE	z-value	p-value
N	(Intercept)	-2.614	0.126	-20.72	$< .001$
	word count	0.133	0.010	12.88	$< .001$
Y	(Intercept)	-2.102	0.188	-11.198	$< .001$
	word count	0.075	0.009	8.61	$< .001$

Table 5: Stratified analyses on self-repair in the two levels of metaphor presence

self-repair rates when word count is held constant, the trend shifts when considering interactions between metaphor presence and word count. Notably, when word count is above 10 words, those without metaphors have generally higher rates of containing self-repair than utterances with metaphors. However, the majority of utterances without metaphors are shorter than 10 words ($1,579$ out of $1,942$) and have lower self-repair rates, which explains the main effect discussed earlier. This finding suggests that different linguistic variables may interact in shaping conversation behaviors, underscoring the need for disfluency research to consider the impact of word count, especially its interaction with other linguistic variables.

Consistent with previous findings (e.g., Oviatt, 1995; Bortfeld et al., 2001), the positive association between word count and the probabilities of disfluencies confirms that the cognitive effort involved in articulating longer utterances is higher, regardless of the presence of metaphors. More interestingly, our results also show that the impact of word count on disfluency rates is less prominent on longer utterances with metaphors. It is possible that the production of longer utterances with metaphors requires more deliberate planning and articulation, which leads to relatively lower disfluency rates. This might also be explained by the presence of compensatory cognitive strategies. A recent study (Qiu et al., 2024, in progress) showed that speakers may employ more compensatory cognitive strategies, such as gestures, in turns with metaphors than those without. These strategies were found to help to alleviate the speakers' cognitive pressure (Kita, 2000) and sustain mutual understanding (Healey et al., 2015). While these strategies may happen at a higher chance in longer utterances, it is possible that they mitigate the impact of cognitive difficulties, resulting in lower disfluency rates.

5 Conclusion

This study compared the probabilities of disfluencies in naturally produced conversational utterances with and without metaphors, taking the impact of word count into account. The findings offer insights into the conversational dynamics in metaphor use and the cognitive mechanisms underlying disfluencies. A strength of the study is that it captures how people talk in everyday life, which would be hard to replicate and control for in psycholinguistic experiments. We also have supportive evidence that the production of turns containing metaphors may pose greater cognitive challenges than those without metaphors.

Several key limitations need to be acknowledged. Firstly, even though utterances examined by this study are thematically consistent, it was not possible to control utterances in spontaneous conversations in terms of semantic content and lengths.

Future studies could consider using experimental designs to compare utterances with and without metaphors on the same topic and of similar word count.

Secondly, this paper focused exclusively on differences in the probabilities of disfluency markers. The placement of disfluency markers, especially in utterances with metaphors, remains to be explored by future research. Furthermore, we do not distinguish between different types of self-repair, for example, whether repetitions or reformulations are more associated with turns containing metaphors. Additionally, more fine-grained analysis distinguishing between, for example "forward-looking" and "backwards-looking" disfluencies (Ginzburg et al., 2014), remains for future work.

Thirdly, in this study, metaphor presence was annotated as a binary variable. In fact, there are some more fine-grained aspects of metaphors that may cause the utterance to be processed with different levels of ease, for example, the number of metaphorical lexical units, the degree of novelty/conventionality (Giora, 2002), and deliberateness of metaphor use (Reijnierse et al., 2018). Future research could explore how these features interact with disfluencies and other aspects of language use. This can be investigated either in spontaneous conversation, or with more controlled psycholinguistic methods like the tangram experiments in Clark and Wilkes-Gibbs (1986).

Despite these limitations, our results show that both word count and metaphor presence are significant factors contributing to the presence of disfluencies. Utterances with metaphors are generally more likely to contain filled pauses and self-repairs compared to those without metaphors. This may stem from heightened cognitive or communicative challenges associated with metaphor use, or potentially reflect the speaker's increased awareness of language use in the conversation (Cienki, 2020). Interestingly, the impact of word count on disfluencies varies between utterances with and without metaphors and across different disfluency markers, highlighting the combined influence of metaphor use and longer utterances on speech disfluencies.

Acknowledgements

This research was supported by ERC Starting Grant DivCon: Divergence and convergence in dialogue: The dynamic management of mismatches (101077927). Howes was further supported by the Swedish Research Council grant (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We are also thankful for the comments from the anonymous reviewers and the advice on statistics from Asad Sayeed.

References

Karl GD Bailey and Fernanda Ferreira. 2007. The processing of filled pause disfluencies in the visual world. In *Eye movements*, pages 487–502. Elsevier.

Mathias Benedek, Roger Beaty, Emanuel Jauk, Karl Koschutnig, Andreas Fink, Paul J Silvia, Beate Dunst, and Aljoscha C Neubauer. 2014. Creating metaphors: The neural basis of figurative language production. *NeuroImage*, 90:99–106.

Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.

Susan E Brennan and Michael F Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of memory and language*, 44(2):274–296.

Alan Cienki. 2020. A multimodal perspective on mca: Cues of (possible) metacommunicative awareness. In *Drawing attention to metaphor: Case studies across time periods, cultures and modalities*, pages 63–92. John Benjamins Publishing Company.

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

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

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

M. Colman and P. G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

Raymond W Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

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

Rachel Giora. 2002. Literal vs. figurative language: Different or equal? *Journal of pragmatics*, 34(4):487–506.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. 1:517–520.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, 97:101–121.

Patrick Healey, Nicola Plant, and Mary Howes, Christine amd Lavelle. 2015. When words fail: Collaborative gestures during clarification dialogues. In *Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium*, Austin, Texas, USA.

Julian Hough and Matthew Purver. 2014. Strongly incremental repair detection. *arXiv preprint arXiv:1408.6788*.

Christine Howes, Julian Hough, Matthew Purver, and Rose McCabe. 2014. Helping, I mean assessing psychiatric communication: An application of incremental self-repair detection.

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*. In press.

Christine Howes, Mary Lavelle, Patrick G. T. Healey, Julian Hough, and Rose McCabe. 2017. Disfluencies in dialogues with patients with schizophrenia. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, UK.

Anna Albertha Kaal. 2012. *Metaphor in conversation*. Phd-thesis - research and graduation internal, Vrije Universiteit Amsterdam.

Sotaro Kita. 2000. How representational gestures help speaking. *Language and gesture*, 1:162–185.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Mary Lavelle, Patrick G. T. Healey, and Rose McCabe. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin*.

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

Robin J Lickley. 2001. Dialogue moves and disfluency rates. In *ISCA tutorial and research workshop (ITRW) on disfluency in spontaneous speech*.

Sharon Oviatt. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1):19–36.

Amy Han Qiu, Christine Howes, and Vladislav Maraev. 2024. Metaphoricity, deliberateness and conversation behaviours: A case study of behavioral patterns in the use of linguistic metaphors. Work in Progress.

W Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J Steen. 2018. Dmip: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2:129–147.

Elena Semino. 2008. *Metaphor in discourse*. Cambridge University Press Cambridge.

Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of international conference on spoken language processing*, volume 96, pages 11–14. Citeseer.

Gerard J Steen. 2023. Thinking by metaphor, fast and slow: Deliberate metaphor theory offers a new model for metaphor and its comprehension. *Frontiers in Psychology*, 14:1242888.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, Tryntje Pasma, et al. 2010. *A method for linguistic metaphor identification*. John Benjamins Publishing Company Amsterdam.

Speaker transition patterns in German: A comparison between task-based and casual conversation in face-to-face and remote conversation

Qiang Xia, Marcin Włodarczak
Department of Linguistics
Stockholm University
{qiang.xia,wlodarczak}@ling.su.se

Emer Gilmartin
Inria, Paris
gilmare@tcd.ie

Abstract

The study describes floor transition patterns in free and task-oriented 'spot the difference' conversations by 10 pairs of German native speakers. Each floor transition was delimited by stretches of longer (> 1 s) intervals of solo speech and included an arbitrary number of intervening intervals corresponding to silences, overlaps and shorter stretches of solo speech. While the effect of video conferencing was minor, the type of task had a large effect on the turn-taking patterns. Compared to the free conversation, the task-oriented dialogues were characterised by more frequent speaker changes, particularly short transitions involving a single gap. In addition, within-speaker transitions with three intervening intervals were very common in this condition, especially those in which the interlocutor provided shorter verbal contributions, possibly corresponding to feedback expressions.

1 Introduction

Although widely described as the fundamental mechanism of spoken interaction, turn-taking is still not very clearly understood. Spoken interaction can vary in multiple ways, including number of speakers involved, purpose, register, setting, and medium. It is likely that the temporal arrangement of speech also varies depending on factors such as those mentioned above. In this study, we address this problem by examining the arrangement of contributions by participants in German task-based and free (casual) conversations held face-to-face and remotely over the Internet.

We base our analyses on *floor state dynamics*, where spoken interaction is represented as a series of floor state intervals, describing who is speaking or remains silent at a particular time. The floor state changes constantly throughout the interaction, and sequences of floor states, or *floor state transitions*, capture speech activity patterns, facilitating a data-driven method to analyse the local dynamics of turn-taking in different types of spoken interaction. They can be used to describe turn-taking patterns of arbitrary complexity and provide a convenient starting point for more specific investigations of conversational structure and content.

We perform a within-subject comparison of the floor state dynamics of conversations from a subset of the Berlin Dialogue Corpus (BeDiaCo), version 2 (Belz et al., 2021), where pairs of German speakers engaged in two conversation types (task-free casual conversation, and 'spot the difference' or Diapix tasks) in two sessions – face-to-face and over an internet connection.

2 Background

In this section, we briefly discuss the contextual factors that might condition the emerging patterns of turn-taking in conversation, including the effects of videoconferencing and the organisation of conversation floor, both of which are of direct interest to this study. We also introduce the paradigm used to describe floor transitions used in this work.

2.1 Contextual effects on turn-taking

Even though Sacks et al. (1974) made it abundantly clear that their turn-taking model did not necessarily apply to all speech exchange systems, much of the work on conversational turn-taking adopts the assumption that "overwhelmingly, one party talks at a time" (Sacks et al., 1974, p. 700) as one of the underlying principles of all verbal interaction. However, this is not necessarily the case as the rules governing the temporal arrangement of turns depend on contextual factors such as task, medium and speakers' familiarity (O'Connell et al., 1990).

In particular, Edelsky (1981) demonstrated that in addition to the "one-speaker-at-a-time" model, conversation floor can also be collaborative with several interlocutors engaging in a "free-for-all" state. In a collaborative constructed floor, turn

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

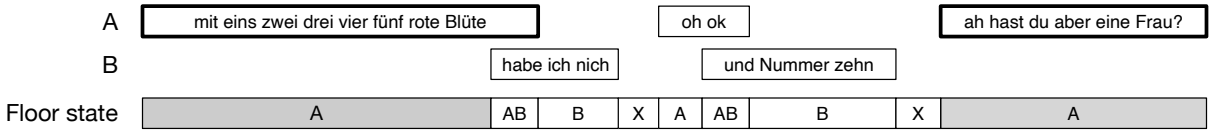


Figure 1: Example of a between-speaker transition. The two top rows represent speakers' talkspurts (*A: With one two three four five red flowers; B: I don't have it; A: oh ok; B: and number ten; A: ah but do you see a woman there?*). The third row represents the floor state with solo-speech intervals longer than one second marked in grey.

length is more evenly distributed compared to a single-floor model, and overlapping speech is considered a sign of participants' active engagement in a shared conversational space. Similarly, Tannen (1980) found that high involvement in conversation is characterized by high speech rate, rapid turn-taking with short gaps and frequent overlaps.

In addition, while our understanding of turn-taking mechanisms and conversational style is predominantly based on face-to-face (and, to some extent, telephone) conversations, the effect of the medium can potentially have a strong effect on temporal patterns of turn exchange. As a case in point, electronically-mediated remote conversations are characterised by an unavoidable electronic transmission delay, which might disrupt the rhythm of conversational turn-taking, causing longer response time in answering polar questions (Boland et al., 2021). Egger-Lampl et al. (2010) found a positive correlation between conversational interactivity and speakers' sensitivity to delay impairments. They demonstrated that in highly interactive telephone conversations, such as random number verification, fewer speaker changes take place under long-delay conditions than under short-delay conditions. This suggests that latency may affect speakers' ability to predict the turn end and they may change their turn-taking behaviours depending on the conversational condition. Indeed, Bailenson (2021) hypothesised that in video conferencing interactions interlocutors need to work harder to send and receive turn-taking cues, which might explain the "Zoom fatigue" reported by some users.

In the present study, we compare speaker transition patterns in conversations characterized by high and low interactivity by contrasting free conversations and the Diapix task (Van Engen et al., 2010; Bullock and Sell, 2022), a spot-the-difference game where participants are each given similar pictures which contain a number of differences and try to find all differences through speech alone. Baker and Hazan (2011) examined Diapix interactions and concluded that it is a valid method for eliciting balanced speech contribution in dyadic conversations. This task allows researchers to analyze conversational dynamics in a controlled but naturalistic setting, providing insights into how participants manage turn-taking in collaborative dialogues. We additionally investigate the effect of the medium by having the same participants conducting both types of interaction face-to-face and using video-conferencing software.

2.2 Analysis paradigm

The analysis of turn-taking patterns in large conversational corpora has a long tradition going back to the seminal work on telephone speech by Norwine and Murphy (1938); Brady (1968); Jaffe and Feldstein (1970), which describes floor transition phenomena in terms of probabilities of solo speech, silence and overlap sequences. This line of research has proven useful for describing temporal properties of turn-taking patterns in interaction (Heldner and Edlund, 2010) and for identifying differences between interactional settings, such as face-to-face and telephone interaction (ten Bosch et al., 2004, 2005). Furthermore, machine learning on speech and silence data from large corpora of dyadic and multiparty speech has been successfully used to infer information about spoken interaction, for example, predicting speaker activity from conversation history (Jaffe et al., 1964; Beebe et al., 1988, 2000), inferring information such as relationships between participants, genre, and features such as personality traits of speakers in dyadic and multiparty interaction (Laskowski, 2011; Gilpin et al., 2018). However, much of this work is built on two assumptions which do not make justice to the complexity of the conversational turn-taking. First, it considers any transition between non-overlapping intervals, however short, as potentially meaningful. Second, it implicitly assumes that speaker change and retention are achieved within a scope of a single interval of silence or overlap.

Consider, for instance, Figure 1, which shows an excerpt from a dyadic conversation. There are nine

floor states – solo speech (three *A*s and two *B*s), overlaps (two *AB*s) and silence (two *X*s). Existing data-driven approaches to turn-taking could treat this stretch as a series of four transitions: two instances of *A_AB_B* from A to B, and two instances of *B_X_A* from B to A. However, looking at the transcript and the speech patterns, it seems more likely that the *longer* stretches of solo speech by speaker A delimit a single more complex transition with A retaining the conversational floor. Such larger conversational structures are routinely overlooked by large-scale corpus studies.

2.3 Floor state transitions

A more detailed approach to describing floor transitions like those in Figure 1 was proposed in Gilmartin (2021). In this approach, longer sequences of speech and silence were captured by concatenating floor state intervals. Floor state transitions were identified as the sequence of intervals between stretches of solo (single-party) speech in the clear (without overlap), in order to gain insight into how turn change and retention is managed by participants. To approximate turn changes or retention, a minimum duration threshold was placed on the solo speech intervals leading into and out of the transitions. Transitions were classified as within- or between-speaker (WST and BST, respectively), depending on whether speaker change occurred or whether the same speaker continued.

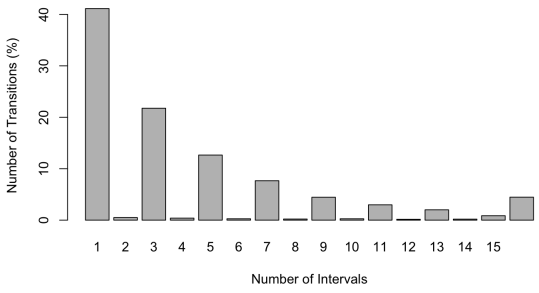


Figure 2: Frequency of transitions with different numbers of intervening intervals in a corpus of casual (free) conversation, reproduced from Gilmartin (2021).

This approach was used by Gilmartin et al. (2020), who identified turn transitions in 24 multiparty conversations in English, Estonian and Swedish. Each transition was characterised in terms of the number of *intervening intervals* (i.e. silences, overlaps and shorter stretches of solo speech) it took to complete a turn transition. The study found that the distribution of floor transitions was similar to that in Figure 2 with 95% of transitions completed in fewer than 16 intervening intervals. One-interval transitions (i.e. consisting of a single instance of silence or overlap) were the most frequent but they nevertheless accounted for less than 40% of all transitions, suggesting that existing accounts of turn-taking might miss much of floor change dynamics. In addition, transitions involving even numbers of intervals were vanishingly rare, due to the very low likelihood of two or more participants starting or stopping at exactly the same moment.

The composition of transitions in Swedish, Estonian and English in terms of incidence and duration of silent, overlapping and solo-speech intervening overlaps was investigated in Włodarczak and Gilmartin (2021). They found that while one-interval transitions are predominantly silent, more complex patterns of speech and silence were more likely with increasing number of intervening intervals. Overlaps in particular became more common as the number of intervening intervals increased, particularly in BSTs. Similarly, longer transitions were found to involve increasingly many interlocutors speaking, with participation by all three speakers more likely in BST than WST. The authors demonstrated that the most common three-interval transitions (which account for about 21% of transitions identified) were similar across the three data sets, both in terms of interval types and in terms of their percentage frequencies. In other words, even though the transitions are quite complex (especially as the number of intervening intervals increases), a relatively small number of labels accounted for a substantial portion of all floor transitions found. A later study on dyadic phone conversations in the Switchboard corpus found that the transition distribution in Switchboard's conversations broadly followed patterns found in multiparty talk, but that there are fewer complex transitions observed.

3 Method

Below we describe the data used, segmentation and processing into floor state transitions.

3.1 Data

The present investigation is based on a subset of the Berlin Dialogue Corpus (BeDiaCo), version 2 (Belz et al., 2021). The material consisted of free talk and task-oriented interactions between 10 pairs

of German native speakers (mean age = 25.7, SD = 3.8, 10 females, 10 males) in two conditions: face-to-face and remote (Zoom-mediated) conversations. Each of the speaker pairs was living together at the time of the recording.

The conversations were recorded in the phonetics laboratory of the Humboldt Universität zu Berlin. In the face-to-face condition, participants sat opposite each other in a sound-attenuated booth and wore neckband headsets (Beyerdynamics Opus 54) to record their speech. In the remote condition, they were located in adjacent offices and spoke to each other via Zoom installed on two tablets (Lenovo; 10.1 inch). Both tablets were connected to the Internet through the university's wireless network (Eduroam). Subjects wore headphones to listen to each other and their speech was recorded by additional microphones placed in the room (Sennheiser Me62, Sennheiser Me64).

The free conversation had participants talking about self-selected topics (e.g., one's favourite place in Berlin, plans for the next holiday) for about 10 minutes. During the task-oriented part, the speakers participated in the Diapix task. The participants were given about 10 minutes to locate 10–13 differences between their pictures.

The participants came to the lab to be recorded twice, with about a week between sessions. In each session, participants solved two Diapix tasks with a free conversation in between via one medium. For each session and speaker pair, the order of conversation media and Diapix tasks was randomised.

According to the post-experiment questionnaire, 13 of the 20 participants reported using Zoom on a "daily" or "weekly" basis, the others "monthly" or "never". 15 of the participants were "comfortable" or "very comfortable" engaging in Zoom interactions, while five were "neither comfortable nor uncomfortable" (Belz et al., 2021).

3.2 Processing - Identifying speaker transitions

Intervals of speech and silence in each speaker's recording were reconstructed from manually corrected word alignments distributed with the corpus, assembled into talkspurts (or interpausal units, IPUs), given a minimum silence threshold of 200 ms.

The resulting talkspurt segmentation was then used to identify floor state intervals, i.e. divide the conversation into continuous segments where a particular subset of speakers is active. Possible floor states include solo speech by one speaker, intervals of overlapping speech by two speakers, or general silence. More generally, for a conversation with n speakers, there are 2^n possible floor state labels - general silence, n different solo speech labels, and various combinations of speakers in overlap.

In the next step, speaker transitions were identified by locating instances of solo speech of at least one second in duration and recording the floor state intervals between those. Each transition was classified as WST or BST and was characterised by the number of intervening intervals it contained.

4 Results

The corpus comprised 8451 floor transitions (floors defined as talkspurts longer than one second) across 60 conversations. As shown in Figure 3, single-speaker floor constitutes the majority of the data, accounting for 69.2% of the conversation time, followed by silent intervals (23.7%) and overlapping speech by two speakers (7.09%). 3520 between-speaker and 4931 within-speaker transitions were found.

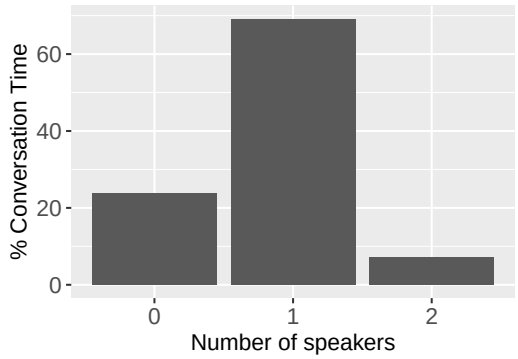


Figure 3: Distribution of the conversation time by the number of speakers.

4.1 General transition patterns

Figure 4 illustrates the percentage of different numbers of between- and within-speaker intervals in Diapix and free conversations in face-to-face (ftf) and Zoom interactions. All groups have more than 98% of transitions completed in less than 15 intervening intervals (Diapix_ftf: 98.25%, Diapix_zoom: 98.26%, free_ftf: 98.50%, free_zoom: 99.34%). In general, the greater the number of intervening intervals involved in the transition, the less frequent they are in the data. For a given number of intervening intervals, there are usually more instances

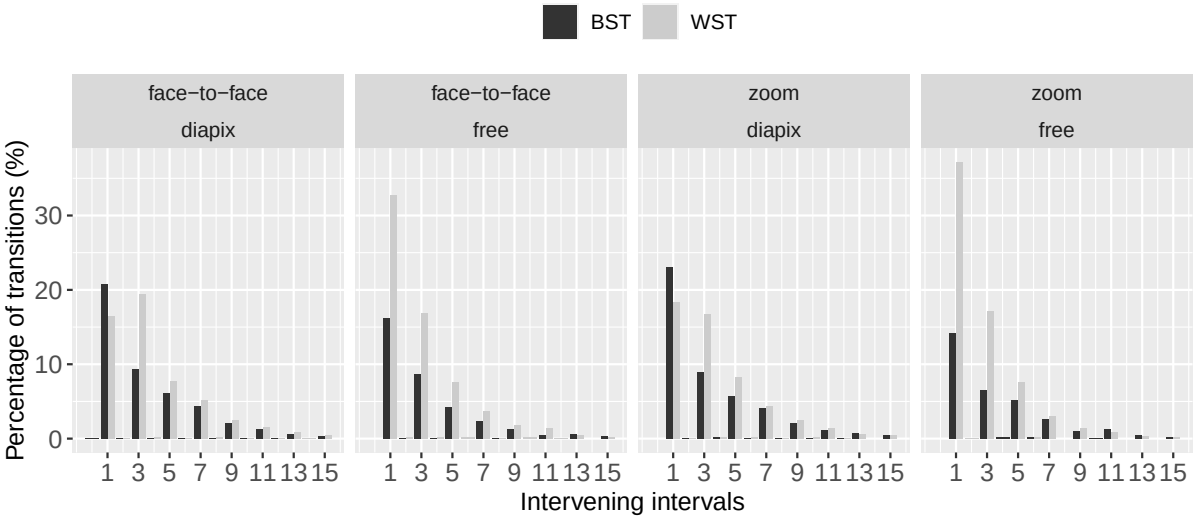


Figure 4: Frequency distribution of speaker transitions in face-to-face and Zoom interactions depending on the number of intervening intervals.

of WST than BST within the group.

Transitions including even numbers of intervening intervals constitute only 0.01% of the data. Such transitions entail two speakers starting or stopping at exactly the same time, with zero gaps and zero overlaps in transition, which is extremely unlikely given the granularity of the manually corrected IPU segmentation.

The cumulative distribution of transitions completed in fewer than 15 intervening intervals is shown in Figure 5. Notably, the difference between the cumulative percentages within each group with the same number of intervening intervals is greater when broken down by task (left panel) than by medium (right panel). Transitions with one intervening interval account for 50.02% of all transitions in free conversation, much higher than the 39.11% in Diapix tasks. Transitions with three to seven intervening intervals exhibit a similar tendency toward a cumulative percentage higher by some distance in free conversations than Diapix. No big differences are found in the cumulative distribution divided by medium, for example, 41.08% for one-interval transitions in face-to-face interactions and 44.55% for Zoom.

In total, 58.35% of transitions are WST. Only 14 conversations have a BST-to-WST ratio above 1, all from Diapix tasks (Figure 6). Compared to free conversations, Diapix tasks have a significantly higher proportion of BST, indicating the Diapix conversations are indeed more interactive and characterised by more frequent speaker change.

Given that the main differences between the media involve floor transitions with one and three intervening intervals, we focus on these to further elucidate the underlying effects of task and medium. Jointly, these cases account for 68.77% of all transitions in the data.

4.2 Transitions with one intervening interval

Unlike the face-to-face and Zoom conversations, which exhibit a similar distribution of intervening intervals per speaker transition, the two tasks show notable differences with respect to transitions consisting of one and three intervening intervals. In the Diapix task, the most common sequence overall was BST with one intervening interval, while WSTs containing one intervening interval were clearly the most frequent sequence in free conversation. In sum, transitions containing only one intervening interval constitute about half of all transition types in free conversations (ftf: 48.50%, zoom: 51.27%), with a slightly lower proportion in Diapix (ftf: 37.01%, zoom: 41.00%).

In order to further elucidate these differences, Figure 7 shows the distribution of all BSTs and WSTs with one intervening interval. In both face-to-face and Zoom interactions, Diapix tasks have a higher proportion of *A_X_B* sequence (between-speaker silences) than *A_X_A* sequence (within-speaker silences); conversely, free conversation shows the opposite pattern. Conversation medium does not seem to affect the frequency of one-interval transitions.

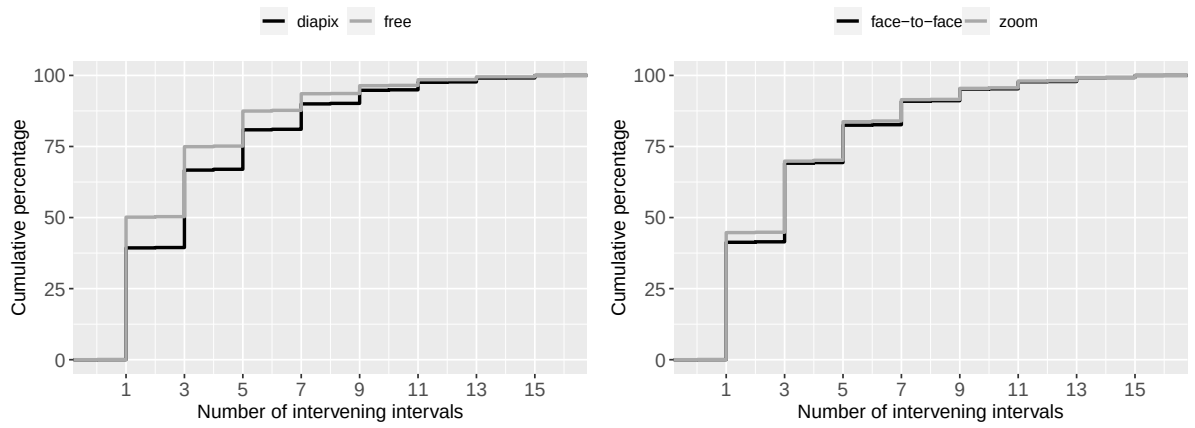


Figure 5: Cumulative distribution of the number of intervening intervals in a speaker transition depending on task (left) and medium (right).

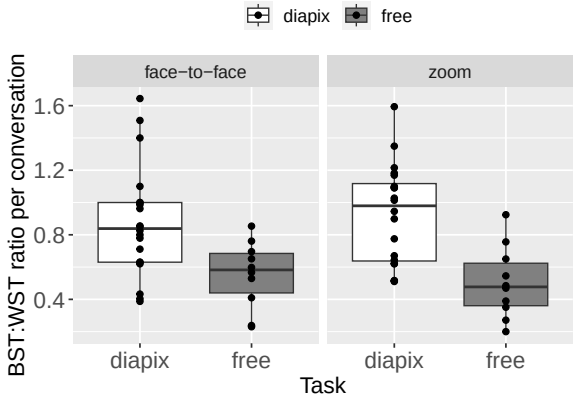


Figure 6: Distribution of BST:WST ratio per conversation.

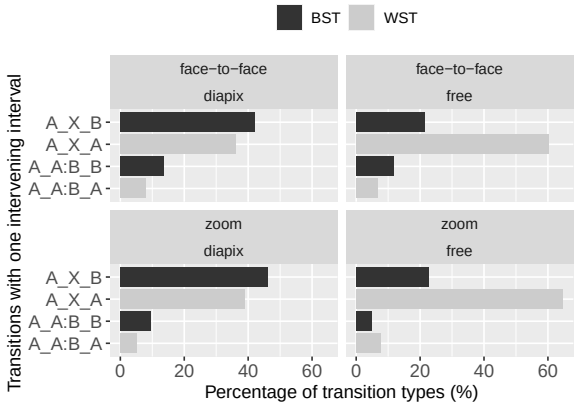


Figure 7: Distribution of floor state sequences in transitions with one intervening interval in face-to-face and Zoom interactions.

4.3 Transitions with three intervening intervals

Overall, transitions containing three intervals account for approximately 25% of all transitions across tasks and media (Diapix_ftf: 28.47%, free_ftf: 25.27%, Diapix_zoom: 25.48%, free_zoom: 23.60%), a slightly higher proportion in Diapix tasks in both media.

Compared to transitions with one intervening interval, there are usually fewer transitions with three intervening intervals across the tasks and media (see Figure 4). Only in the Diapix face-to-face interactions are WSTs containing three intervening intervals more frequent than WSTs containing one intervening interval. However, this difference is not present in Zoom interactions.

In Figure 8, we examine the BSTs and WSTs containing three intervening intervals in more detail. The most frequent transition types are similar for each task, with smaller differences between the media: the most common three-interval sequence used in Diapix conversations is the WST *A_X_B_X_A*, followed by its BST counterpart *A_X_B_X_B* (for an example, see Excerpt 1); while free conversations have a stronger preference for *A_X_A_X_A* sequence, followed by *A_X_B_X_A* in face-to-face interactions and *A_X_A_A:B_A* in Zoom (see Excerpts 2 and 3).

Excerpt 1: Sequences of *A_X_B_X_A* (line 1-3) and *A_X_B_X_B* (line 3-4).

```
1 A: ach so ja aber da sind drei runtergefallen
2 B: nein
3 A: und es hat ein ROTes Rad die Schubkarre
4 B: ja (0.4) und dahinter sind so zwei Stöcker

  A: oh yes, but three of them fell down
  B: no
```

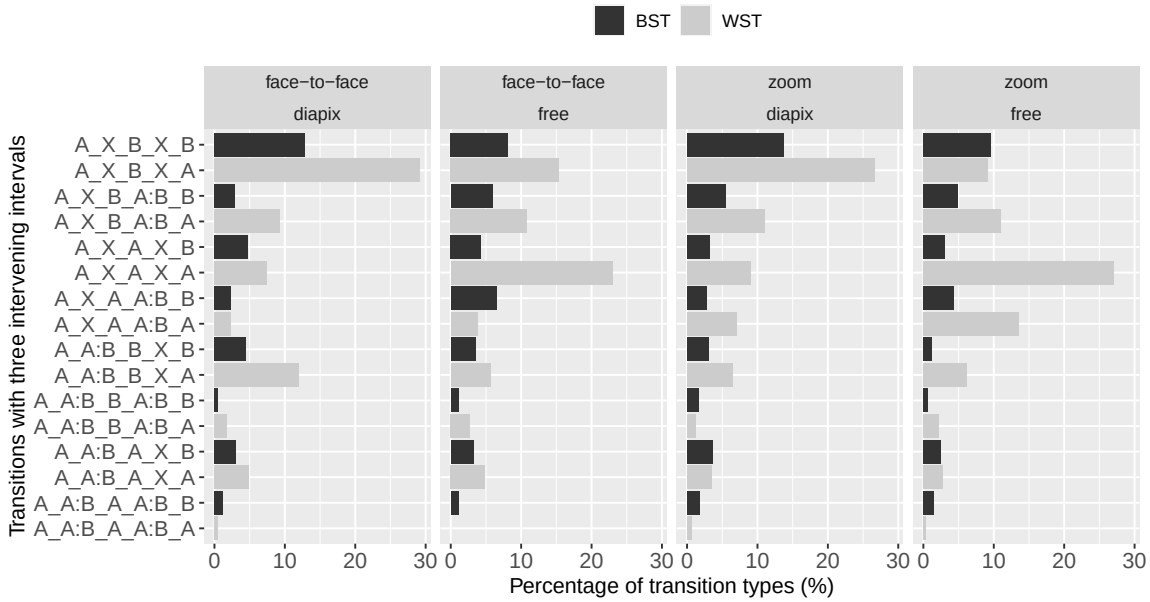


Figure 8: Distribution of floor state sequences in transitions with three intervening intervals in face-to-face and Zoom interactions.

```
A: and it has a RED wheel the wheelbarrow
B: yeah (0.4) and behind it are two sticks
                        [ba_z_diapix2_f2f1:315-323]
```

Excerpt 2: Sequence of *A_X_A_X_A*.

```
1 A: also ich find keine Ahnung mein mein
2    Lieblingsort in Berlin sind so (1.2) Orte zu
3    denen man sehr oft eigentlich hingeht (1.4)
4    Mercedes Benz Arena is für MICH voll schön

  A: well I don't know, my favorite places in
     Berlin are (1.2) actually places that you
     frequently visit (1.4) for ME Mercedes Benz
     Arena is quite nice
                        [bd_z_frei_m8f7:337-350]
```

Excerpt 3: Sequence of *A_X_A_A:B_A*.

```
1 A: diese Frage ganz anders beantworten (0.2)
2    auf seine eigene [Art und Wei]se
3 B:                  [ja=        ]

  A: answer this question quite differently (0.2)
     in his [own way]
  B:        [yeah   ]
                        [bd_z_frei_m8f7:502-506]
```

Upon closer examination of WSTs containing three intervening intervals, as shown in Figure 9, we can see that these sequences appear to fall into two distinct groups, depending on whether the interlocutor B is involved during the transitions. Transition including the involvement of the other interlocutor is preferred in all groups. Yet, transitions without B's involvement constitute a higher percentage in free conversations, regardless of medium. The implication of these results will be discussed in the next section.

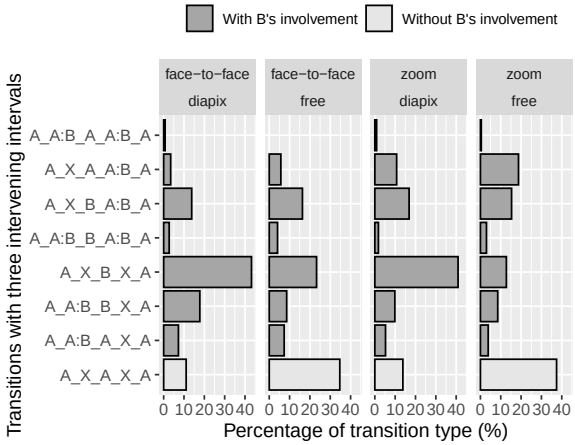


Figure 9: Distribution of WSTs with three intervening intervals categorised by B's involvement.

5 Discussion

The results of the present study show a striking similarity to previous studies (Gilmartin et al., 2020; Gilmartin, 2021), especially the transition pattern in free conversations. First, the distribution frequency of BSTs and WSTs declines sharply when the number of intervening intervals increases. Second, there are always more WSTs than BSTs within the group with the same number of intervening intervals.

However, Diapix task exhibits a distinct feature in both BSTs and WSTs containing one and three intervening intervals. Among the one-interval tran-

sitions, the BST occurrences outnumbered those of WST. To be more specific, there are more *A_X_B* sequences compared to the *A_X_A* sequences. This difference indicates that Diapix conversations are indeed characterised by high interactivity, with interlocutors changing the floor more frequently to facilitate intensive information exchange. Rapid turn-taking leads to a high number of between-speaker gaps, while monologic utterances are marked by numerous within-speaker pauses.

In the case of WSTs containing three intervening intervals, their majority consists of transitions where B produced a short utterance during A's monologic stretch, see Figure 9. The results suggest that Diapix tasks prompt speakers to provide more short utterances (e.g. backchannel, acknowledgement) than free conversations. These findings highlight that task-based conversations exhibit different turn-taking dynamics compared to free conversations. Our results thus reflect the distinctive characteristics of conversations with different levels of interactivity.

Compared to the task, the medium seems to play a less important role in speaker transition patterns. We expected that online-mediated conversations would reduce interlocutors' engagement, resulting in fewer speaker changes. A noteworthy difference is observed in the three-interval WSTs in Diapix tasks, where their occurrences surpassed those of one-interval WSTs in face-to-face interactions, but not over Zoom. We assume that interlocutors provide more feedback in the back channel when conversing face-to-face, while on Zoom, due to the latency and remoteness, backchannel-like utterances are avoided to prevent misinterpretation as a turn-starter, which could cause unintended interruption.

Beyond this, the media did not alter the general speaker transition patterns within the same task. This may be attributed to interlocutors' increased familiarity with remote conversations (see Section 3.1), leading them to adapt to the new conversation dynamics.

Another potential explanation for the minor influence of remote conversation is that the transition delay of audio signal does not reach the threshold needed to create noticeable disruptions, such as the 800 ms delay in telephone conversation suggested by Egger-Lampl et al. (2010). Unfortunately, we were not able to obtain the exact delay in real-time conversation, as Zoom does not provide access to this data. Consequently, this forms a new area of focus for future work where the latency in remote conversations will be examined.

It is worth pointing out that data-driven analysis such as that described above cannot capture all the details of real conversations. Based on the task settings, we assume that the three-interval WSTs are primarily short feedback utterances, such as acknowledgement, short answers and backchannelling. Nonetheless, instances of unsuccessful floor competition and premature relinquishment would also be included in these sequences. A qualitative analysis of these cases is needed to determine the exact distribution of backchannelling and other potential turn-taking behaviours.

We plan to build on our analyses by exploring the role of the duration of the intervening intervals in transitions and indeed the stretches of solo speech bounding the transitions in order to deepen our understanding of how speech is arranged by participants, and also to extend our analyses to a variety of spoken interaction types. We hope that the insights gained by these studies will contribute to a better understanding of human-human spoken interaction and will aid in specifying more effective artificial dialogue technologies.

Acknowledgments

Emer Gilmartin's work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-II220043,Adaptive Personality for Intelligent Agents)

References

Jeremy N. Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior*, 2(1).

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

Beatrice Beebe, Diane Alson, Joseph Jaffe, Stanley Feldstein, and Cynthia Crown. 1988. Vocal congruence in mother-infant play. *Journal of Psycholinguistic Research*, 17:245–259.

Beatrice Beebe, Joseph Jaffe, Frank Lachmann, Stanley Feldstein, Cynthia Crown, and Michael Jasnow. 2000. Systems models in development and psychoanalysis: The case of vocal rhythm coordination and attachment. *Infant Mental Health Journal*, 21(1-2):99–122.

Malte Belz, Alina Zöllner, Megumi Terada, Robert Lange, Lea-Sophie Adam, and Bianca Sell. 2021. Dokumentation und Annotationsrichtlinien für das Korpus BeDiaCo.

Julie E. Boland, Pedro Fonseca, Ilana Mermelstein, and Myles Williamson. 2021. Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, pages 1272–1282.

Paul T. Brady. 1968. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91.

Oliveira Maggie Bullock and Bianca Sell. 2022. PDF and PSD files of DiapixGEtv picture materials – German version adapted to elicit tense vowels.

Carole Edelsky. 1981. Who's Got the Floor? *Language in Society*, 10(3):383–421.

Sebastian Egger-Lampl, Raimund Schatz, and Stefan Scherer. 2010. It takes two to tango - Assessing the impact of delay on conversational interactivity on perceived speech quality. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pages 1321–1324.

Emer Gilmartin. 2021. *Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis*. Ph.D. thesis, Trinity College, Dublin, Dublin, Ireland.

Emer Gilmartin, Kätlin Aare, Maria O'Reilly, and Marcin Wlodarczak. 2020. Between and within speaker transitions in multiparty conversation. In *Proceedings of Speech Prosody 2020*, pages 799–803, Tokyo, Japan.

Leilani H. Gilpin, Danielle M. Olson, and Tarfah Alrashed. 2018. Perception of speaker personality traits using speech signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page LBW514. ACM.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Joseph Jaffe, Louis Cassotta, and Stanley Feldstein. 1964. Markovian model of time patterns of speech. *Science*, 144(3620):884–886.

Joseph Jaffe and Stanley Feldstein. 1970. *Rhythms of dialogue*. Academic Press, New York.

Kornel Laskowski. 2011. *Predicting, detecting and explaining the occurrence of vocal activity in multiparty conversation*. Ph.D. thesis, Carnegie Mellon University.

A. C. Norwine and O. J. Murphy. 1938. Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, 17(2):281–291.

Daniel C. O'Connell, Sabine Kowal, and Erika Kaltenbacher. 1990. Turn-taking: A critical analysis of the research tradition. *Journal of psycholinguistic research*, 19(6):345–373.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Langauge*, 50(4):696–735.

Deborah Tannen. 1980. Toward a Theory of Conversational Style: The Machine-Gun Question. Technical report, Southwest Educational Development Laboratory, 211 East 7th Street, Austin, Texas 78701.

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.

Louis ten Bosch, Nelleke Oostdijk, and Jan Peter de Ruiter. 2004. Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Proceedings of 7th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic.

Kristin J. Van Engen, Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. 2010. The wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 53(4):510–540.

Marcin Włodarczak and Emer Gilmartin. 2021. Speaker transition patterns in three-party conversation: Evidence from English, Estonian and Swedish. In *Proceedings of Interspeech 2021*, pages 801–805.

Poster Abstracts

A Multi-party Dialogue Dataset for Dialogue Goal Tracking in a Hospital Setting and How It Can Be Used in LLM Prompt Engineering Experiments

Weronika Sieińska, Angus Addlessee, Daniel Hernández García, Nancie Gunson, Marta Romeo, Christian Dondrup, Oliver Lemon
Heriot-Watt University, Edinburgh, UK
{w.sieinska, a.addlesee, d.hernandez_garcia, n.gunson, m.romeo, c.dondrup, o.lemon}@hw.ac.uk

Abstract

We describe a multi-party dialogue dataset, which we collected, annotated, and released on GitHub for public use. The dataset is specifically designed for the task of dialogue goal tracking. It consists of transcriptions of 35 conversational interactions between 2 human speakers and a humanoid social robot called ARI in a hospital setting. The robot is there to alleviate the workload of medical staff by providing patients with information related to the hospital. In the dataset, each utterance that states a goal of a human speaker, e.g., to go to the reception or to find out where they can get a cup of coffee, is explicitly annotated with that goal. In this paper, we also describe a computational experiment we conducted with the use of the dataset to illustrate how it can be used. We prompt engineered 5 large language models for the task or dialogue goal tracking. While some of the models performed very poorly, others were able to grasp the task quite well and predicted most goal annotations correctly.

1 Introduction

Today's voice assistants are typically dyadic, with a single user interacting with a single system. However, as dialogue systems are getting deployed on social robots and placed in public spaces (Gunson et al., 2022; Moujahid et al., 2022), these systems are increasingly required to deal with challenges of multi-party dialogue. Importantly in this paper, they need to track user goals, which can be shared between multiple people or even answered not by the system but by other human speakers.

Regardless of the number of users, in order for a conversational system to work, it needs to contain a control mechanism for tracking the state of a dialogue, which is a separate, however similar, challenge. Researchers have been interested in tracking the state of a dialogue for years (Larsson and Traum, 2000; Williams and Young, 2007; Wang and Lemon, 2013; Ren et al., 2018; Balaraman et al., 2021). In 2013, Williams et al. (2013) started a series of scientific competitions called Dialogue State Tracking Challenge[1] (DSTC). In 2024, the dialogue research community can participate in the competition for the 12th time[2].

Dialogue goal tracking, on the other hand, is a form of dialogue system evaluation, especially in task-oriented (also called *goal*-oriented) dialogues, which creates the need for robust goal tracking strategies and suited datasets.

Researchers have been collecting multi-party dialogue data for years, some of which is even multimodal (Robinson et al., 2004; Djalali et al., 2012; Yamasaki et al., 2012; Mahajan and Shaikh, 2021; Reverdy et al., 2022). The existing variety of datasets also serve various purposes. Some datasets were constructed for the task of building common ground between different parties (Furuya et al., 2022), whereas others – for modeling social phenomena in discourse (Shaikh et al., 2010). Chen et al. (2020) built a multi-party dialogue dataset for the analysis of emotions and interpersonal relationships between speakers. To our knowledge, however, there are no available datasets built specifically for the task of speaker's goal tracking in human-robot interaction.

2 Multi-party Dialogue Dataset

In this paper, we describe a novel multi-party dialogue dataset consisting of transcriptions of 35 interactions between 2 human speakers and a humanoid robot called ARI (Cooper et al., 2020) in a hospital setting. The robot is there to alleviate the workload of medical staff by providing patients and their companions with information related to the hospital. We de-

[1]The competition is now known as Dialogue System Technology Challenge.

[2]https://dstc12.dstc.community/

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

signed our dataset specifically for the task of multi-party dialogue goal tracking and released it as a GitHub repository: https://github.com/wsieinska/multi-party-dialogue-dataset.

We annotated the data for speakers, addressees, and goals of speakers such as to get a cup of coffee, to find lifts, to go to the toilet, etc. We differentiate between *individual* goals – when only 1 speaker has the goal; and *shared* goals – when both speakers have the same goal (e.g., they both want to eat something). We think that, in multi-party interactions, the distinction between individual and shared goals may affect the way they are answered, and, hopefully, make the interactions feel more natural.

We used ELAN[3] for annotation, which is a tool for annotating audio and video recordings (Brugman and Russel, 2004). We describe in detail how the data was collected in Appendix A and how it was annotated in Appendix B. Dataset statistics can be found in Appendix C. Appendix D contains an example dialogue from our dataset.

3 Computational Experiment

We conducted a computational experiment with the use of our dataset. We prompt engineered 5 large language models (LLMs) to perform goal tracking in multi-party conversations, namely: GPT-4o, GPT-4 Turbo, GPT-3.5 Turbo, Vicuna-13b-v1.5-16k, and Llama-2-13b-chat-hf-16k. The prompt we used can be found in Appendix E.

We took the few-shot learning approach and added 3 training examples to the prompt (dialogues 1, 11, and 21) leaving 32 dialogues for testing (3 was the highest possible number due to memory limitations). For each test dialogue file, we created a copy and replaced goal annotations with blanks. The task for the LLMs was to return these dialogues with blanks filled in with their predictions of goal annotations. It can be divided into two subtasks: (1) return the same text of the given test dialogue, (2) replace blanks with predictions of goal annotations.

We evaluated performance at subtask 1 by computing similarity scores between generated dialogues and dialogues from our dataset. We used `python3 difflib.SequenceMatcher` as our metric. Then, to evaluate performance at subtask 2, we extracted predicted goal annotations and compared them to gold annotations from our dataset with the use of the same metric. However, due to the fact that the LLMs did not perform very well at (it would seem straightforward) subtask 1 (especially Llama-2-13b-chat-hf-16k), some generated dialogues needed to be slightly altered to enable automatic extraction of predicted goal annotations. Both altered and unaltered dialogues are available for comparison in our GitHub repository.

Table 1 presents our experimental results. Each result is a mean of results obtained for all 32 dialogues used for testing. GPT-4o obtained the best results at both subtasks reaching 84% at subtask 1 and almost 80% at subtask 2. Llama-2-13b-chat-hf-16k performed the worst and did not even reach 5% of goal annotations predicted correctly.

Model	Subtask 1	Subtask 2
Llama-2-13b-chat-hf-16k	31.04 ± 17.49	4.89 ± 11.85
Vicuna-13b-v1.5-16k	61.02 ± 21.22	36.99 ± 38.03
GPT-3.5 Turbo	73.37 ± 19.92	63.54 ± 34.37
GPT-4 Turbo	77.71 ± 23.32	66.09 ± 39.22
GPT-4o	**84.09 ± 20.25**	**79.33 ± 30.89**

Table 1: Experimental results for subtasks 1 and 2.

4 Conclusions and Future Work

Multi-party dialogue goal tracking is a complex and challenging task. In order to solve it, multi-party dialogue data must be collected and annotated for speakers' goals. Therefore, we hope that our dataset will be a valuable contribution.

In our experiment, we tested the ability of 5 state-of-the-art LLMs to track goals of speakers in multi-party interactions. Some of the models were able to grasp the task quite well, however, there is still a lot of room for improvement.

In the future, it would be interesting to repeat our experiment with other prompts, e.g., a more detailed prompt explaining the reasoning behind how goals are annotated, and more training examples.

In this work, we were solely interested in the task of tracking goals of speakers. However, our dataset could be annotated for split utterances, coreferences, anaphoras, ellipses, and clarification requests; and used for other tasks.

Lastly, we appreciate that the size of our dataset is rather small. Hence, another useful follow-up to our work would be further data collection.

Acknowledgements

This research was funded by the EU H2020 program under grant agreement no. 871245 (SPRING project https://spring-h2020.eu/).

[3] https://archive.mpi.nl/tla/elan

References

Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent Neural Methods on Dialogue State Tracking for Task-Oriented Dialogue Systems: A Survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, Singapore and Online. Association for Computational Linguistics.

Hennie Brugman and Albert Russel. 2004. Annotating Multi-media/Multi-modal Resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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, Marseille, France. European Language Resources Association.

Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. ARI: the Social Assistive Robot and Companion. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 745–751, virtual. IEEE.

Alex Djalali, Sven Lauer, and Christopher Potts. 2012. Corpus Evidence for Preference-Driven Interpretation. In *Logic, Language and Meaning*, pages 150–159, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yuki Furuya, Koki Saito, Kosuke Ogura, Koh Mitsuda, Ryuichiro Higashinaka, and Kazunori Takashio. 2022. Dialogue Corpus Construction Considering Modality and Social Relationships in Building Common Ground. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4088–4095, Marseille, France. European Language Resources Association.

Nancie Gunson, Daniel Hernández García, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2022. Developing a Social Conversational Robot for the Hospital Waiting Room. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1352–1357.

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3–4):323–340.

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.

Meriam Moujahid, Helen Hastie, and Oliver Lemon. 2022. Multi-party Interaction with a Robot Receptionist. In *17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 927–931. IEEE.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards Universal Dialogue State Tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, Brussels, Belgium. Association for Computational Linguistics.

Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R. Cowan, and Naomi Harte. 2022. RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2517–2527, Marseille, France. European Language Resources Association.

Susan Robinson, Bilyana Martinovski, Saurabh Garg, Jens Stephan, and David Traum. 2004. Issues in Corpus Development for Multi-party Multi-modal Task-oriented Dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Samira Shaikh, Tomek Strzalkowski, Aaron Broadwell, Jennifer Stromer-Galley, Sarah Taylor, and Nick Webb. 2010. MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Zhuoran Wang and Oliver Lemon. 2013. A Simple and Generic Belief Tracking Mechanism for the Dialog State Tracking Challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432, Metz, France. Association for Computational Linguistics.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Jason D. Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

Shota Yamasaki, Hirohisa Furukawa, Masafumi Nishida, Kristiina Jokinen, and Seiichi Yamamoto. 2012. Multimodal Corpus of Multi-party Conversations in Second Language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 416–421, Istanbul, Turkey. European Language Resources Association (ELRA).

A Data Collection

We collected a dataset of 35 interactions between 2 human speakers and a humanoid robot called ARI in a hospital setting. We did that in the "Wizard of Oz" setup. Each interaction was recorded using cameras both on ARI itself and external ones.

In the videos, one can see two human speakers (the participants of the data collection) standing next to each other. Please note that the released dataset only contains written transcriptions of the interactions. We were not able to release videos due to privacy regulations (videos contained personally identifiable information of data collection participants – their faces).

Human speakers were assigned particular roles. One of them was a patient who came to the hospital to attend an appointment with a medical doctor, whereas the other was their companion.

Participants were also given tasks to complete in each interaction with ARI. They were supposed to: retrieve information about the location of lifts, room 17, and toilets; as well as find out where they can get something to eat, where they can get a cup of coffee, and what time they should expect their appointment to commence at. In the dataset, goal annotations often reflect the tasks participants were trying to complete. Figure 1 presents picture representations of the tasks given to the participants. The pictures allowed us to avoid suggesting the use of any particular words and fostered more diverse wording in the dataset.

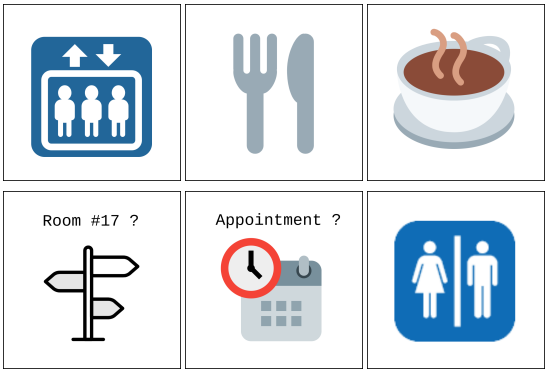


Figure 1: Picture representations of tasks given to participants during data collection.

The tasks were supposed to give participants an idea about what kind of information they can retrieve from ARI, however, participants were welcome to make other hospital-related requests, e.g., ARI was asked whether the hospital cafeteria serves cakes and whether consultations are covered by social security health insurance.

B Data Annotation

We annotated the data for speakers, addressees, and goals of speakers. All of the data was annotated by the first author of this paper, and 20% of the data was also annotated by the second author. Overall, the authors agreed with each other's annotations in 96.08%.

B.1 Speaker Annotation

Speaker is the participant who uttered the given utterance. It is either patient (`Pat`), companion (`Com`), or ARI (`ARI`). Speaker annotations were determined by the analysis of videos, in particular: head and body movements, and voice timbres. Unfortunately, it was not possible to determine who is speaking at the given moment by looking at participants' lips as they were covered by face masks. The inter-annotator agreement for speaker annotations is 100.00%.

B.2 Addressee Annotation

Addressee is the participant who the given utterance is addressed to. Similarly to speaker annotation, addressee annotation required the analysis of videos, head and body movements in particular. Sometimes, the speaker would address someone by their name, making the addressee annotation task trivial, e.g., "So, Mrs Companion, do you know what I'll be eating today?", "ARI, I've been waiting a long time, I'm tired.". Possible values of the addressee annotation are the following: ARI (`ARI`), patient (`Pat`), companion (`Com`) – one addressee; patient and companion (`Pat+Com`) – ARI addressing both human speakers; patient and ARI (`Pat+ARI`), companion and ARI (`Com+ARI`) – a human speaker addressing ARI and the other human speaker. The inter-annotator agreement for addressee annotations is 98.53%.

B.3 Goal Annotation

In each interaction, the patient and the companion have certain goals, which often reflect the tasks the participants were given during data collection (to get a cup of coffee, to find lifts, to go to the toilet, etc.). The inter-annotator agreement for goal annotations is 89.71%.

If a patient (`Pat`) has a goal to go to the hospital reception, the syntax of the goal annotation is the following: `G(Pat, go-to(reception))`. All

goal annotations from this dataset are listed below (each of the annotations can represent a goal of any human speaker Pat/Com):

- G(Pat, drink(<ARG>)) – the patient is thirsty and they specified that they would like to drink <ARG>, where <ARG> is, e.g., coffee, hot chocolate, tea, water, etc.;
- G(Pat, eat(<ARG>)) – the patient is hungry and they specified that they would like to eat <ARG>, where <ARG> is, e.g., a piece of cake, croissant, sandwich, etc.;
- G(Pat, get-info(<ARG>)) – the patient would like to get information about <ARG>, where <ARG> is, e.g., their appointment, day schedule in the hospital, cafeteria opening times, etc.;
- G(Pat, go-to(<ARG>)) – the patient would like to go to <ARG>, were <ARG> is, e.g., the cafeteria, courtyard, lift, reception, toilet, vending machine, etc.;
- G(Pat, sit-down()) – the patient is tired and would like to sit down.

If an argument is missing in the G(Pat, drink()) or the G(Pat, eat()) goal annotations, it means that the patient is thirsty/hungry but did not specify what they would like to drink/eat. In the dataset, the argument is always present for the G(Pat, get-info(<ARG>)) and the G(Pat, go-to(<ARG>)) goal annotations. G(Pat, sit-down()) does not take an argument.

Other goal annotations, which are rare but also occur in the dataset, are: G(Pat, request-escort(<ARG>)) – here <ARG> is a location and is always specified, G(Pat, request-volume-up()), and G(Pat, get-help()) which do not take an argument.

B.4 Types of Goal Annotations

There are 5 types of goal annotations (each of the annotations can represent a goal of any human speaker Pat/Com):

- G(Pat, get-info(cafeteria(location))) – "open goal" – used when the patient asks for the location of the hospital's cafeteria by saying, e.g., "Where can I find the cafeteria?".
- AGP(Pat, get-info(cafeteria(location))) – "answer goal (positive)" – used when ARI or the companion gives the patient the information they requested by saying, e.g., "There's a cafeteria on the ground floor, near the courtyard.".
- AGN(Pat, get-info(cafeteria(location))) – "answer goal (negative)" – used when ARI or the companion expresses their inability to provide requested information by saying, e.g., "Sorry, I don't have this information.".
- CGP(Pat, get-info(cafeteria(location))) – "close goal (positive)" – used when the patient acknowledges they have received the requested information by saying, e.g., "Ok, great, thanks.".
- CGN(Pat, get-info(cafeteria(location))) – "close goal (negative)" – used when the patient acknowledges they will not receive the information they requested by saying, e.g., "Oh well, thanks anyway.".

Each utterance that states a goal is explicitly annotated with that goal – even if that particular goal has already occurred before and is still open. There is no need for more types of goal annotations: RG(Pat, go-to(reception)) – "reopen goal" – is not necessary because it can be treated just like opening a new goal (it does not matter that the same goal has already occurred in the dialogue and that it is closed). We decided to take this approach for simplicity.

B.5 Shared Goal Annotation

All goal annotations described so far are examples of *individual* goal annotations – they describe goals of individual participants (the patient or the companion). Some goal annotations, however, describe goals, which are *shared* by the participants. We think that, in multi-party interactions, the distinction between individual and shared goals may affect the way they are answered, and, hopefully, make the interactions feel more natural, e.g., if a shared goal was opened, it could be more natural for ARI to address both participants while answering it, not just the one who was the speaker and opened it. Addressing both participants instead of just one of them could be reflected in the wording of the answer, ARI's head pose, ARI's gestures, etc.

Shared goals are built similarly to individual ones. Participants sharing a goal are joined by the "+" sign: Pat+Com (the order does not matter,

however, it is always Pat+Com (not Com+Pat) in the dataset (for simplicity)), and if their goal is to eat a sandwich, the annotation is the following: G(Pat+Com, eat(sandwich)).

A goal counts as shared when the speaker uses the word "we", e.g.:

- Pat: "How does it work here? We don't have any information. Is there any schedule for the day?" →G(Pat+Com, get-info(day_schedule))
- Pat: "And how do we get to the cafeteria?" →G(Pat+Com, get-info(cafeteria(directions)))
- Pat: "Could we have a little hot chocolate?" →G(Pat+Com, drink(hot_chocolate))

A goal also counts as shared when the speaker says "Me too." (or the like) following the specification of the other participant's goal, e.g.:

- Com: "I'm thirsty. I would like a glass of water." →G(Com, drink(water))
  Pat: "Oh yes, me too. Do you think there's a water fountain?" →G(Pat+Com, drink(water))
- Com: "Could you wait for me here? I need to go to the toilet." →G(Com, go-to(toilet))
  Pat: "I need to go too. I'll go with you." →G(Pat+Com, go-to(toilet))
- Pat: "I would grab a bite, I'm getting hungry." →G(Pat, eat())
  Com: "So am I. ARI, where can we get something to eat?" →G(Pat+Com, eat())

C Dataset Statistics

We analysed the data in terms of the number of turns, number of tokens (words), and the number of goal annotations. Table 2 presents statistics describing our dataset. On average, a dialogue from our dataset consists of 29.8 turns and includes 271.71 tokens and 8.17 individual G-type goal annotations.

D Example Dialogue

Table 3 presents an example dialogue from our dataset. In the dialogue, the patient and the companion want to eat a snack, go to the cafeteria, go to the toilet, and find out their appointment time. Their goals are opened, answered, and closed.

	Mean	St.Dev.	Min	Max
Turns	29.80	15.20	12	67
Tokens	271.71	162.19	86	766
Ind. G	8.17	6.89	1	30
Ind. AGP	3.74	3.32	0	11
Ind. AGN	1.71	2.30	0	9
Ind. CGP	1.69	1.43	0	4
Ind. CGN	0.60	0.81	0	2
Sh. G	2.63	2.28	0	8
Sh. AGP	1.51	1.63	0	6
Sh. AGN	0.60	0.91	0	3
Sh. CGP	0.74	0.92	0	3
Sh. CGN	0.23	0.60	0	3

Table 2: Dataset statistics (Ind. – Individual, Sh. – Shared, St.Dev. – Standard Deviation).

E The Prompt

I will give you a dialogue between two people, whose names are Pat and Com, and a robot, whose name is ARI. The dialogue will consist of multiple dialogue turns in the following format: "turn speaker->addressee: *utterance* @goal$". If the dialogue is "01 Pat->ARI: *I would like a cup of coffee, please.* @G(Pat, drink(coffee))$", then '01' is the turn number, 'Pat' is the speaker, 'ARI' is the addressee,"*I would like a cup of coffee, please.*" is the utterance, and "@G(Pat, drink(coffee))$" is an annotation of the goal of the speaker. However, each goal annotation will be replaced with the '@[BLANK]$' tag. I want you to guess missing goal annotations and return the dialogue with blanks filled in. You will find this dialogue between the '<START>' and '<END>' tags. Do not return any other text. I will also give you three example dialogues to learn from. Do not return the text of example dialogues or any other text. Remember, your task is to return the text between the '<START>' and '<END>' tags with the '@[BLANK]$' tags replaced by your guesses of goal annotations.
Example dialogue 1:
{example_dialogue_1}
Example dialogue 2:
{example_dialogue_2}
Example dialogue 3:
{example_dialogue_3}
Here is the dialogue, which I want you to return with blanks filled in:
<START>
{dialogue_with_blanks}
<END>

T.	Sp.→Add.	Utterance	Goal
01	ARI→Pat+Com	Hello, how can I help you?	–
02	Com→ARI	Hello.	–
03	Pat→Com	Well, it's my first time here, I don't know if this is the case for you. I'd really like to um... to be able, to be able to eat a little bit before going to my appointment, do you know where that is?	G(Pat, get-info(food(location)))
04	Com→Pat	Oh, I don't know. We'll ask the question.	AGN(Pat, get-info(food(location)))
05	Com→ARI	I don't know what your name is. Can you give us information and tell us what your name is?	G(Com, get-info(ari(name)))
06	ARI→Com	Hello, my name is ARI. How can I help you?	AGP(Com, get-info(ari(name)))
07	Com→ARI	Where's the cafeteria?	G(Com, get-info(cafeteria(location)))
08	ARI→Com	There are a few options available as part of your visit, and there's also a cafeteria on the ground floor.	AGP(Com, get-info(cafeteria(location)))
09	Com→Pat	Ok, on the ground floor.	CGP(Com, get-info(cafeteria(location)))
10	Pat→ARI	And how do we get to the cafeteria?	G(Pat+Com, get-info(cafeteria(directions)))
11	ARI→Pat+Com	You have to enter the second building behind you. Then it's the second door on the left.	AGP(Pat+Com, get-info(cafeteria(directions)))
12	Pat→ARI	Right, the second building.	CGP(Pat+Com, get-info(cafeteria(directions)))
13	Com→ARI	I'd really like to know what's on the menu.	G(Com, get-info(menu))
14	ARI→Com	Today we have an endive salad to start, followed by a ham and cheese escalope with a side of mashed potato and a fruit compote for dessert.	AGP(Com, get-info(menu))
15	Pat→ARI	Yes, but that's for lunch. We would like a snack, wouldn't we? Grab a coffee or breakfast, since we're early for our appointment.	G(Pat+Com, eat(snack))
16	ARI→Pat+Com	Of course. There are a few options available as part of your visit, and there's also a cafeteria on the ground floor.	AGP(Pat+Com, eat(snack))
17	Com→Pat	And I'd really like to go to the toilet too.	G(Com, go-to(toilet))
18	Com→ARI	Can you tell us where the toilets are?	G(Com, get-info(toilet(location)))
19	ARI→Com	Very easy! You have to go to the corridor at the end of the room. It's the second door on the right.	AGP(Com, get-info(toilet(location)))
20	Com→ARI	Second door on the right. Oh well, we'll find it. Thank you.	CGP(Com, get-info(toilet(location)))
21	Pat→ARI	Next to... So it's next to the cafeteria?	G(Pat, get-info(toilet(location)))
22	ARI→Pat	Absolutely.	AGP(Pat, get-info(toilet(location)))
23	Com→Pat	I feel like it's further away. That, that's not on the ground floor...	–
24	Pat→ARI	I'm having a difficulty understanding what you're saying. I can't hear very well. Could you repeat where the toilets are, please?	G(Pat, get-info(toilet(location)))
25	ARI→Pat	Very easy! You have to go to the corridor at the end of the room. It's the second door on the right.	AGP(Pat, get-info(toilet(location)))
26	Pat→ARI	In the corridor at the end of the room, ok...	CGP(Pat, get-info(toilet(location)))
27	Pat→ARI	Um, how do I know what time my appointment is at?	G(Pat, get-info(appointment(time)))
28	ARI→Pat	Sorry, I don't have this information.	AGN(Pat, get-info(appointment(time)))
29	Pat→ARI	That's a shame. Um...	CGN(Pat, get-info(appointment(time)))
30	Com→Pat	Maybe it's at the reception. Maybe there's a service, a receptionist who will let you know, who will tell you. Let's go there.	G(Com, go-to(reception))
31	Pat→Com	Ok, I'll go and ask at the reception then.	G(Pat, go-to(reception))
32	ARI→Pat+Com	Thank you. See you around. Have a nice day.	–
33	Com→ARI	Thank you.	–
34	Pat→ARI	Have a nice day.	–

Table 3: Example dialogue from our dataset where individual and shared goals are opened, answered, and closed (T. – Turn, Sp. – Speaker, Add. – Addressee).

Using LLMs to Generate Training Data for Dialogue System NLUs

Bogdan Laszlo
University of Gothenburg
Master in Language Technology (MLT)
guslasbo@student.gu.se

Staffan Larsson **Asad Sayeed**
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
staffan.larsson@ling.gu.se
asad.sayeed@gu.se

Abstract

This paper explores using Large Language Models (LLMs) to generate dialogue datasets for training lightweight Natural Language Understanding (NLU) models for use in modular task-oriented dialogue systems. Employing a schema-guided framework and prompt engineering, we explore how synthetic dialogue data compares to MultiWoZ data on NLU tasks.

1 Introduction

LLMs are impressive in their capability to participate in open-domain dialogue, including understanding user utterances. At the same time there are problems with LLMs, such as producing misleading or false output ("hallucinations"), failure to adhere to instructions, sensitivity to small nuances in prompt design, costs and environmental impact (Rillig et al., 2023), and reliance on constant calls to proprietary LLMs in the cloud.

For many practical, domain-specific applications, a more lightweight controllable modular dialogue system may still be a viable alternative. However, it may often be desirable also in modular systems to make use of the advantages of LLMs. Using LLMs to generate training data for lightweight NLU models is one example of this. NLU models are designed to e.g. determine user intent, identify key entities and/or decipher sentiment.

Collecting datasets of human-human dialogue is labour-intensive, expensive, and may involve privacy concerns. Wizard-of-Oz (WoZ) data collection (Budzianowski et al., 2018) also requires manual effort for data cleaning and annotation. (Budzianowski, 2019).

Synthetic data generation offers a potentially viable and affordable solution for NLU training. However, synthetic datasets in general may exhibit biases in data distribution, may contain incomplete data and inconsistent annotations, and lack may diversity and nuance (Hao et al., 2024; Li et al., 2023).

In this paper, we investigate how NLU models trained on synthetic data compare to models trained on real-world data, when both are tested against real-world data.

2 Method and dataset

To generate synthetic dialogues, we use a schema-guided framework inspired by (Li et al., 2023) combined with strategic prompt engineering (Rastogi et al., 2020). The schema-guided approach involves defining a structured framework that outlines the possible states and transitions in a dialogue, ensuring that the generated dialogues are viable and aligned with specific conversational objectives.

3 Using LLMs to generate dialogues

As explored in Steindl et al. (2023) and Park et al. (2023), LLMs can produce dialogues that closely mimic human conversations. LLM dialogue generation can be fine-tuned for specific applications, such as asking relevant and context-specific questions (Horiuchi and Higashinaka, 2022), replicating complex dialogue patterns across various domains, Liu et al. (2023) and answer retrieval for a retrieval-based conversational character (Chen and Artstein, 2024).

4 Data and Models

Previous approaches to generating synthetic dialogue data have but encountered significant issues. These include models deviating from given templates (Steindl et al., 2023), generating contextually irrelevant responses (Liu et al., 2023), and facing scalability challenges (Rastogi et al., 2020).

The method proposed here tries to address these problems by enforcing strict dialogue schemas through prompt engineering, ensuring models adhere to templates. Additionally, the dialogue-

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

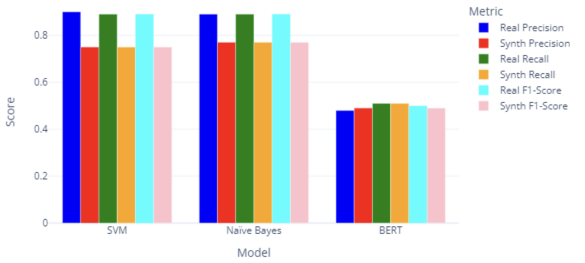


Figure 1: Experiment I — Domain Classification

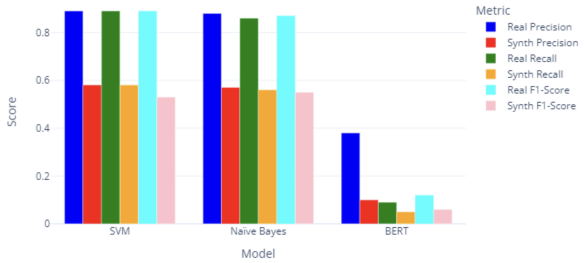


Figure 2: Experiment II — Intent Classification

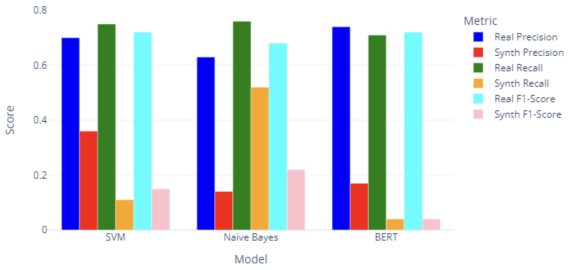


Figure 3: Experiment III — Slot labeling

generating model is exposed to the entire dialogue history in each iteration to prevent out-of-context utterances, in order to improve the coherence and relevance of synthetic dialogue.

Several dialogue datasets have been used for training NLU models. A prominent dataset is MultiWOZ, a multi-domain wizard-of-oz dataset (Budzianowski et al., 2018) that includes several annotations useful for training NLUs. We constructed a synthetic dataset that mirrors the structure and selected domains of MultiWoZ 2.2. We used 458 *train* and 500 *hotel* domain dialogues. The synthetic dataset was similarly constrained to approximately 516 *train* and 500 *hotel* domain dialogues. The creation of the synthetic dataset involved the following steps[1]: schema generation (using GPT-3.5), dialogue generation from schemas (using GPT-4), dialogue clean-up (to remove inconsistencies and errors introduced in generation), alignment of annotation alignment with MultiWoZ structure, and splitting the dataset into training, validation, and testing subsets with proportions of 80%, 5%, and 15%, respectively.

5 Experiment and results

We trained three different NLU models: Support Vector Machine(SVM), Naive Bayes, and BERT on both our synthetic dataset and MultiWoZ. Each model is evaluated on 3 tasks: domain classification, mulitclass intent classification, and slot multi-labelling.

On the domain classification task (Figure 1), the models trained on MultiWoZ perform better than those trained on the synthetic dataset, with the exception of the BERT model which performs poorly overall. However, models trained on the synthetic dataset perform better than a random baseline model. On the intent classification task (Figure 2), the MultiWoZ-trained models in general outperform the synthetic-trained ones.

In slot labeling (Figure 3), models detect the presence of slots without extracting their values. The synthetically-trained models have an unsatisfactory performance. Some slots were not labelled at all. SVM was not always able to beat the baseline model (which assigned the *book_train* category to every utterance). The Naïve Bayes model predicted the same label for every utterance.

6 Error analysis

Error analysis reveals some limitations of the method used, including inconsistent quality, where generated dialogues often lacked the complexity that characterizes natural dialogue. Another limitation is bias, causing repetitiveness with respect to phrasing and chosen topics and converging on a limited number of scenarios focusing primarily on Eurocentric settings (e.g. constantly referencing cities such as London and Cambridge). Furthermore, annotation quality is a concern, and the LLMs introduce hallucinated slots and intents that do not conform to the established MultiWOZ annotation guidelines.

7 Conclusion and future work

We conclude that synthetic data is useful for NLU training, and more so for low-granularity tasks, but not as useful as human data. This is in line with e.g. Chen and Artstein (2024).

[1]The source code for the dialogue generation framework is available at https://github.com/Devix71/nlu_dialogue_dataset_generator

Acknowledgments

This work was supported by Swedish Research Council grant 2014-39, Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Paweł Budzianowski. 2019. The magic triangle of dialogue data collection. PolyAI Blog. [Online; accessed 2024-01-15].

Elizabeth Chen and Ron Artstein. 2024. Augmenting training data for a virtual character using gpt-3.5. In *The International FLAIRS Conference Proceedings*, volume 37.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*.

Sota Horiuchi and Ryuichiro Higashinaka. 2022. Learning to ask specific questions naturally in chat-oriented dialogue systems. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 263–276. Springer.

Bogdan Laszlo. 2024. Creating synthetic dialogue datasets for nlu training. an approach using large language models. Master's thesis, Master in Language Technology Programme, University of Gothenburg.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Mengjuan Liu, Chenyang Liu, Yunfan Yang, Jiang Liu, and Mohan Jing. 2023. Promoting open-domain dialogue generation through learning pattern information between contexts and responses. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 351–362. Springer.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

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.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Sebastian Steindl, Ulrich Schäfer, and Bernd Ludwig. 2023. Generating synthetic dialogues from prompts to improve task-oriented dialogue systems. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 207–214. Springer.

Boosting Questions' Effectiveness in Medical Interviews

Davide Mazzaccara
CIMeC, University of Trento
davide.mazzaccara@unitn.it

Alberto Testoni
ILLC, University of Amsterdam
a.testoni@uva.nl

Raffaella Bernardi
CIMeC, DISI, University of Trento
raffaella.bernardi@unitn.it

Abstract

Questions are a fundamental tool for acquiring information, from children's learning to complex tasks. Recent work has shown that the informativeness of questions by large language models (LLMs) can be enhanced through Direct Preference Optimization (DPO) and Expected Information Gain (EIG). In this study, we evaluate the effectiveness of a DPO-trained model in the context of medical interviews. Our findings indicate that DPO training improves success rates in medical interviews, thereby demonstrating the broader applicability and generalizability of this approach.

1 Introduction

Questions in language serve as requests for information (Hiż, 1978). The speaker lacks information in their knowledge state and asks questions to gain this information. This process of acquiring information through questioning is essential for children to learn about the world (Ruggeri and Lombrozo, 2015) and for adults to solve complex problems (Geva et al., 2021). A complex problem is a medical interview: the doctor asks questions to elicit the patient's signs and symptoms. Once enough information has been collected, the doctor identifies the disease and proceeds with treatment.

Despite their remarkable language and reasoning abilities (Kojima et al., 2022), Large Language Models (LLMs) have been observed to generate low informative questions (Bertolazzi et al., 2023), evaluated through the 20 Questions Game and Expected Information Gain (EIG). Based on the intuition that LLMs are good at generating diverse questions and providing answers to these close-ended questions (Testoni et al., 2023), Hu et al. (2024) propose an inference time probabilistic reasoning strategy (see also Piriyakulkij et al. 2023). The authors make the LLM generate different questions via sampling, then selecting the question maximizing the EIG measure. Alternatively, Mazzaccara et al. (2024) use probabilistic reasoning to create a dataset of sampled low and high-informative questions. By training on these data with preference optimization, the authors conclude that LLMs could learn to reason with informativeness.

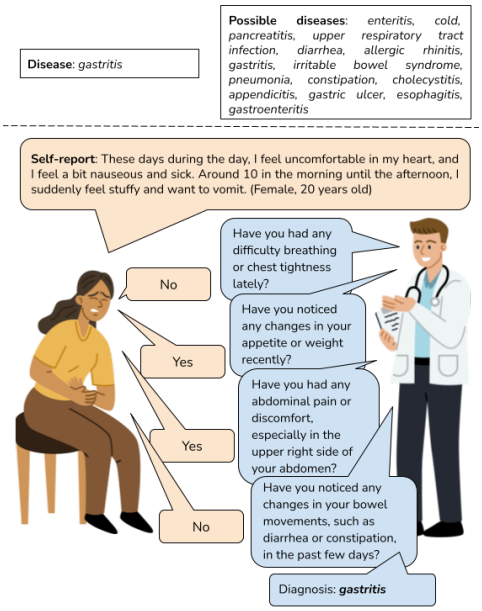


Figure 1: Example of a Medical Interview (MedDG). The dialogue is machine-generated: LLAMA 2 DPO plays the role of the doctor and GPT-3.5 the patient.

Mazzaccara et al. (2024) concludes that LLM's reasoning with informativeness generalises across different domains. Our study delves into this conclusion by testing the trained model on a different domain and task, i.e., medical interviews. Medical interviews are task-oriented dialogues, where the doctor collects information through question-answer pairs to make a diagnosis. As illustrated in Fig. 1, the doctor is provided with the possible diseases and a patient's self-report. The doctor asks questions about the patient's signs and symptoms to identify the patient's disease. LLMs, trained to ask informative questions, could assist doctors towards more efficient and effective medical interviews.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

2 Setting

The 20 Questions Game and Medical Interviews comprise two roles, a Questioner and an Answerer. The Questioner ask yes/no questions to collect information and identify the candidate in a list of possible candidates. The Answerer guides this process, providing truthful yes/no answers. In our setting, a *game* consists of the candidate set with the target candidate; a *dialogue* is the series of question-answers exchanges. A dialogue is considered *successful* if the target is reached within the first 20 questions.

To train an LLM to ask informative questions, Mazzaccara et al. (2024) creates 20 Question games with common concepts from the following categories: mammal, bird, clothing, weapon, fruit, and vegetables. Questions are sampled from the chosen LLM, LLAMA 2-CHAT (7B), and then evaluated in terms of EIG by the same model. The resulting low and high-EIG questions are employed to tune the same LLAMA 2 with Direct Preference Optimization (Rafailov et al., 2023). Trained to ask more informative questions, the resulting model is more efficient (fewer turns to reach the target) and more effective (higher success rate) in the 20 Questions game in different domains. We compare LLAMA 2-CHAT (7B) Zero-shot and after DPO as Questioner, the Answerer is GPT-3.5-TURBO-0125.

In the task of Medical Interview, the Questioner asks yes/no questions to identify the patient's disease.[1] Medical interviews differ from the 20 Questions game in that the Questioner is initially provided with a self-report from the patient. This implies that a medical interview game comprises: self-report, possible candidate diseases, and the target disease. The self-report is provided to the Questioner before the first turn alongside the possible diseases. In our evaluation setting, we test the trained model with and without the self-report in two medical datasets.

The medical datasets employed for testing are DX (Xu et al., 2019) and MedDG (Liu et al., 2022). The English versions of the datasets are provided by Hu et al. (2024). The self-reports of both DX and MedDG have been extracted from human online doctor-patient interviews. We employed the test set of DX, consisting of 104 games with 5 possible diseases. For MedDG, we use the 10% of the selected games by Hu et al. (2024). The resulting MedDG dataset consisting of 50 games with 15 possible diseases.

[1] Simplifying our setting to yes/no questions and answers allows for easier computation of EIG, while representing a good approximation of the task

3 Results

We evaluate the training's impact on efficiency and effectiveness in medical interviews. The Average number of Questions (**AQ**) measures efficiency as the number of questions the model needs to reach the target. The Success rate at 1 (**S@1**) measures effectiveness as the percentage of times the model achieves the target within the first tentative.

The results are reported in Table 1. DPO training seems not to positively impact informativeness, as shown by lower AQ in all settings. In terms of effectiveness, instead, the DPO training leads to higher S@1 for both DX and MedDG. In DX medical interviews, the DPO outperforms the Zero-shot by an absolute difference of $+12.5\%$ in S@1 without the self-report and $+28.8\%$ S@1 with the self-report. Overall this is a rather positive result given that DPO has been trained on radically different concept domains. When comparing the same setting with and without the self-report, we see that for large candidate sets, MedDG, both Zero-Shot and DPO improve their Success rate, as one would expect; interestingly, DPO improves its efficiency more than Zero-shot (the AQ decreases -3 vs. -0.6). Maybe surprisingly, with smaller candidate sets, DX, both DPO and Zero-shot improve in efficiency, when the self-report is provided, but their success rate decreases with DPO suffering less (-5.8 vs. -22.1).

Setting	Method	DX		MedDG	
		AQ ↓	S@1 ↑	AQ ↓	S@1 ↑
w/o self-report	Zero-shot	5.5	42.3%	7.6	6.0%
	DPO	6.3	54.8%	9.9	12.0%
self-report	Zero-shot	4.4	20.2%	7.0	18.0%
	DPO	4.5	49.0%	6.9	22.0%

Table 1: Results for LLAMA 2-CHAT (7B) zero-shot and DPO in DX and MedDG. In the first row are reported the results for the setting without self-report. In the second row, with the self-report.

References

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. ChatGPT's information seeking strategy: Insights from the 20-

questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162, Prague, Czechia. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9.

Henry Hiż, editor. 1978. *Questions*. Reidel, Dordrecht/Boston.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Proceedings*, pages 447–459, Germany. Springer Science and Business Media Deutschland GmbH.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. 2024. Learning to ask informative questions: Enhancing llms with preference optimization and expected information gain. *Preprint*, arXiv:2406.17453.

Top Piriyakulkij, Volodymyr Kuleshov, and Kevin Ellis. 2023. Asking clarifying questions using language models and probabilistic reasoning. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Azzurra Ruggeri and Tania Lombrozo. 2015. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216.

Alberto Testoni, Raffaella Bernardi, and Azzurra Ruggeri. 2023. The efficiency of question-asking strategies in a real-world visual search task. *Cognitive Science*, 47(12).

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7346–7353.

Inferring Partner Models for Adaptive Explanation Generation

Amelie Robrecht
Social Cognitive Systems
Bielefeld University
arobrecht@techfak.de

Heike Buhl
Educational Psychology
Paderborn University
heike.buhl@uni-paderborn.de

Stefan Kopp
Social Cognitive Systems
Bielefeld University
skopp@techfak.de

1 Introduction

While most current approaches focus on explanations as single-turn answers to why-questions (Chandra et al., 2024; Lewis, 1986; Anjomshoae et al., 2019), we conceive them as a co-constructive process that may encompass different explanatory questions, including *What?, How?*, and *Why?* (Rohlfing et al., 2021; Axelsson et al., 2022; El-Assady et al., 2019; Lombrozo, 2006; Miller, 2019). Crucially, in human-human explanations, we see interlocutors adapting their utterances to what they think supports their partner best. Although it has long been argued that this requires a model of the partner (Clark and Wilkes-Gibbs, 1986), it is unclear, which features this partner model (PM) consists of. We argue that such a model is a dynamic and fuzzy representation of the interlocutor needed to maintain a shared understanding or grounding of a task. It is influenced by characteristics, experiences, expectations, and stereotypes (Brennan et al., 2010) and must encompass the modeling of a partner's knowledge. That is, a PM is not a box with facts, but rather a mosaic of relevant known fragments about the partner (Dillenbourg et al., 2016). This stresses the importance of inferring the correct fragments for a given task, initially and then constantly updated during the interaction.

Previous work on conversational human-agent interaction has mainly looked at adapting the interaction to the user's evolving understanding (Buschmeier and Kopp, 2018). We extend this view by considering two processes, *inferring* the assumed features in a PM from the user's behavior (Chandra et al., 2024), and then *applying* the PM in a non-stationary decision process to determine the best communicative action. Our previous work focused on applying the simplyfied PM in the decision process (Robrecht and Kopp, 2023). Now we focus on how it is inferred, both in the sense of determining relevant features and then inferring their values during the interaction. That is, we focus on the relations of features and observable information in the PM (analogous to Dillenbourg et al. (2016)). To that end, we go beyond the user's knowledge by examining four additional features hypothesized to be decisive in explanations. We employ a Dynamic Bayesian Network (DBN) to model this inference in order to keep track of the central goal of an explanation: grounding the explanandum.

2 Features of the Partner Model

If the considered features of the user go beyond pure knowledge, the explanation becomes more personal and efficient. The more explicit the state of a dimensional feature is inferred, the more informative they are as an instrument for user adaptation. The belief about a feature is dynamic, independent of the feature's invariance to time. Each feature can be tracked when receiving explicit feature-directed statements or implicitly in the course of interaction. The more meaningful explicit determination is rare, due to its higher costs, while the more fuzzy implicit determination can be executed continuously.

We hypothesize user's **expertise** E to play a significant role in tailoring the explanation to them, as it influences the depth of information required for understanding. Unlike local knowledge, expertise is considered as prior knowledge which is persistent and does not fluctuate during the interaction. Expertise can be observed through explicit user statements S_e or implicit through the frequency of positive user feedback FB_p. A high level of expertise increases the improvement of understanding, as the user already has domain-specific knowledge and can transfer structures and relationships. When adapting to this feature, the agent therefore expects a user with a higher level of expertise to understand more quickly.

Cognitive load L describes the amount of a person's limited working memory resources used in

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

a specific task (Chandler and Sweller, 1991). We assume that adapting an explanation to the personal cognitive load is relevant in order not to leave the listener hanging or bored. Making an adaptive system sensitive to the user's cognitive load is an established approach (Khawaja et al., 2014) and linguistic measures are established in HAI (Khawaja et al., 2014; Arvan et al., 2023). Cognitive load can explicitly be derived from user statements S_L, which are considered the most reliable (Khawaja et al., 2014). Specific linguistic features, such as word count (higher load = longer sentences) (Khawaja et al., 2014), Type-Token Ratio (higher load = lower ratio) (Arvan et al., 2023), or Gunning Fog Index (Gunning, 1968; Khawaja et al., 2014) (higher load = higher index), are proven to correlate with the user's cognitive load and can be used as an implicit measure FB_c. When adapting to the cognitive load of the user, the amount of information per utterance can be varied.

We expect attentiveness to be a relevant feature, as we presume a user with high attentiveness to have a low probability of missing a given information. Just like the cognitive load, the users' **attentiveness** A can change while the explanation evolves. Although there is no explicit measure for attentiveness, there are different ways of implicit measurement: In addition to eye movement and prosody, the frequency of feedback (FB_p and FB_n) is a predictor of attentiveness (Buschmeier et al., 2011; Oertel et al., 2016). Consequently, a low level of attentiveness leads to a higher probability of fully missing an utterance when applying this feature.

According to Allwood et al. (1992), feedback can be illustrated as a ladder with four rungs: contact, perception, understanding, and attitudinal reactions. While attentiveness mainly deals with the lower levels of feedback (contact and perception), **cooperativeness** (C) represents the user's willingness to express understanding and attitude. Consequently, cooperativeness mainly considers so-called *substantive contributions* (Chi et al., 2008), where the user takes the turn. We assume cooperativeness to be an important feature as we expect a highly cooperative user to autonomously interrupt and report non-understanding. The dynamic feature of cooperativeness can be indicated by explicit utterances S_c ,or measured implicitly through the frequency of substantive feedback FB_s. When applying this feature, a higher level of cooperativeness leads to a higher improvement of understanding if no feedback is provided.

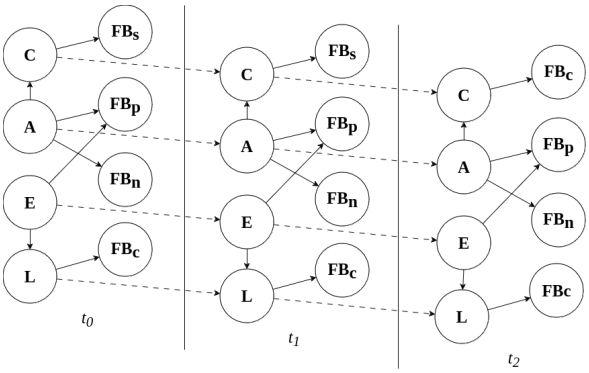


Figure 1: DBN to infer PM. Only implicit Feedback is displayed. See text for variable introduction.

Chandra et al. (2024) argue that a PM is inferred and repaired. Similarly, we look at inferring the PM using a DBN (see Fig. 1). A Bayesian Network is a graphical formalism for representing joint probability distributions, while DBNs are specifically designed to model changes over time, assuming a stationary underlying process with the previous state as a prior (Murphy, 2002). A time step always consists of an explanation move and the associated user feedback. Even if no response is given, the frequency of feedback (FB_n, FB_p, FB_s) changes and the DBN needs to be updated. Certain features are assumed to influence others: As shown in (Khawaja et al., 2014) and discussed earlier, expertise impacts the cognitive load of the user. At the same time, attentiveness is a requirement for cooperativeness (Allwood et al., 1992), which explains their dependency.

3 Discussion and Future Work

This paper explores the relevant features needed for a PM to effectively adapt an explanation. It focuses on features that go beyond pure knowledge. In a second step, it introduces a DBN as a potential tool for implementing such a PM in human-machine interaction. In a next step, the PM will be evaluated by eliminating individual features or combinations of these. The explanations created in this process will be compared with each other and with human-generated explanations, to confirm that the selected features have the hypothesized influence. Additionally, we will merge the improved PM with the current decision model (Robrecht and Kopp, 2023) and assess it in a user study.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Mohammad Arvan, Mina Valizadeh, Parian Haghighat, Toan Nguyen, Heejin Jeong, and Natalie Parde. 2023. Linguistic Cognitive Load Analysis on Dialogues with an Intelligent Virtual Assistant. In *Proceedings of the 45th Annual Conference of the Cognitive ScienceSociety*.

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling Feedback in Interaction With Conversational Agents—A Review. *Frontiers in Computer Science*, 4:744574.

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two Minds, One Dialog. In *Psychology of Learning and Motivation*, volume 53, pages 301–344. Elsevier.

Hendrik Buschmeier and Stefan Kopp. 2018. Communicative Listener Feedback in Human–Agent Interaction: Artificial Speakers Need to Be Attentive and Adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden.

Hendrik Buschmeier, Zofia Malisz, Marcin Włodarczak, Stefan Kopp, and Petra Wagner. 2011. 'are you sure you're paying attention?' - 'uh-huh' communicating understanding as a marker of attentiveness. In *Interspeech 2011*, pages 2057–2060. ISCA.

Paul Chandler and John Sweller. 1991. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4):293–332.

Kartik Chandra, Tony Chen, Tzu-Mao Li, Jonathan Ragan-Kelley, and Joshua Tenenbaum. 2024. Cooperative Explanation as Rational Communication. *arXiv preprint*.

Michelene T. H. Chi, Marguerite Roy, and Robert G. M. Hausmann. 2008. Observing Tutorial Dialogues Collaboratively: Insights About Human Tutoring Effectiveness From Vicarious Learning. *Cognitive Science*, 32(2):301–341.

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

Pierre Dillenbourg, Séverin Lemaignan, Mirweis Sangin, Nicolas Nova, and Gaëlle Molinari. 2016. The symmetry of partner modelling. *International Journal of Computer-Supported Collaborative Learning*, 11(2):227–253.

Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. 2019. Towards XAI: Structuring the Processes of Explanations. *Proceedings of the ACM Workshop on Human-Centered Machine Learning, Glasgow, UK*, 4:13.

Robert Gunning. 1968. *The Technique of Clear Writing*. New York, McGraw-Hill.

M. Asif Khawaja, Fang Chen, and Nadine Marcus. 2014. Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *International Journal of Human-Computer Interaction*, 30(5):343–368.

David Lewis. 1986. Causal Explanation. *Philosphical Papers*, 2:214–240.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38. ArXiv: 1706.07269v3.

Kevin P Murphy. 2002. Dynamic Bayesian Networks.

Catharine Oertel, Joakim Gustafson, and Alan W. Black. 2016. Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Feedback Utterances. In *Interspeech 2016*, pages 2915–2919. ISCA.

Amelie Robrecht and Stefan Kopp. 2023. SNAPE: A Sequential Non-Stationary Decision Process Model for Adaptive Explanation Generation:. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 48–58, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Hab-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728. Publisher: Institute of Electrical and Electronics Engineers Inc.

Modeling the Use-Mention Distinction in LLM-Generated Grounding Acts

Milena Belosevic
German Linguistics, Faculty of Linguistics and Literary Studies,
Bielefeld University
milena.belosevic@uni-bielefeld.de

Hendrik Buschmeier
Digital Linguistics Lab, Faculty of Linguistics and Literary Studies,
Bielefeld University
hbuschme@uni-bielefeld.de

Abstract

Given that large language models (LLMs) are systems that do not understand human language in a human-like way, LLM-generated grounding acts, such as explicit claims of understanding (e.g., "I understand"), can lead to overtrust in the capabilities of LLM chatbots, supporting their perception as human interlocutors (Shaikh et al., 2024). This paper argues for enriching these grounding acts with metalinguistic markers (e.g., scare quotes) that motivate users to perceive them as 'mentioned' and not as 'used' language (use–mention distinction; Sperber and Wilson, 1981). We illustrate how different types of meta-language can be enriched with (non)verbal metalinguistic units to mark LLM-generated grounding acts as mentioned language.

1 Introduction

Shared understanding is crucial for effective dialogues in human interactions and, arguably, interactions with artificial interlocutors. Therefore, a growing body of research deals with the role of common ground in interactions with LLMs (Jokinen et al., 2024; Mohapatra, 2023; Shaikh et al., 2024; Pilán et al., 2024). Defining common ground in the context of LLMs is challenging because it is still unclear what (if anything) LLMs understand and whether they have human-like understanding capabilities (Bender et al., 2021). At first sight, LLM-based chatbots can generate human-like grounding acts (e.g., acknowledgments) and exhibit attentiveness and adaptiveness to their interlocutor's feedback and needs (Buschmeier and Kopp, 2018). However, LLM-generated grounding acts often mislead users into ascribing human-like capabilities to them. This contrasts with theories claiming that LLMs are systems without communicative intents that merely produce statistically likely continuations of word sequences (Shanahan, 2024). The system, thus, produces grounding acts because LLMs perform well on formal linguistic competence (Mahowald et al., 2024). This paper assumes that "LLMs do not exhibit the kind of understanding that requires commonsense knowledge, but simply make inferences based on statistically significant syntactic patterns" (Saba, 2023). Therefore, the system cannot understand a question in a human-like manner, eventually producing grounding acts that should not be perceived verbatim. The lack of LLM's functional linguistic competence may lead to overreliance and unsafe use of LLMs (Bender et al., 2021). For this reason, the concept of common ground needs to be modified.

2 (Non)verbal Metalinguistic Indicators of Use–Mention Distinction

This short paper proposes modifying common ground in interactions with LLMs based on the user's metalinguistic knowledge. Our approach reconciles the incapability of LLMs to understand language in a human-like manner on the one hand and their ability to produce linguistic patterns formally identical to those used by human interlocutors in naturalistic contexts on the other hand. It also aims to shift users' perception of LLM-generated grounding acts as human-like signals of conversational grounding toward the assumption that these grounding acts signal a gap between the meanings that humans project onto the LLM-generated texts and what the texts in fact mean (Hayles, 2023). To avoid users' overreliance on the system and support them in modifying their expectations regarding the LLMs' understanding capabilities, the concept of common ground (Clark and Schaefer, 1989) should be adjusted to LLMs' capabilities. To this end, metalinguistic (non)verbal markers could help users perceive LLM-generated grounding acts as 'mentioned' and not as 'used' language (i.e., employing a linguistic expression to talk about the expression itself rather than to talk about some aspect of the

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

world; see Moore, 2019, pp. 12–13 and Sperber and Wilson, 1981).

The distinction between used and mentioned language is based on the human ability to take a linguistic item as an object of scrutiny (Anderson et al., 2002; Wilson, 2011). In human interactions, one of the main functions of metacommunicative markers, such as metalinguistic commentaries (e.g., "What I was trying to say was . . . "), or quotations (Jaworski et al., 2004), is to indicate the use–mention distinction. In addition, metalinguistic skills are central for monitoring one's own and making inferences about other's state of understanding (Anderson et al., 2002). Therefore, LLM-generated output that comprises anthropomorphic linguistic units (Abercrombie et al., 2023) should be explicitly marked as the mentioned language. Accordingly, these units should be perceived as the mentioned language.

3 Modelling LLM-Generated Grounding Acts as Mentioned Language

We propose to modify a corpus-based classification schema of meta-language in naturally occurring human conversations (Anderson et al., 2004) to the context of human-LLM interactions. To model the communicative incapabilities of LLMs, this schema could be specified by (non)verbal metalinguistic oral and written markers proposed by Hyland (2018, pp. 33–34). These markers are appropriate because, in conversations with chatbots, the message is transmitted by written communication and conceptualized as a spoken language (Koch and Oesterreicher, 1985). We hypothesize that three of the five types of metalanguage proposed by Anderson et al. (2004) could be relevant to human-LLM interaction and can be modified for this context. These are illustrated with an example in Table 1, and it can be seen that each type can be specified by several (non)verbal metalinguistic units to mark LLM-generated grounding acts as mentioned language.

The metalinguistic units could be produced by explicitly instructing (via prompts) the system to generate them or by including a second agent in the human–LLM interaction. This agent could initiate meta-dialogues (Traum and Andersen, 1999) or serve as a 'reflection assistant' (Kim et al., 2023) motivating users to prompt the generation of metalinguistic markers. (Non)verbal metalinguistic units are more or less explicit and can be combined with each other across all three types of meta-language.

Types of meta-language	**Examples of (non)verbal metalinguistic units**
Simulate clarification or correct the word meanings produced by users: User: Can you solve this math problem? Chatbot: You mean *generate a solution*? / What does the word "solve" mean?	Intonation, stress, voice quality; font style, weight, and type; quotes; mention-significant nouns and verbs (*mean, say, word, term,* etc.)
Simulate monitoring one's own ongoing utterance: User: Can you solve this math problem?; Chatbot: Yes, I can "help" you./I can help you (I "said": help).	quotes and air quotes; instances of meta-dialogue
Simulate commenting on users' or own words: User: Can you solve this math problem?; Chatbot: "Can you solve [!] this math problem?" / Yes, I can solve [sic] it.	mention-significant nouns and verbs (*mean, say, word, term,* etc.); exclamation marks; quote-similar expressions ([sic])

Table 1: Potential markers of LLM-generated grounding acts as mentioned language.

The cases presented in Table 1 are thus not exhaustive. For example, to correct the anthropomorphic user's input, the chatbot could be instructed to combine a font style with the mention-significant verb (Wilson, 2011, 43–50) "mean", which is less implicit than explicitly asking about the meaning of the verb "solve". Similarly, simulating monitoring of one's own language use with emojis is more implicit than the instances of meta-dialogue: "I can help you (I "said": help)." Finally, the chatbot can repeat (some parts) of the user's input to comment on it and implicitly motivate users to critically reflect on their language use.

4 Conclusions and Outlook

This paper illustrates how metalinguistic markers could guide users to adopt a metalinguistic critical stance towards LLM-generated grounding acts. Their practical application should be tested in naturally occurring human-LLM interactions. Given that grounding acts in human interactions can be described as metadiscursive (since they are used to check and manage understanding (Kopple, 1985; Verdonik, 2022; Verdonik et al., 2023), we will test experimentally whether they can be perceived as metalinguistic markers without being marked with metalinguistic units discussed above.

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore.

Michael L. Anderson, Andrew Fister, Bryant Lee, Luwito Tardia, and Danny Wang. 2004. On the types and frequency of meta-language in conversation: A preliminary report. In *14th Annual Meeting of the Society for Text and Discourse*, pages 1–4, Chicago, IL, USA.

Michael L. Anderson, Yoshi Okamoto, Darsana Josyula, and Don Perlis. 2002. The use-mention distinction and its importance to HCI. In *Proceedings of the 6th Workshop on the Semantics and Pragmatics of Dialog*, pages 21–28, Edinburgh, UK.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual, Canada.

Hendrik Buschmeier and Stefan Kopp. 2018. Communicative listener feedback in human–agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1213–1221, Stockholm, Sweden.

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

Katherine N. Hayles. 2023. Afterword: Learning to read AI texts. Critical Inquiry Blog.

Ken Hyland. 2018. *Metadiscourse*. Bloomsbury Academic, London, UK.

Adam Jaworski, Nikolas Coupland, and Dariusz Galasinski. 2004. Metalanguage: why now? *Language Power and Social Process*, 11:3–10.

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards harnessing large language models for comprehension of conversational grounding. In *Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*, Sapporo, Japan.

Yeongdae Kim, Takane Ueno, Katie Seaborn, Hiroki Oura, Jacqueline Urakami, and Yuto Sawa. 2023. Exoskeleton for the mind: Exploring strategies against misinformation with a metacognitive agent. In *Proceedings of the Augmented Humans International Conference 2023*, Glasgow, UK.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1):15–43.

William J. Vande Kopple. 1985. Some exploratory discourse on metadiscourse. *College Composition and Communication,*, 36:82–93.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28:517–540.

Biswesh Mohapatra. 2023. Conversational grounding in multimodal dialog systems. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 706–710, Paris, France.

Andrew W. Moore. 2019. How significant is the use/mention distinction? In Andrew W. Moore, editor, *Language, World, and Limits: Essays in the Philosophy of Language and Metaphysics*, page 11–16. Oxford University Press, Oxford, UK.

Ildikó Pilán, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2024. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In *Proceedings of the 25th Meeting of the Special Interest Group on Discourse and Dialogue*, Kyoto, Japan.

Walid S. Saba. 2023. Stochastic LLMs do not understand language: Towards symbolic, explainable and ontologically based LLMs. In *International Conference on Conceptual Modeling*, pages 3–19. Springer.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 6279–6296, Mexico City, Mexico.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use–mention distinction. In Peter Cole, editor, *Radical Pragmatics*, pages 295–318. Academic Press, New York, NY, USA.

David R Traum and Carl F Andersen. 1999. Representations of dialogue state for domain and task independent meta-dialogue.

Darinka Verdonik. 2022. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28:144–171.

Darinka Verdonik, Simona Majhenič, and Andreja Bizjak. 2023. Are metadiscourse dialogue acts a category on their own? In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue – Poster Abstracts*, Maribor, Slovenia.

Shomir Wilson. 2011. *A Computational Theory of the Use-Mention Distinction in Natural Language*. Ph.D. thesis, University of Maryland, College Park, MD, USA.

MedExpDial: Machine-to-Machine Generation of Explanatory Dialogues for Medical QA

Andrea Zaninello
Fondazione Bruno Kessler
Free University of Bolzano (Italy)
azaninello@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

1 Motivations and Background

We describe a pilot study on generating synthetic explanatory dialogues for the medical domain, based on a pre-existing medical dataset of multiple-choice questions with human-written explanations. We use an instruction-tuned large language model (LLM) to generate dialogues between a medical student and a teacher/doctor helping answer questions about clinical cases. We inject varying degrees of background knowledge into the teacher prompt and analyze the effectiveness of these dialogues in terms of whether the student is able to get to the correct answer and in how many turns. This method has potential applications in developing and evaluating argument-based explanation models for medical question answering (QA).

Currently, medical QA systems and health-related AI systems are increasingly being used to provide patients with access to reliable information, support healthcare professionals in their decision-making processes, or for educational purposes (Kell et al., 2024; Alonso et al., 2024; Yagnik et al., 2024; García-Ferrero et al., 2024). A key challenge in this field is providing explanations that are both accurate and understandable to the user (Li'evin et al., 2022), as they play a crucial role in building trust and transparency in AI systems, particularly in critical domains like healthcare (Hossain et al., 2023).

On the one hand, traditional approaches to explanation generation in medical QA often involve providing static summaries, rule-based or template-based explanations (Budler et al., 2023). However, these approaches are only partially able to capture the reasoning involved in medical diagnosis and treatment (Li'evin et al., 2022; Molinet et al., 2024). On the other hand, by engaging the users in a conversation, dialogue systems can provide more interactive explanations, adapting to the user's specific needs and understanding, which can

```
<?xml version="1.0" ?>
<full_question id="23_113" type="INFECTIOUS">
  A 71-year-old woman with a history of rheumatoid arthritis on sulfasalazine,
prednisone and etanercept. She goes to the emergency room for 72 hours of clinical
manifestations compatible with facial herpes zoster affecting the right hemiface,
auricular pavilion, respecting the forehead and conjunctival chemosis. What would be
the appropriate treatment?
  <full_answer>I think this question is not clearly in the Infectious Diseases
syllabus, and may overlap with OFT and DERMA, but as I understand it, in an
immunocompromised patient and also with data of ocular involvement, admission for
intravenous treatment would be indicated due to the high risk of possible
complications.</full_answer>
  <option num="1">Symptomatic treatment of pain only.</option>
  <option num="2">Topical treatment with acyclovir.</option>
  <option num="3">Outpatient treatment with acyclovir, valacyclovir or oral
famciclovir.</option>
  <option num="4" correct="YES">
    Hospital admission and treatment with acyclovir or famciclovir iv.
    <explanation char_ranges="[[135, 310]]" word_ranges="[[24, 50]]">in an
immunocompromised patient and also with data of ocular involvement, admission for
intravenous treatment would be indicated due to the high risk of possible
complications.</explanation>
  </option>
  <option num="5">Parenteral Ig and vaccination.</option>
</full_question>
```

Figure 1: An xml-coded question, answers and explanations from the CasiMedicos dataset.

be dynamically tailored through interactions and feedback in a dialogue flow (Wachsmuth and Alshomary, 2022). However, because of the highly sensitive nature of medical records, ecological data are extremely difficult to collect in this domain.

To fill this gap, we explore the generation of dialogue-based medical explanations in an educational setting (Anonymous, 2024), as a way to enhance the explainability of medical QA systems, contributing to developing effective medical dialogue models.

2 Explanatory Dialogue Generation

Our explanatory dialogues are based on *CasiMedicos*, a pre-existing dataset of medical questions and answers with human-written explanations (Agerri et al., 2023), which contains questions in Spanish, English, French, Basque, and Italian, covering various medical specialties. Every language corresponds to a train, test, dev splits of 434, 125, and 63 questions each. Each question consists of a clinical case followed by a question on the case, 5 multiple-choice options of which one is the correct answer, and a human-written explanation for the correct answer and/or for the reason why the other

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

options are not correct. An example question from *CasiMedicos* is provided in Figure 1.

The first step is to identify the questions in CasiMedicos that a state-of-art LLM is *unable* to correctly answer, under the assumption that its internal knowledge alone is not sufficient to answer them. To do this, we prompt an instance of GPT-4 (OpenAI, 2023) to answer the 125 questions of the English split of the CasiMedicos test set, without any help (0-shot). We parse the model's answers with regular expressions and compare them with the CasiMedicos correct answers. GPT-4 was able to answer 105 over 125 questions correctly, yelding an initial accuracy of 84%.

Then, we use the 20 answers that the model was unable to answer correctly and two independent instances of GPT-4, a medical *Teacher* and a medical *Student*, to generate dialogues. The Teacher is prompted to help a student prepare for the USMLE exam, and incrementally provided with more information from the knowledge base, while the Student is only prompted to play the role of the student with no additional information[1].

We experiment with four different modes of dialogue generation corresponding to the information provided to the Teacher instance. Specifically, the Teacher is only provided with the clinical case without the correct answer (More 0), or incrementally with the correct answer (Mode 1), the alternative options (Mode 2), and the human-written explanation (Mode 3).

The Teacher is allowed to use any of the provided information as she wishes to guide the conversation and help the Student reach the correct answer. The Teacher is also prompted to end the conversation when the final answer is reached, outputting an <END> tag once the Student identifies the correct answer. For each question, 2 different dialogues are generated for each mode, ranging from a minimum of 6 turns to a maximum of 10 turns, for a total of 160 dialogues. We split the generated dialogues into an 80-dialogue test and dev sets.

Finally, students from the University of Bologna manually annotated each dialogue of the test set for the following elements: 1. *Answer Detection*, i.e., the text fragment within the dialogue where the Student provides her final answer; 2. *Option Mapping*: a mapping between the Student's final answer and the original question's option[2]; 3. *Answer Correctness*: whether the Student's answer is correct based on the knowledge base. We manually and semi-automatically revise the annotation and evaluate the effectiveness of the dialogues in the different modes by measuring the accuracy of each dialogue mode as well as the number of turns it takes for the Student to get to the correct answer. A lower number of turns should in fact indicate a more effective dialogue.

[1]Code, data and example dialogues are provided at https://github.com/andreazaninello/MedExpDial

[2]With value = 0 if the answer is not among the options

Mode	Correct	Accuracy	Mean Turns
Mode 0	9	0.45	4.5
Mode 1	13	0.65	5.1
Mode 2	17	0.85	5.0
Mode 3	19	0.95	5.3

Table 1: Explanation-based dialogue effectiveness.

3 Results

The baseline dialogue effectiveness results are reported in Table 1. As expected, injecting more information corresponds to better performances. However, it is to be highlighted that the model, initially unable to answer 0-shot, in our dialogical setting is able to answer correctly 9 of the 20 initial incorrectly answered questions. Moreover, we notice the larger accuracy rise from mode 1 to mode 2, indicating that providing the model with alternative options is particularly effective in guiding the student to the correct answer, results that are even outperformed when providing the model with human-written explanations. This confirms the need for carefully curated data in order to develop efficient explanatory dialogue systems, especially in the medical domain.

4 Conclusions

We presented an approach for developing synthetic explanatory dialogues for medical QA, highlighting the potential of dialogue-based explanations to develop and evaluate argument-based explanation models for medical QA systems. Baseline results suggest that dialogue-based explanations are a promising approach to improving the understandability of medical QA systems. In future work, we plan to move to open models, extend the approach to several languages, as well as analyze the arguments presented by both the Teacher and the Student to identify common argumentation strategies and their impact on the Student's understanding and ability to get to the correct answer.

References

Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, Iker García-Ferrero, Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, German Rigau, and Anar Yeginbergenova. 2023. Hitz@antidote: Argumentation-driven explainable artificial intelligence for digital medicine. In *SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing*.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Preprint*, arXiv:2404.05590.

Anonymous. 2024. Rewired: Instructional explanations in teacher-student dialogues. ACL ARR 2024 February Blind Submission.

Leona Cilar Budler, Lucija Gosak, and Gregor Stiglic. 2023. Review of artificial intelligence-based question-answering systems in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1487.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in biology and medicine*, 155:106649.

Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. 2024. Question answering systems for health professionals at the point of care—a systematic review. *Journal of the American Medical Informatics Association*, 31(4):1009–1024.

Valentin Li'evin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *Patterns*, 5.

Benjamin Molinet, Santiago Marro, Elena Cabrio, and Serena Villata. 2024. Explanatory argumentation in natural language for correct and incorrect medical diagnoses. *Journal of Biomedical Semantics*, 15(1):8.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so i could understand": A dialogue corpus for learning to construct explanations. *Preprint*, arXiv:2209.02508.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. 2024. Medlm: Exploring language models for medical question answering systems. *ArXiv*, abs/2401.11389.

Ackowledgements

We would like to thank the 2023/2024 students of the Language Technology Seminar at the University of Bologna for contributing to the annotation of this dataset. This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and by the ANTIDOTE project (CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002).

Are conversational large language models speakers?

Paul Piwek
The Open University, United Kingdom
paul.piwek@open.ac.uk

> *Fundamental understanding, you can hardly argue with that.*
> Kees van Deemter (van Deemter and Mineur, 1994, 58)

With the advent of large language models (LLM), and in particular their framing as chatbots – that is, conversational agents – the original and time-honoured test for determining whether machines can think, the Turing test (Turing, 1950), has been called into question. We have reached a point where current generations of conversational LLM can pass time-limited versions of the test (Jones and Bergen, 2023). Additionally, the very ability of machines to pass the test is no longer considered to be a genuine indicator of thinking, though it may be a good indicator of the capability for deception (Biever, 2023).

Recently, informal arguments, such as the Octopus test thought experiment (Bender and Koller, 2020) have been put forward purporting to show that systems that are trained only on (language) form cannot understand language. In this paper we will refrain from taking a stance on this argument, and instead raise a further question which considers conversational LLMs from the point of language generation or production rather than understanding. The question we aim to address is: 'Are large language models speakers?' Conversational LLM have brought back to attention fundamental questions about what it means to be a language user and, in line with the quote at the beginning of this paper, we believe this is a good thing.

We start by considering the foundational contribution to linguistic pragmatics made by H.P. Grice (Grice, 1957). Grice investigated what is involved in a speaker meaning something when they use language. In fact, Grice subsumes speaker meaning under, what he calls, non-natural meaning, in contrast with natural meaning. As examples of natural meaning, Grice provides regularities in nature such as smoke meaning fire and a rash meaning measles. Grice proposes that non-natural meaning is fundamentally different from natural meaning. As an example of a situation involving non-natural meaning, Grice asks us to consider that three rings on a bus, at the least in England at the time Grice wrote his paper, meant non-naturally (meant$_{NN}$) that the bus is full. As a first approximation, Grice suggests that such an 'utterance' u has a non-natural meaning if it was intended by its utterer to induce a belief in some 'audience'. Grice then proceeds to refine this description of non-natural meaning by considering cases that reveal the shortcomings of this first approximation: 'I might leave *B*'s handkerchief near the scene of a murder in order to induce the detective to believe that *B* was the murderer; but we should not want to say that the handkerchief (or my leaving it there) meant$_{NN}$ anything or that I had meant$_{NN}$ by leaving it that *B* was the murderer.' (Grice, 1957, 381-382) After further rounds in which Grice considers other limitations of the initial formulation, he eventually arrives at the proposal that *A* meant non-naturally something is equivalent to *A* uttered u with the intention of inducing a belief by means of the recognition of this intention.

Gricean non-natural meaning allows us to characterise speakers as producers of non-natural meanings. The definition does however assume a prior understanding of the notions of belief, intention and recognition. It is tempting to interpret these as psychological states or processes. However, the treatment of such folk psychological notions as foundations for science has been criticised from various angles, e.g., by problematising the concept of belief as foundation for cognitive science (Stich, 1983) and our common sense understanding of conscious experiences (Frankish, 2016). Similarly, the notion of intentions or psychological reasons has not escaped scrutiny: 'Why do you think this? Why did you do that? We answer such questions by giv-

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

ing reasons, as if it went without saying that reasons guide our thoughts and actions and hence explain them. (. . .) It is based, however, on a convenient fiction: most reasons are after-the-fact rationalizations.' (Mercier and Sperber, 2017, 109)

Returning to the topic of conversational LLMs, it is also not clear how to apply these folk psychological concepts to conversational LLMs. It seems somewhat too convenient to simply dismiss the possibility of conversational LLMs as speakers on the basis that they don't have intentions or goals. It is not prima facie clear that they completely lack intentions or at least functionally equivalent states. Though LLM training (i.e. pretraining) is limited to the next word prediction task, conversational LLMs are finetuned in ways that arguably do instill implicit goals on how to follow instructions and avoid inappropriate responses (Ouyang et al., 2022). Furthermore, explicit user prompts or hidden system prompts/context could also be argued to introduce goals.

To be fair to Grice, he specifically writes that he does not want to 'peopl[e] all our talking life with armies of complicated psychological occurrences' (Grice, 1957, 386) and gestures at what is 'normally conveyed', 'refer[ence] to the context', and 'asking the utterer afterward' (Grice, 1957, 387). This line of thought is suggestive of an alternative approach to the question whether conversational LLMs are speakers grounded in a view of language use as participation social practices or Wittgensteinian language games (Wittgenstein, 1953).

A potentially fruitful twist to this approach is proposed by Robert Brandom (Brandom, 1994, 2000), who works out in detail how the language game of giving and asking for reasons is fundamental to all other language games in that this specific game explains the representational power of language - i.e. the language – world relationship. Doing so, he espouses an unusual explanatory move from pragmatics to semantics.

In a nutshell, the game of giving and asking for reasons – for partial formalisations see (Kibble, 2006; Piwek, 2011, 2014) – puts certain normative demands on interlocutors, in particular, an assertion (e.g., 'It rains') results, downstream, in commitments (e.g. prohibiting inconsistent assertions such as 'It doesn't rain' or 'It snows') and, upstream, in potential challenges about the entitlement to or justification for that assertion ('The tiles wet.').

Mastery of this game of giving and asking for reasons may provide us with some insight into the extent to which conversational LLMs are speakers. Interestingly, in as far as commitments and consistency are concerned, conversational LLMs have and continue to struggle with negation (e.g. tests with the prompt 'I do not have two apples. I give one away. How many apples do I have?') causes chatGPT to produce correct responses about 3 out of 5 times, but also bizarre incorrect ones such as 'You have on apples left (...)' (ChatGTP4o, 5 July 2024). Testing Gemini and ChatGPT4o for their way of dealing with contradictions – i.e. challenging its assertions – we found that, after challenging the result of calculating the product of two large numbers, Gemini always concedes that the user is right (even if they clearly aren't) whereas ChatGPT4o, after each challenge, responds with 'To ensure absolute accuracy, I will recompute once again'. Both are appropriate machine responses, but nothing like the behaviour of a speaker who cares about their contribution to the conversation and is sensitive the assessment by others.

This final point is fundamental, resting on the view of speaking (**S**) as a contribution by a person to a language game, i.e. a normative social activity requiring (i) sensitivity to, i.e. caring about, peer assessment of one's contributions and (ii) engagement with peer assessment of others' contributions.

In contrast, automatic natural language generation (**A**) is the algorithmic generation of output strings that we take to be English or French or Chinese or . . . , given a (more or less formal) specification of requirements on the output (e.g. a prompt, logic formula or other).

We'd like to conclude by proposing that the current perspective on speaking and generation raises both a concern and challenge. Let's start with the concern, which can be seen as our variation, and attempt at clarification, of the Eliza effect (Weizenbaum, 1966) and the more general Media Equation (Reeves and Nass, 1996):

The chatbot conceit = *the design of systems that do* **A** *but appear to be in the business of doing* **S** *by framing interactions as dialogue.*

On the positive side, for researchers in pragmatics a daunting but also invigorating challenge remains and has, arguably, been rekindled by the recent advent of conversational LLM:

The pragmatics challenge: *What are the ingredients* **I** *such that* **A** + **I** = **S***?*

Acknowledgments

The argument presented in this paper was originally prepared for an informal gathering on the 6^{th} of July 2024 in honour of Kees van Deemter, who I'd like to thank for the many discussions we've had and will hopefully have in future about meaning, language generation and many other things. The way Kees succeeds in caring for and combining both open-mindedness and rigour exemplifies to me what it means to be a speaker in the sense of this short paper.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Celeste Biever. 2023. ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*, 619:686–689.

Robert Brandom. 1994. *Making It Explicit: reasoning, representing, and discursive commitment*. Harvard University Press, Cambridge, Massachusetts.

Robert Brandom. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, Massachusetts.

Keith Frankish. 2016. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12):11–39.

Herbert Paul Grice. 1957. Meaning. *Philosophical Review*, 66(3):377–388.

Cameron Jones and Benjamin Bergen. 2023. Does GPT-4 Pass the Turing Test? *arXiv preprint*. ArXiv:2310.20216 [cs].

Rodger Kibble. 2006. Reasoning about propositional commitments in dialogue. *Research on Language and Computation*, 4(2-3):179–202.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. ArXiv:2203.02155 [cs].

Paul Piwek. 2011. Dialogue structure and logical expressivism. *Synthese*, 183(1):33–58.

Paul Piwek. 2014. Towards a computational account of inferentialist meaning. In *Proceedings of the 50th Anniversary Convention of the AISB*.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press.

Stephen Stich. 1983. *From folk psychology to cognitive science: The case against belief.* The MIT Press.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Kees van Deemter and Anne-Marie Mineur. 1994. Kees van Deemter interviewed by Anne-Marie Mineur: "Fundamental begrip, daar kun je bijna niet tegen zijn.". *Ta! studentenblad computationale taalkunde*, 4(2):58–69.

Joseph Weizenbaum. 1966. Eliza a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Ludwig Wittgenstein. 1953. *Philosophical investigations. Philosophische Untersuchungen*. NY, Macmillan. (1953). x, 232 pp.

Pre-Generative Conversational AI

Staffan Larsson
Gothenburg University and Talkamatic AB
Gothenburg, Sweden
staffan.larsson@ling.gu.se, staffan@talkamatic.se

Abstract

LLMs such as ChatGPT have raised expectations on Conversational AI (CAI) applications, yet deployment is often hindered by controllability problems. This paper, to be accompanied by a live demo, describes Pre-Generative Conversational AI (PGCAI) and its implementation in Talkamatic Dialog Studio, a tool suite for creating high-quality controllable conversational AI application without the need for coding, prompting or manual dialogue building.

1 Introduction

Generative AI in the form of Large Language Models such as ChatGPT is currently re-shaping the conversational AI landscape, and is generally taken to enable a multitude of practical Conversational AI applications in many different areas, including customer service, education, and more.

However, many companies and organisations are also hesitant when it comes to using an LLM-driven conversational agent to (for example) represent them on their website, or engage in one-to-one educational dialogue with children in schools. One reason for this is a host of well-known problems deriving from the overall problem of controlling the behaviour of LLMs. This may result in generating outputs that do not adhere to the desired agent behaviour (Kann et al., 2022).

For many applications of LLMs, such problems can be handled by manually checking the output of the LLM before using it (e.g. publishing a text or sending an email). However, in conversational AI applications, this is typically not an option, as the LLM interacts directly in real time with users.

This paper proposes a solution to this problem in the form of "Pre-Generative Conversational AI". Instead of letting the user talk directly to a generative AI, with the risks that entails, we instead use generative AI to generate dialogues *before* they are published. At runtime, the dialogue can be handled without using LLMs at all, or using them only for limited tasks such as NLU.

In essence, PGCAI enables using our normal preferred way of working with LLMs (generate-curate-publish) also for conversational AI.

2 Key components

PGCAI has to three key components: a dialogue generator, a dialogue editing tool, and a flexible but controllable dialogue manager.

2.1 Dialogue generator

The dialogue generator uses LLMs to create dialogues based on some content. Of course, there are many types of dialogue one could have about some content. Hence, the dialogue generator relies on distinguishing different dialogue types, or *genres* (Larsson, 2002; Ginzburg and Wong, 2024). Examples of such genreas are educational, instructional, question-answering and negotiative dialogue.

The task of the dialogue generator is to take some content (a text, a database or something else) a specification of a dialogue genre, and produce a dialogue blueprint which can then be used by the dialogue manager to engage in a flexible dialogue. For each type of dialogue, it uses genre-specific prompts to produce dialogues of the type selected by the dialogue designer.

2.2 Dialogue curation tool

Since PGCAI does not require designing or implementing a dialogue (in the form of code or using a GUI), nor requires any prompt writing, we do not use the term "dialogue designer". Instead, the role of the human in building a dialogue application is to *curate* the dialogue, in the sense of taking an existing dialogue blueprint and adapting and perfecting it for the precise use it will be put to. To aid in this process, a dialogue curation tool is needed.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

After a dialogue has been generated, it can immediately be tested by interacting with it. If the curator is unhappy with some aspect of the dialogue, they can go in and inspect and edit the dialogue blueprint. The precise structure of this blueprint will depend on the dialogue genre. For question-answering dialogue, the main component is a list of question-answer pairs. For education dialogue, it is a pedagogical interaction consisting mainly of questions of various kinds (right/wrong questions asking about information offered explicitly in the text, or requiring some inference on the part of the user, more open questions asking the user to reflect, and more). Other elements are also present, such as a list of potentially difficult words that the system can explain if needed. For other types of dialogue, other structures are available for curation. Importantly, these structures are quite simple and editing them does not require any deep technical understanding of conversational AI or even of human dialogue. However, genre-specific competence can often be useful, such as pedagogical skills in the case of educational dialogue.

2.3 Flexible dialogue management

LLMs are widely recognised as going considerably beyond the state of the art when it comes to NLU. For this reason, we allow for using LLMs to take care of NLU even when not using them to generate responses to the user. A similar approach is taken in Rasa (Bocklisch et al., 2024). Talkamatic Studio allows the dialogue designer to decide what NLU to use, offering LLMs as options but also non-LLM technologies.

LLMs are also quite adept at handling many different kinds of dialogue in a flexible way, meaning that they often respond appropriately to less expected or less routine user behaviours. The success of PGCAI depends crucially on the ability of the system to achieve dialogue behaviour on par with or surpassing an LLM. Hence, we need to achieve a high level of flexibility in PGCAI, despite the fact that the dialogue blueprints are not generated at runtime. This poses considerable challenges for the dialogue manager. Talkamatic have developed the Talkamatic Dialogue Manager (Larsson and Berman, 2016) which supports a wide (and growing) variety of conversational behaviours across several dialogue genres, including the ones mentioned above.

TDM is based on the Information State Update approach to dialogue management, and more specifically Issue-Based Dialogue Management (Larsson, 2002). As part of a series of research projects and later in Talkamatic, TDM has been gradually extended to cover an increasing range of dialogue behaviours and dialogue genres.

3 Talkamatic Studio

Talkamatic Studio[1] is a comprehensive software service offering all the components needed for PGCAI. It offers a dialogue generator, a dialogue curation tool, a runtime frontend and backend using TDM for dialogue management, an LLM control panel, and a dialogue analytics tool.

4 Related work

Of course, the control problem for LLMs is not new and a lot of work is being done to address it. The absolute majority of methods for dealing with this problem is of the "guardrails" type. In LLM-based Conversational AI, however, the user is still interaction with an LLM at runtime, and it is difficult or impossible to guarantee that guardrails always work. Ayyamperumal and Ge (2024) discuss various guardrail approaches such as layered protection models, system prompts, Retrieval-Augmented Generation (RAG) architectures and bias mitigation, and observe that "[c]rucial challenges remain in implementing these guardrails." Xu et al. (2024) show that hallucination is not just a temporary glitch, but are in fact inevitable in LLMs.

We believe that in many applications, including using Conversational AI agents for education in schools and to represent companies and organisations, there will be a strong preference for zero risk solutions, i.e. solutions that can *guarantee* there will be no bad output from the system.

5 Conclusion and future work

We have presented Pre-Generative Conversational AI and its implementation in Talkamatic Studio. This approach and implementation addresses a central problem with using LLMs for Conversational AI - the lack of control. To the best of our knowledge, Talkamatic Studio is the only solution that combines dialogues generated by LLMs, control (including complete control with no LLM output generation at runtime), curation (putting a human in the loop), and flexible dialogue across several dialogue genres, going beyond form-filling dialogue.

[1] https://talkamatic.se

References

Suriya Ganesh Ayyamperumal and Limin Ge. 2024. Current state of llm risks and ai guardrails. *arXiv preprint arXiv:2406.12934*.

Tom Bocklisch, Thomas Werkmeister, Daksh Varshneya, and Alan Nichol. 2024. Task-oriented dialogue with in-context learning. *arXiv preprint arXiv:2402.12234*.

Jonathan Ginzburg and Kwong-Cheong Wong. 2024. Language games and their types. *Linguistics and Philosophy*, 47(1):149–189.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*. Association for Computational Linguistics.

Staffan Larsson. 2002. *Issue-Based Dialogue Management*. Ph.D. thesis, Department of Linguistics, Goteborg University.

Staffan Larsson and Alexander Berman. 2016. Domain-specific and general syntax and semantics in the talkamatic dialogue manager. *Empirical Issues in Syntax and Semantics*, 11:91–110.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

It is difficult, but not impossible: Measuring Scalar Activation in Language Models

David Arps and **Yulia Zinova**
Heinrich Heine University Düsseldorf
david.arps@hhu.de, zinova@hhu.de

Abstract

We explore the behaviour of language models on adjectival scales by analyzing activation changes when prompted with related and unrelated adjectives. We find evidence for scale activation, which aligns with results from human priming experiments.[1]

1 Introduction

Scalar diversity has been extensively studied in experimental setups with human participants when testing implicature endorsement rates (Van Tiel et al., 2016; Sun et al., 2018; Gotzner et al., 2018; Ronai and Xiang, 2022). Priming experiments (e.g. Lacina and Gotzner, 2024) explore the link between implicature computation and lexical priming. They find that priming with a weak scalemate leads to faster recognition of the strong scalemate.

Hu et al. (2023) show that pragmatic inference tasks pose great challenges for language models (LMs). Nizamani et al. (2024) show that DeBERTa models perform poorly on scalar implicatures, even after fine-tuning. As the availability of alternatives is considered to be the basis of implicature computation (Gotzner and Romoli, 2022), we analyze the activation of scalar adjectives in the LM.

2 Experimental Setup

Activation of strong adjectives In our first experiment, we follow the design used in Lacina and Gotzner (2024) and Ronai and Xiang (2023). In human priming experiments, participants were presented with sentences that carry either a related (1-a) or an unrelated (1-c) adjective. After that, participants were asked to perform a lexical decision task and their reaction times were recorded. Both Ronai and Xiang (2023) and Lacina and Gotzner (2024) found that participants recognized stronger adjectives as existent words faster when the preceding sentence contained the weak scalar item.

[1]We will release our code upon publication.

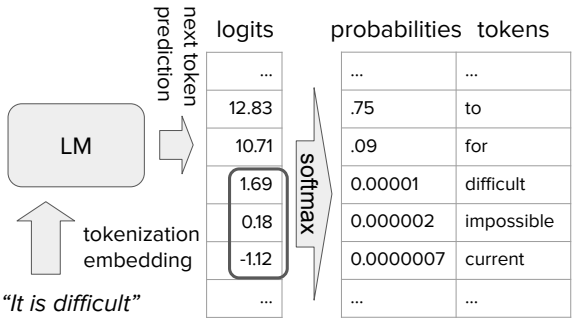


Figure 1: We collect next word prediction logits from the LM to measure activation of scalar concepts.

(1) a. It is difficult. (WEAK)
b. It is impossible. (STRONG)
c. It is current. (UNRELATED)

To test whether a similar effect can be observed within a LM, we collect activations for strong adjectives after either a weak or an unrelated adjective has been processed. The hypothesis that corresponds to human behaviour is that the activation of the strong adjective should be higher after the model processes a weak adjective, in comparison with processing an unrelated adjective.

Activation of weak adjectives We invert the prime/target adjectives from the previous setup and collect activations of the weak adjectives given either a related (1-b) or an unrelated (1-c) prompt.

Activation difference We use both setups above to check whether LM behaviour aligns with the results of De Carvalho et al. (2016) for humans, who found that weak terms activate the respective strong ones more than strong terms activate weak ones (in French).

Activation of unrelated adjectives As a control condition, we collect activations of unrelated adjectives after prompts with weak and strong adjectives.

Activation without context Ronai and Xiang (2023) did not find evidence for priming when par-

activation condition	Lacina and Gotzner (2024) of strong weak, unr.	of weak strong, unr.	diff.	of unr. strong, weak	No Context of strong weak, unr.	of weak strong, unr.	diff.	of unr. strong, weak
125M	****	****	n.s.	n.s.	****	****	n.s.	n.s.
350M	****	***	n.s.	*	****	*	*	*
1.3B	****	****	n.s.	n.s.	****	****	n.s.	n.s.
2.7B	****	****	n.s.	n.s.	****	****	n.s.	n.s.
6.7B	****	****	n.s.	n.s.	****	****	n.s.	n.s.

Table 1: Significance test results, where * stands for $p < 0.05$, *** for $p < 0.001$ and **** for $p < 0.0001$.

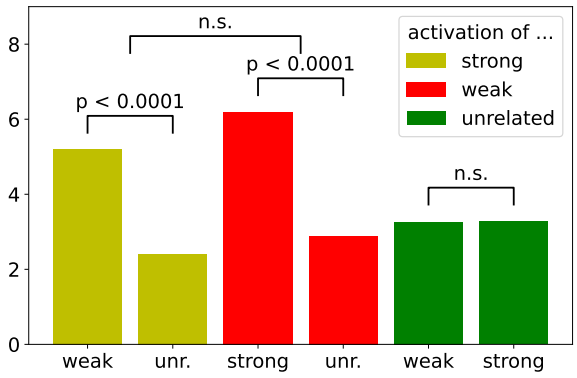


Figure 2: Scalar activation of adjectives after various prompts (OPT 125m). Individual bars show conditioning of the scalar terms.

ticipants were presented with isolated lexical items. To check whether LMs are sensitive to this, we repeat all of the above settings such that the LM is presented with the adjectives in isolation.

2.1 Scale selection

We use experimental materials from Lacina and Gotzner (2024), which contains constant sentence frames and focuses on one grammatical class (adjectives). To mitigate tokenization effects, we exclude 18 of 64 scales where any of the adjectives is split into more than one subword token.

3 Activation for language models

We use next token prediction models, which assign weights (logits) that indicate how well a token is activated by the context (Fig. 1). The softmax function transforms the logits into a probability distribution over the vocabulary, which is used for next word prediction. As an effect, tokens that are ranked among the top 10% of continuations receive low probabilities (see adjectives in Fig. 1). Compared to softmax probabilities, logits for individual tokens are relatively independent from each other. We calculate the activations of strong adjectives from both probabilities and logits, and use the paired sample t-test for the condition effect for the strong adjective activation. We find that the effect is significant for logits (p<0.0001) but not probabilities (p=0.21). This finding aligns with the research on the internal prediction construction process of LMs (Geva et al., 2022). In what follows, we use logit values as the activation measure.[2]

4 Results

We test OPT (Zhang et al., 2022) models of varying sizes from 125M to 6.7B parameters. All but one model demonstrate similar behaviour both with and without context (Tab. 1). Fig. 2 presents results for the smallest model: Activation of strong and weak adjectives is significantly higher after a related adjective; activation difference and activation of unrelated adjectives do not vary significantly.

The only exception is the second smallest (350M) model. Because we do not have insights into the training process of the models, we refrain from making claims about the reason for this unexpected behaviour.

5 Discussion

The presented setup allows to study the activation of vocabulary items beyond discrete token predictions. This allows to test whether linguistic concepts (e.g., scalar activations) are captured by the LM. As a next step, we will examine scalar activation in more complex contexts as in Sun et al. 2018 and Nizamani et al. 2024, and track the development of activations at several points in the sentence. We will also test whether linguistic features of the scales (e.g., boundedness) correlate with the magnitude of the activation effect for LMs, and whether the difference between the models is reflected in fine-tuning results.

[2]The absolute logit value depends not just on the vocabulary item, but also other factors such as sentence length. We subtract the mean of the logits over the vocabulary for presentational reasons. This does not affect the significance of the described effects.

References

Alex De Carvalho, Anne C Reboul, Van der Henst, Anne Cheylus, Tatjana Nazir, et al. 2016. Scalar implicatures: The psychological reality of scales. *Frontiers in psychology*, 7:203305.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicole Gotzner and Jacopo Romoli. 2022. Meaning and alternatives. *Annual Review of Linguistics*, 8(Volume 8, 2022):213–234.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, 9:1659.

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, Toronto, Canada. Association for Computational Linguistics.

Radim Lacina and Nicole Gotzner. 2024. Exploring scalar diversity through priming: A lexical decision study with adjectives. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795, Torino, Italia. ELRA and ICCL.

Eszter Ronai and Ming Xiang. 2022. Three factors in explaining scalar diversity. In *Proceedings of sinn und bedeutung*, volume 26, pages 716–733.

Eszter Ronai and Ming Xiang. 2023. Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, 2:229–240.

Chao Sun, Ye Tian, and Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology*, 9:2092.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of semantics*, 33(1):137–175.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Getting to the point: Contrasting Directness and Warmth in Motivational Embodied Conversational Agents

Michael O'Mahony, Cathy Ennis, Robert Ross
School of Computer Science, Technological University Dublin, Ireland
michael.t.omahony@mytudublin.ie, {cathy.ennis, robert.ross}@tudublin.ie

Abstract

Enhancing long-term engagement with conversational agents remains a significant challenge. Controlling the perceived warmth or directness of an agent's personality through the style of its generated text could be used to increase user likeability. This paper reports an investigation of a Wizard-of-Oz (WoZ) mediated study of two variants of a motivational embodied conversational agent to measure user perception of and attitudes towards warmth in interaction style[1]. Results show a significant effect of users preferring an agent with a "more direct" personality for this scenario, though this effect is in many ways nuanced.

1 Introduction

There have been many advancements in data-to-text generation in recent years, especially through the use of neural networks (Lin et al., 2020; Su et al., 2021), and more recently, Large Language Models (LLMs). Most of this work has focused on content fidelity rather than text style (Lin et al., 2020; Su et al., 2021). However, LLMs have significantly improved our ability to style content. Since the release of ChatGPT in November 2022, dialogue systems have been applied to many more tasks, but the challenge of keeping a user engaged with an agent and understanding how the nuances of the agent can be tailored to enhance specific conversational goals, for example around motivation intervention, remains a very real research challenge. Moreover, the style of generated text can change the perceived personality of an agent and hence impact likeability and engagement.

Text style is an important aspect of generated text as a wide variety of applications require that information is given in a certain way. Considering that text generation plays a large role in the user satisfaction of a dialogue system (Peng et al., 2020), dialogue systems that aim to imitate a human agent can appear to have a consistent personality through a reliable, controllable style of conversation.

While style, personality and its relationship to engagement and likeability is of relevance to semantics and pragmatics study in general, in this paper we are particularly focused on the domain of Motivational Interview (MI) agents. MI is a counselling technique used to increase a person's motivation to change their behaviour. Some other studies have researched the impact of MIs delivered by intelligent agents on users.

We hypothesise that agent "warmness" vs "directness" will impact participant likeability leading to differences in responses to the general agent ratings (see section 3). We also hypothesise that, from these ratings, there will be a preference of one simulated personality over the other.

2 Related Work

The impact of personality variations in the healthcare domain has already been the subject of significant study. Many of these works look at agent empathy (Barange et al., 2022; Chauvin et al., 2023) but there are also works focused on other aspects such as humour (Olafsson et al., 2020) and adaptivity (Egede et al., 2021).

A few studies have employed an agent to deliver MIs to participants to increase their motivation to eat healthier (Olafsson et al., 2020, 2019), exercise more (Olafsson et al., 2020, 2019; Galvão Gomes da Silva et al., 2020, 2018; Chauvin et al., 2023), or quit unhealthy behaviours such as excessive alcohol consumption (Olafsson et al., 2023). Some of these studies used animated conversational agents (Olafsson et al., 2019, 2020; Chauvin et al., 2023), a NAO robot (Galvão Gomes da Silva et al., 2018), or video recordings of human actors (Galvão Gomes da Silva et al., 2020) to build the ECA.

[1]Data will be available at https://github.com/Michael-OMahony/getting-to-the-point-data/ after 01/01/2025.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

3 Experiment Design and Methodology

To investigate the perception and impact of directness variation in MI agents, we conducted an online WoZ between-subjects user study to measure the likeability of the ECA. The interaction scenario was an MI delivered by a virtual agent to increase users' motivation to change their exercise behaviour. Participants interacted with the ECA via voice though mediated through an online interface. We recruited 25 participants from local communities. Participants were also given a questionnaire, and a number of concrete metrics were collected alongside recordings of the interactions.

The interview script was adapted from an earlier study that used a NAO robot to deliver an MI to participants to help increase their motivation to change their exercise habits (Galvão Gomes da Silva et al., 2018). While our experiment used a WoZ setup, in the original study the participant would control when the next utterance was delivered by pressing a button on the robot's head. The authors designed the script so that each question should make sense to the user, irrespective of how they answered previous questions. In practice this method mostly worked, but there were instances where a somewhat broad question lead to some confusion.

Building on the existing corpus, we created two conditions by altering the text style of parts of the original script using ChatGPT to create "warmer" and "more direct" versions of the agent script. In practice, we only changed the beginning and end of the script, aside from a minor change in the first question for clarity, we did not alter any of the questions as designing a counselling intervention was outside the scope of this work, and we believe the start and end of an interaction are influential on user satisfaction. There were no options to change the next utterance based on the participants responses, but we could repeat the last question upon request.

Participants answered a questionnaire before and after the interaction. The pre-interaction questionnaire included demographics, exercise frequency, Ten-Item Personality Inventory (TIPI) for the participant (Gosling et al., 2003), familiarity and attitudes towards virtual agents. The post-interaction questionnaire included the TIPI for the virtual agent (Gosling et al., 2003), general agent ratings (Olafsson et al., 2019, 2020), and an open-ended feedback box. The general agent rating questions were Q1:"I am satisfied with the agent", Q2:"I would continue talking with the agent", Q3:"I trust the agent", Q4:"I like the agent", Q5:"The agent was knowledgeable", Q6:"The conversation was natural", Q7:"I have a good relationship with the agent", and Q8:"I am similar to the agent". Participants rated the agent using a five point Likert scale.

As the focus of our work was on embodied agents rather than text or speech only based interaction, the agent was given a virtual appearance. For this we used the Unity game engine, along with a Ready Player Me avatar. Moreover, we used the Talking With Hands dataset for the talking gestures (Lee et al., 2019), Ready Player Me animation library for the idle animation, and Salsa Lip Sync. Google's Cloud AI text-to-speech was used for the agent's voice, where we selected a female avatar as some studies suggest that men slightly prefer a female therapist to a male one or do not care, and women are much more likely to prefer a female therapist (Liddon et al., 2018; Seidler et al., 2022). In general, each experiment lasted 20-30 minutes, with the interaction lasting 5-15 minutes.

4 Results

The interaction times between the "warmer" and "more direct" conditions were not statistically significant (means=680, 646s). Table 1 presents mean Likert Ratings for each of the key likeability questions. The means for the responses to every question were higher for the "more direct" condition though when analysed on a question by question basis, the only question which demonstrated statistically significant difference was Q7. Potential limitations were the sample size, and our inability to alter most of the agent script. Future work will focus on nuancing the qualities of directness and warmness in speech and embodying these in a more automated agent with evaluation of effectivness as well as engagement.

Q	Warm	Direct	Difference	Sig.
1	3.46	3.83	-0.37	0.4494
2	3.08	3.33	-0.26	0.5729
3	3.15	3.50	-0.35	0.4058
4	3.53	4.00	-0.46	0.0953
5	3.23	3.50	-0.27	0.5685
6	2.85	3.00	-0.15	0.5495
7	**2.85**	**3.42**	**-0.57**	**0.0379***
8	2.15	2.25	-0.10	0.8193

Table 1: Means and significance for each question.

5 Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Mukesh Barange, Sandratra Rasendrasoa, Maël Bouabdelli, Julien Saunier, and Alexandre Pauchet. 2022. Impact of adaptive multimodal empathic behavior on the user interaction. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA. Association for Computing Machinery.

Rachel Chauvin, Céline Clavel, Nicolas Sabouret, and Brian Ravenet. 2023. A virtual coach with more or less empathy: impact on older adults' engagement to exercise. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, IVA '23, New York, NY, USA. Association for Computing Machinery.

Joy Egede, Maria J. Galvez Trigo, Adrian Hazzard, Martin Porcheron, Edgar Bodiaj, Joel E. Fischer, Chris Greenhalgh, and Michel Valstar. 2021. Designing an adaptive embodied conversational agent for health literacy: a user study. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, IVA '21, page 112–119, New York, NY, USA. Association for Computing Machinery.

Joana Galvão Gomes da Silva, David J Kavanagh, Tony Belpaeme, Lloyd Taylor, Konna Beeson, and Jackie Andrade. 2018. Experiences of a motivational interview delivered by a robot: Qualitative study. In *J Med Internet Res*.

Joana Galvão Gomes da Silva, David J. Kavanagh, Jon May, and Jackie Andrade. 2020. Say it aloud: Measuring change talk and user perceptions in an automated, technology-delivered adaptation of motivational interviewing delivered by video-counsellor. *Internet Interventions*, 21:100332.

Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 763–772.

Louise Liddon, Roger Kingerlee, and John A. Barry. 2018. Gender differences in preferences for psychological treatment, coping strategies, and triggers to help-seeking. *British Journal of Clinical Psychology*, 57(1):42–58.

Shuai Lin, Wentao Wang, Zichao Yang, Xiaodan Liang, Frank F. Xu, Eric Xing, and Zhiting Hu. 2020. Data-to-text generation with style imitation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1589–1598, Online. Association for Computational Linguistics.

Stefan Olafsson, Teresa O'Leary, and Timothy Bickmore. 2019. Coerced change-talk with conversational agents promotes confidence in behavior change. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth'19, page 31–40, New York, NY, USA. Association for Computing Machinery.

Stefan Olafsson, Teresa K. O'Leary, and Timothy W. Bickmore. 2020. Motivating health behavior change with humorous virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Stefan Olafsson, Paola Pedrelli, Byron C. Wallace, and Timothy Bickmore. 2023. Accomodating user expressivity while maintaining safety for a virtual alcohol misuse counselor. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, IVA '23, New York, NY, USA. Association for Computing Machinery.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 172–182, Online. Association for Computational Linguistics.

Zac E. Seidler, Michael J. Wilson, David Kealy, John L. Oliffe, John S. Ogrodniczuk, and Simon M. Rice. 2022. Men's preferences for therapist gender: Predictors and impact on satisfaction with therapy. *Counselling Psychology Quarterly*, 35(1):173–189.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, page 895–909, Punta Cana, Dominican Republic. Association for Computational Linguistics.

"No, you listen!" A pilot experiment into escalation devices in confrontational conversation

Sara Amido
Universitat Pompeu Fabra
sara.amido@upf.edu

Vladislav Maraev
University of Gothenburg
vladislav.maraev@gu.se

Christine Howes
University of Gothenburg
christine.howes@gu.se

Abstract

In conversation, interlocutors are often friendly or polite and generally socially collaborative. However, it is not uncommon that interlocutors get upset, defensive, and antagonistic, engaging in confrontational conversation. Given that we are able to intuitively perceive the contrast between confrontational and non-confrontational conversation, our goal is to find out whether there is a linguistically manifested contrast between the two contexts. A corpus of confrontational conversation was analysed for potentially escalating linguistic devices. In this paper we propose an exploratory experiment where we test these devices to find out whether they, in fact, correspond to the perceived escalation of confrontation in a conversation.

1 Introduction

We define confrontational conversation (CC henceforth) generally as an antagonistic exchange characterised by blaming, insults, personal attacks, and so on, where interlocutors express themselves in a non-collaborative and even combative manner (see e.g. Walton 1998 on *eristic dialogue*). The contrast between CC and non-CC is often intuitively clear to speakers, therefore the overarching aim of this research is to explain the roots of this intuition. We tackle this question by attempting to find distinctive linguistic features of CC.

To study the linguistic properties of CC, a relevant contribution in the literature is that of impoliteness strategies (Lachenicht 1980, Austin 1990, Culpeper 1996, Bousfield 2008, a.o.) since these are concerned with attacking face (Culpeper, 1996), where face is loosely defined as "one's public self-image" (Brown and Levinson, 1987, 61). We assume that face-attack (or face-threat) in interaction escalates confrontation in conversation provided certain conditions that enable it are met, including a balanced power dynamic between interlocutors, similar cultural background, etc.

Culpeper (1996) proposes impoliteness *super-strategies*, which are classified according to how they interact with face (whether they threaten face directly or indirectly, whether they threaten negative or positive face, etc.) in a symmetrically opposite fashion to the taxonomy of superstrategies in Politeness theory (Brown and Levinson, 1987). Embedded hierarchically within superstrategies are *output strategies*, which are an open-ended list of ways to achieve the former. Examples of these include *seeking disagreement* or *using taboo words*. At a macro-level, Culpeper (2011, 136) proposes impoliteness formulae, which are concrete linguistic structures which have been attested to correlate with impoliteness, e.g. *shut* [*the fuck*] *up*.

Our research directly relates to impoliteness formulae in Culpeper (2011). We aim to determine whether explicit linguistic devices are perceived as impolite/aggressive/face-threatening. We present a bottom-up approach to testing corpus-sourced devices for their potential to escalate confrontation.

In order to do so, we must supply two things: context and interpretation (Culpeper, 2016). The importance of context has long been discussed with respect to impoliteness strategies. Tracy and Tracy (1998) propose that those which are perceived as impolite across most contexts are *context-spanning* strategies, whereas strategies which are perceived as impolite only in some contexts are *context-tied*. Essentially we are testing whether devices found in CC data are escalating when used in a new context (context-spanning) or not systematically perceived as offensive (context-tied). Secondly, impoliteness strategies also require that interlocutors actually perceive impoliteness (/aggression/face-threat), which we will assess by measuring interlocutors rapport "the experience of harmony, fluidity, synchrony, and flow felt during a conversation" (Gratch and Kang, 2015). Since CC is antagonistic, we assume that confrontation and rapport are negatively correlated.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

2 Corpus analysis

The data used for the corpus analysis were selected transcribed dialogues from British reality television show *The Only Way is Essex* (TOWIE corpus[1]). Turns were annotated as potentially aggressive if they seemed to escalate confrontation, i.e. if reactions to those turns, as well as subsequent turns, provided indication that aggression/face-threat was perceived.

These turns were grouped as different generalised devices which included: repetition (full or partial) of the interlocutor's previous turn; returning the speech act, particularly directives; second person reports, i.e. a statement about the addressee for which the latter has epistemic authority; insults; rhetorical questions; patronising commands; vocatives, specifically turn-final addressee's name.

2.1 Selected devices and examples

Three devices were selected for testing: second person reports, patronising commands, and turn-final addressee's name. They are exemplified in (1), (2), and (3) respectively, with devices in boldface. The following are adapted from the TOWIE corpus.

(1) MEG: I react by screaming and shouting. Like, I can't help it. It's who I am.
CHL: Okay, well. **You can help it. You can.**
MEG: Well, I can't! It's the way I am when I've been hurt! ((indignant expression))

(2) YAZ: Now let's be honest.
LCK: Listen.
YAZ: You message me as well.
LCK: No, no no, **listen**.
YAZ: No, no no, you listen!

(3) MEG: You're boring.
CHL: Who are you talking to?
MEG: You're boring**, Chloe.**
CHL: Good, you don't have to be around me!

2.2 Research question

In the corpus, these devices do not occur on their own. Since we are asking whether they should be classified as context-spanning escalating devices, we must take into account that in the corpus they are concurrent with other potentially confounding phenomena. For example, the second person report in (1) and the patronising command in (2) are coupled with disagreement, whereas the turn-final vocative in (3) is preceded by an insult. Therefore, our aim is to test these devices in different contexts. Our research question is whether these linguistic devices – second person reports, patronising commands, and turn-final addressee's name – escalate confrontation in interaction.

3 Method

3.1 DiET chat tool

The Dialogue Experimental Toolkit (DiET) chat tool (Healey et al., 2003) is a text-based chat interface into which interventions, such as adding fake turns, can be introduced into a dialogue in real time, thus causing a minimum of disruption to the 'flow' of the conversation. For this experiment we will use the version of DiET which runs through the messenger app Telegram.[2]

3.2 Participants and procedure

40 fluent English speakers will be recruited and grouped into 20 pairs, with 10 pairs in the intervention condition and 10 pairs as controls (with no interventions). Participants will be prompted to discuss the balloon task (see Section 3.3) via chat on Telegram for 20 minutes.

In this experiment, the three selected devices in Section 2.1 will be automatically inserted into the chat via DiET with a set number of turns between interventions. The devices inserted via DiET appear to be sent by the participants themselves. That is, when messages are sent or modified on behalf of p1, they appear to p2 as sent by p1, and are not visible at all to p1; likewise for when messages are sent on behalf of p2.

All participants will subsequently be asked to fill in a survey evaluating the rapport with their interlocutor based on the chat.

3.3 Task

The balloon task is an ethical dilemma, where participants must discuss which of four people in a hot air balloon should be sacrificed to save the other three. There are potential reasons for saving or sacrificing each person, and the task usually leads to lively discussions (see e.g. Howes et al., 2021). Since such ethical dilemmas give rise to questions and opinions concerning sensitive topics (where the notion of face is salient in the interaction), it provides a context that allows for CC to ensue.

[1] https://www.sara-amido.com/research/resources

[2] https://dialoguetoolkit.github.io/chattool/

References

Paddy Austin. 1990. Politeness revisited - the dark side. In Alan Bell and Janet Holmes, editors, *New Zealand Ways of Speaking English*, pages 277–293. Multilingual Matters, Clevedon.

Derek Bousfield. 2008. *Impoliteness in Interaction*. John Benjamins Publishing Company.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics*, 25(3):349–367.

Jonathan Culpeper. 2011. *Impoliteness: Using Language to Cause Offence*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Jonathan Culpeper. 2016. Impoliteness strategies. In Alessandro Capone and Jacob L. Mey, editors, *Interdisciplinary Studies in Pragmatics, Culture and Society*, pages 421–445. Springer International Publishing, Cham.

Jonathan Gratch and Sin-Hwa Kang. 2015. Gratch lab rapport measures. online.

P. G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, Massachusetts.

Christine Howes, Ellen Breitholtz, Mary Lavelle, and Robin Cooper. 2021. Justifiable reasons for everyone: Dialogical reasoning in patients with schizophrenia. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*. SEMDIAL.

Lance G. Lachenicht. 1980. Aggravating language a study of abusive and insulting language. *International Journal of Human Communication*, 13(4):607–687.

Karen Tracy and Sarah J. Tracy. 1998. Rudeness at 911: Reconceptualizing face and face attack. *Human Communication Research*, 25(2):225–251.

Douglas N. Walton. 1998. *The New Dialectic: Conversational Contexts of Argument*. University of Toronto Press.

To Your Left: A Dataset and a Task of Spatial Perspective Coordination

Mattias Appelgren
FLoV and CLASP
University of Gothenburgh
mattias.appelgren@gu.se

Simon Dobnik
FLoV and CLASP
University of Gothenburgh
simon.dobnik@gu.se

Abstract

In dialogue speakers speak about the same scene while looking at it from different points of view. Who's view is utilised in utterances shifts inside the same conversation and is coordinated by participants as part of their common ground. However, current AI systems are generally trained on a single perspective or multiple random perspectives and are incapable of such coordinations. In this paper we propose a novel artificial dataset that we are developing as a part of our ongoing work with the purpose of evaluating the current state of the art on their ability to learn to recognise and generate spatial descriptions where the speaker and listener have different points of view.

1 Introduction

When humans communicate with each other we have to consider whose Point of View (POV) or Frame of Reference (FoR) (in this paper we use these terms interchangeably) a description is given from (Levinson, 2003). For example, "The tiger is hiding in the bushes to the right of the child" in this example there are at least three different POVs to consider: the speakers, the listeners, and the child's. The listener would need to infer which POV to use in order to complete its intended task, e.g. aiming a tranquilizer at the correct bush. Furthermore, if the listener later becomes the speaker in the same conversational and situational context, what perspective they would take in their utterance? Current state of the art models struggle with spatial relations on their own (Kelleher and Dobnik, 2017; Liu et al., 2023), and very few consider FoR explicitly (some notable exceptions include Lee et al. (2022); Hua et al. (2018); Steels and Loetzsch (2006)). However, Dobnik (2009) found that even when participants are asked to use a fixed FoR they would shift it in response to different situations. Dobnik et al. (2020) further study this phenomanon in human dialogues and find that people will shift FoR throughout extended dialogues, often without explicitly marking the shift.

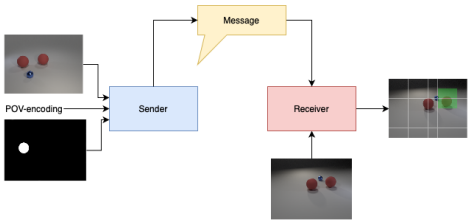


Figure 1: The speaker sees the image, a mask to identify the target, and the listener's POV encoded as a 1-hot vector. It produces a message referring to the target object. The listener sees the same scene from a different POV and receives the message and must predict the region which contains the described object.

In order for robots and other AI systems to communicate successfully with humans they need the capability to generate and interpret referring expressions from different FoRs and in continuous conversational and situational contexts. In this paper we propose an artificial dataset and task which will diagnose systems' ability to consider FoR in spatial descriptions and test conditons under which FoR can be learned by them. We describe work in progress, which means we have not completed the development of this data nor any experiments.

2 Dataset and task

2.1 Task

In our task two agents must communicate about a scene which they are viewing from different POVs. The agents take on the roles of speaker or listener. The speaker is shown a visual scene and an object within the scene that it must refer to. The listener sees the scene from a different POV. The speaker must generate a message describing the target object and the listener must interpret the message and predict which region of the image the object is in.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

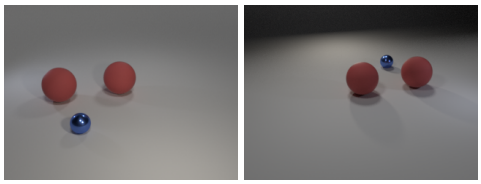


Figure 2: Two views of the same scene.

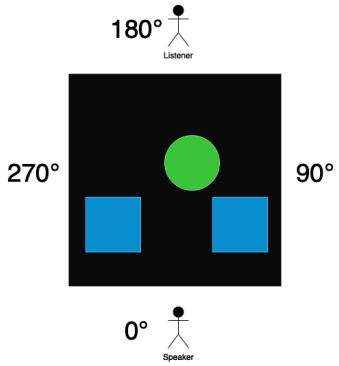


Figure 3: The listener could view the world from four different angles relative to the speaker

Figure 1 shows a diagram of the task set-up.

2.2 Data

We opt for artificial scenes so that we can control precisely the contextual attributes of the interaction environment. The first consideration is that the target object must not be uniquely identifiable from its visual attributes. In Figure 2, if the target was the blue sphere it would be enough to describe it as such to identify it. However, if the target is one of the two red spheres spatial descriptions would have to be used, e.g. "the leftmost red sphere". As such, each image will contain a target object and one or more distractors that share all of the same visual features as the target, in addition to landmark objects which have different visual features to the target, such as the blue sphere in Figure 2. We will capture the scene from four directions, as shown in Figure 3. In different sub-tasks we will experiment with showing the speaker and listener different combinations of views, for example to give the agents the ability of egocentric perspective shifts (Levinson, 2003).

We will use the code that generated the CLEVR dataset (Johnson et al., 2016) to generate the images, potentially extending it to more general objects as done by Lee et al. (2022), both use the Blender graphics software to render images of objects. Figure 2 shows an example of the same scene from two opposite perspectives.

2.3 Experiments

We will implement the speaker and listener in the EGG toolkit (Kharitonov et al., 2019) which is designed to train emergent-language agents from language games. We use the emergent language setting to evaluate current model architecture's ability to learn to communicate while restricting certain contextual properties, like viewing scenes from different POVs. Given we allow the agent's to create any language it is important that we design the task in such a way that they actually have to solve the intended task.

We intend to answer the following questions:

1. Can current model architectures learn to communicate with differing POVs
2. Can we improve models' ability to learn through special pre-training
3. Given contextual priming, do the emergent languages show properties of human language

After these initial experiments we want to see if we can transfer these learnings to models which use human language. We can do this by generating labels for our underlying data.

3 Related Work

Spatial Relations have been studied on without FoR e.g. Cheng et al. (2024); Kelleher and Dobnik (2017); Fu et al. (2024); Liu et al. (2023); Kuhnle and Copestake (2017); Kordjamshidi et al. (2011). Liu et al. (2023) allow annotators to use camera or intrinsic FoR but do not model them explicitly. Lee et al. (2022) model intrinsic FoR, e.g. "plane left of elephant" from the elephant's FoR. This is complementary to our data which poses different challenges to models. Steels and Loetzsch (2006) have robots view events from different perspectives and perform a language game, creating a similar scenario to ours, however, their model architectures are quite out of date so we are due a new look at the problem. Fu et al. (2024) propose several visual benchmarks for visual language models, one is multi-view reasoning, however the task is simply to identify how the camera has moved (left or right) with no spatial reference task. Dobnik et al. (2020) gather dialogues with spatial descriptions from different FoR, however, the number of dialogues is too small to train modern models on and the task is more complex, this proposed data is a first step towards solving this more complex task.

Acknowledgments

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language model. *ArXiv*, abs/2406.01584.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik, John D. Kelleher, and C. Howes. 2020. Local alignment of frame of reference assignment in english and swedish dialogue. In *Spatial Cognition*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *ArXiv*, abs/2404.12390.

Hua Hua, Jochen Renz, and X. Ge. 2018. Qualitative representation and reasoning over direction relations across different frames of reference. In *International Conference on Principles of Knowledge Representation and Reasoning*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12 –13 June*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52, Gothenburg, Sweden. Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, CLASP, Centre for Language and Studies in Probability.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. Egg: a toolkit for research on emergence of language in games. *ArXiv*, abs/1907.00852.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):4:1–4:36.

Alexander Kuhnle and Ann A. Copestake. 2017. Shapeworld - a new test methodology for multimodal language understanding. *ArXiv*, abs/1704.04517.

Jae Hee Lee, Matthias Kerzel, Kyra Ahrens, Cornelius Weber, and Stefan Wermter. 2022. What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning. In *International Joint Conference on Artificial Intelligence*.

Stephen C. Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press, Cambridge.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Luc L. Steels and Martin Loetzsch. 2006. Perspective alignment in spatial language. In *Spatial Language and Dialogue*.

Discourse markers for topic change

Paola Herreno Castaneda and Mathilde Dargnat
CNRS, Université de Lorraine
ATILF
paola.herreno-castaneda@univ-lorraine.fr
mathilde.dargnat@univ-lorraine.fr

Jonathan Ginzburg
CNRS, Université Paris-Cité
Laboratoire de Linguistique Formelle
yonatan.ginzburg@u-paris.fr

1 Introduction

In conversation, new utterances can address a topic distinct from the current discourse topic (henceforth DT) (or Question Under Discussion (QUD), see Ginzburg, 2012; Roberts, 2012). This phenomenon is usually called *topic change* (henceforth TC). TC is a pivotal issue for both spoken dialogue and written text, but it is quite tricky to offer an account especially for theories derived from text analysis such as RST (Mann and Thompson, 1987; Taboada and Mann, 2006) or SDRT (Asher and Lascarides, 2003). In these frameworks, TC is not *a priori* specified as a genuine discourse relation like Cause or Contrast.[1] However, in some studies on connectives (Roze, 2009; Roze et al., 2012), it is possible to find some items associated with a Digression or a Detachment relation.

In this paper, we focus on the question of TC markers in dialogue, especially with respect to three French Discourse Markers (DMs): *à propos*, *au fait*, and *d'ailleurs*. We conduct an empirical study highlighting three of their properties: anaphoric status, position in the utterance, and the permanent/temporary nature of TC. This was the starting point for what we believe is the first formal account of French topic change DMs (TCDMs).

2 Semantic properties of French TCDMs

The literature on DMs is considerable and there is much discussion regarding which properties to take into account to define MD as a category. Nevertheless, some features are generally agreed on[2] The main functions of DMs can be categorized on the basis of three properties: discourse structuring, manifestation of the speaker's knowledge or emotions, and interaction management. In many discourse theories, DMs are associated with discourse relations which can be, for instance, of a temporal or logical type. Here, we are interested in the discourse relations that the three DMs under investigation can encode.

À propos has been characterised as an enunciative connective (Prévost, 2011; Richard and Drouet, 2016) or as a rupture(-inducing) enunciative marker (Molinier, 2003). Pretheoretically, its meaning indicates a change of DT—introducing a new topic in a way that ensures discourse continuity and avoids an abrupt topic transition.

Au fait as a DM is relatively infrequent in our data, but it seems to work in a similar way to *à propos*. It can be placed in an elaborative context, taking a previous element and elaborating on it (Defour et al., 2010). It also occurs with interrogative structures that can be used as a starting point for a new topic (D'Hondt, 2014).

D'ailleurs is often studied in the argumentation field. In the polyphony-based approach developed by Ducrot et al. (1980), it is characterized by the formula $r : P$ *d'ailleurs* Q, where r is a conclusion for which P and Q are arguments; Q is presented as a non-necessary argument for r, a sort of side comment. As a DM, it introduces a complementary and independent argument or even a digressive comment. It is also seen as possessing an additive function, where the elements it adds are detached from the DT, capable of creating a textual discontinuity or a potential TC (Baider, 2018).

3 Empirical results

In the CODIM project framework, a collection of French corpora was compiled. Here, we use spoken data from six corpora CFPP, CLAPI, ESLO, FRA80, MPF, and TCOF (for details, see the references section). We hoped to find 25 examples for each DM in each corpus, but ended up with some-

[1] Asher (2004) does mention in passing that some discourse markers, like *sinon*, *bon* or *au fait*, can trigger a topic shift.

[2] For recent syntheses see a.o., Anscombre et al., 2013; Brinton, 2017; Heine et al., 2021; Dargnat, 2023; Hansen and Visconti, 2024.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

what less—238 tokens in total: 150 for *d'ailleurs*, 75 for *au fait*, and 13 for *à propos*.

In order to analyse the anaphoric cues, we distinguish several kinds of semantic relations between utterances that host the DM and previous ones. While explicit anaphora do of course occur, we also find other types of semantic relations such as hypernymy, hyponymy, meronymy, etc.

Regarding the TC span, the possibilities are quite diverse. Some examples show that the new topic continuation span only covers the host utterance (mostly *d'ailleurs*), while in other cases the span is much larger (*à propos* and *au fait*).

We observe that the syntactic position of the DM can lead to differences in meaning or function. When the DM appears at the beginning, its function seems to involve initiating an illocutionary act by capturing the participant's attention, followed by a short-term digression, side comment, or TC, depending on the nature of the DM. However, when they are placed at the end, their use seems to be less essential and more focused on yielding the turn or just closing an idea whose nature is parenthetical (see Koev, 2022).

4 Formalization

Our goal is to propose a formal account of French TCDMs. We follow the basic approach to dialogue coherence detailed in Ginzburg et al. (2022), in particular the characterization of responses that effect TC. Ginzburg et al. (2022) point out that for ChangeTopic the simplest analysis would involve allowing a response specific to an arbitrary question. The obvious problem this would raise is massive ambiguity since many responses from other classes would be analyzable in such terms. To avoid that, Ginzburg et al. introduce the additional restriction that an *irrelevance* relation (IrRel) (Ginzburg, 2012) should hold between the topic changing utterance and the Dialogue GameBoard, creating a lack of coherence with the current context. What would this amount to? Being neither QSpecific concerning q_1 (i.e., a partial answer to or sub-question of q_1) uttered by A to B, nor being co-propositional with a clarification question generated by q_1's utterance[3] nor QSpecific with respect to the issues ?WishDiscuss(B,q1) or λxKnowAnswer(x,q1) (implicated in metadiscursive and metaepistemic responses, respectively.).

We formulate the basic form of a TCDM in (1a). We treat such DMs as dialogue move indicators, whose force is that the next utterance will involve a new topic. We think that the force of such DMs is not abrupt—here we allow for arbitrary non-related issues, but this might be too permissive —we leave a more empirically based decision to future work. Nonetheless, at least in the case of permanent topic change markers like *au fait* and *a propos*, a downdate is required of the previous MaxQUD, which the speaker feels is exhausted; this is not the case for *d'ailleurs*, so the force it signals will be somewhat different. We assume for now that such DMs are compatible with various kinds of moves (assertions, queries, commands), but we could make the specification more restricted if the data suggested the need. The specification that the DM's complement is verbal and I(ndependent)C(lause):+ means such DMs select for matrix clauses. By specifying information about the LatestMove, we capture the apparent generalization that exophoric triggers are incompatible with such DMs, as shown by our data. The force of such an utterance is $ChangeTopic$, explicated in (1b), with two subcases (i) permanent topic change (applicable to *à propos* and *au fait*) and (ii) temporary topic change (applicable to *d'ailleurs*). In both cases an utterance concerning a question irrelevant to the previous dialogue gameboard is licensed. In case (i) the previous MaxQUD is downdated, whereas in case (ii) both the new and the old issues are maintained as maximal in QUD.

(1) a.
$$\left[\begin{array}{l}\text{arg-struc}: \left\langle \left[\begin{array}{l}\text{cat}: \text{v}\\ \text{IC}: +\\ \text{cont} = \text{p1}: \text{IllocProp}\end{array}\right] \right\rangle \\ \text{dgb-params}: \left[\begin{array}{l}\text{spkr}: \text{Ind}\\ \text{addr}: \text{Ind}\\ \text{t}: \text{TIME}\\ \text{c0}: \text{addressing(spkr,addr,t)}\\ \text{LatestMove.cont}: \text{IllocProp}\\ \text{MaxQUD=q}: \text{Question}\end{array}\right] \\ \text{cont} = \text{ChangeTopic(spkr,p1,q)}: \text{IllocProp}\end{array}\right]$$

b. **Permanent/Temporary ChangeTopic**

$$\left[\begin{array}{l}\text{pre}: \left[\text{QUD} = \left\langle \text{q1, Q} \right\rangle : \text{poset(Question)}\right] \\ \text{effects}: \left[\begin{array}{l}\text{spkr} = \text{pre.addr}: \text{Ind}\\ \text{addr} = \text{pre.spkr}: \text{Ind}\\ \text{r}: \text{Question} \vee \text{Prop}\\ \text{q2}: \text{Question}\\ \text{R}: \text{IllocRel}\\ \text{Moves} = \left\langle \text{R(spkr,addr,r)} \right\rangle \oplus \\ \text{pre.Moves}: \text{list(LocProp)}\\ \text{c1}: \text{Qspecific(R(spkr,addr,r),q2)}\\ \text{QUD} = \text{(i)} \left\langle \text{q2, Q} \right\rangle \\ \text{(ii)} \left\langle \begin{array}{l}\text{Max} = \left\{ \text{q2,q1} \right\}, \\ \text{Q}\end{array} \right\rangle : \\ \text{poset(Question)}\\ \text{c2}: \text{IrRel(q2,pre)}\end{array}\right]\end{array}\right]$$

[3]Here *CoPropositionality*, for two questions means that, modulo their domain, the questions involve similar answers: for instance 'Whether Bo left', 'Who left', and 'Which student left' (assuming Bo is a student.) are all co-propositional.

Corpora

CFPP: Corpus de Français Parlé Parisien. http://cfpp2000.univ-paris3.fr
CLAPI: Corpus de Langue Parlée en Interaction. http://clapi.ish-lyon.cnrs.fr
CRFP: Corpus de Référence du Français Parlé. http://www.up.univ-mrs.fr/delic/corpus/index.html
ESLO: Enquêtes Sociolinguistiques à Orléans. http://eslo.huma-num.fr
FRA80: Corpus de Français des Années 80. CREDIF, ENS de Saint-Cloud.
MPF: Multicultural Paris French. https://www.ortolang.fr/market/corpora/mpf
Scientext: https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/scientext
TCOF: Traitement de Corpus Oraux en Français. http://www.cnrtl.fr/corpus/tcof
Wikiconflits: see (Poudat et al., 2017)

Acknowledgments

We acknowledge the financial support of the ANR CODIM Project (https://www.codim-project.org/) and also of the French *Investissements d'Avenir-Labex EFL* program (ANR-10-LABX-00). We thank the reviewers of TrentoLogue for helpful comments.

References

Jean-Claude Anscombre, Maria Luisa Donaire, and Pierre Patrick Haillet. 2013. *Opérateurs discursifs du français*. Peter Lang Verlag, Lausanne, Suisse.

Nicholas Asher. 2004. Discourse topic. *Theoretical Linguistics*, 2-3(30):163–201.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Fabienne Baider. 2018. *Par ailleurs* et *d'ailleurs*: marqueurs linguistiques de « rupture » textuelle ou marqueurs de continuation argumentative ? In *6e Congrès Mondial de Linguistique Française*, volume 46, pages 01–14. EDP Sciences.

Laurel J. Brinton. 2017. *The Evolution of Pragmatic Markers in English: Pathways of Change*. Cambridge University Press.

Mathilde Dargnat. 2023. *Lexique et discours*, synthèse d'HDR. Université Paris 8 Vincennes Saint-Denis.

Tine Defour, Ulrique D'Hondt, Anne-Marie Simon-Vandenbergen, and Dominique Willems. 2010. In fact, en fait, de fait, au fait: A contrastive study of the synchronic correspondences and diachronic development of english and french cognates. *Neuphilologische Mitteilungen*, pages 433–463.

Ulrique D'Hondt. 2014. Au fait, de fait et en fait: analyse de trois parcours de grammaticalisation. *Revue Romane. Langue et littérature. International Journal of Romance Languages and Literatures*, 49(2):235–263.

Oswald Ducrot, Danière Bourcier, Sylvie Bruxelles, and [...]. 1980. *Les mots du discours*. Editions de Minuit.

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

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.

Maj-Britt M. Hansen and Jacqueline Visconti, editors. 2024. *Manual of Discourse Markers in Romance*. De Gruyter, Berlin.

Bernd Heine, Gunther Kaltenböck, Tania Kuteva, and Haiping Long. 2021. *The Rise of Discourse Markers*. Cambridge University Press.

T. Koev. 2022. *Parenthetical Meaning*. Oxford studies in semantics and pragmatics. Oxford University Press.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Christian Molinier. 2003. Connecteurs et marqueurs énonciatifs: Les compléments figés formés à partir du nom *propos*. *Lingvisticae Investigationes*, 26(1):15–31.

Céline Poudat, Natalia Grabar, Camille Paoloque-Berges, Thierry Chanier, and Jin Kun. 2017. Wikiconflits : un corpus de discussions éditoriales conflictuelles du wikipédia francophone. In Ciara R. Wigham and Gudrun Ledegen, editors, *Corpus de communication médiée par les réseaux, Construction, structuration, analyse*, pages 19–36. L'Harmattan.

Sophie Prévost. 2011. *A propos* from verbal complement to discourse marker: a case of grammaticalization? *Linguistics*, 49(2):391–413.

Élisabeth Richard and Griselda Drouet. 2016. Confirmer pour mieux détourner : marqueurs d'acceptation et modalités de transition. *Testi e linguaggi*, pages 131–141.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Charlotte Roze. 2009. Base lexicale de connecteurs du français.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. Lexconn: A french lexicon of discourse connectives. *Discours*, 10:01–13.

Maite Taboada and William C. Mann. 2006. Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

The Linguistic Interpretation of Non-emblematic Gestures Must be agreed in Dialogue: Combining Perceptual Classifiers and Grounding/Clarification Mechanisms

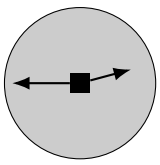
Andy Lücking and **Alexander Mehler** and **Alexander Henlein**
Goethe University Frankfurt, Text Technology Lab
{luecking,mehler,henlein}@em.uni-frankfurt.de

1 Introduction

Non-emblematic manual gestures pose a double challenge for semantic theories: Firstly, gestures are instances of *visual communication*, so their interpretation requires a means of perceptual classification. Secondly, according to gesture studies, non-emblematic gestures lack "standards of form" (McNeill, 1992, p. 22). In other words, there is no lexicon of such gestures (as opposed to emblematic ones). Accordingly, the linguistic interpretation of gestures – that is, the classification of a gesture occurrence by means of verbal labels from a natural language – leaves room for interpretation. If this room for interpretation is to be resolved, it must be negotiated in dialogue ("What does the speaker/gesturer mean by the gesture?"). Therefore, the linguistic meaning of non-emblematic gestures, if unclear or important for the understanding of the utterance, must be agreed in dialogue.

2 Perceptual Classification and CVM

Following formal semantics work on spatial language (Zwarts, 1997, 2003) and the psychophysics of biological movement (Johansson, 1973; Johansson et al., 1980), a uniform, imagistic extension of semantic models, respectively the lexical semantics of certain predicates, is accomplished in terms of vector sequences. For instance, the spatial preposition *near* has the vector denotation in (1) (Zwarts, 2003). The reference object is represented by the black rectangle, the two arrows indicate two of the vectors from the denotation (the gray area; boundaries should be fuzzy, of course).

(1) $[\![\text{near}]\!](\blacksquare) =$ 

Johansson and colleagues showed that the perception of dynamic events, that is, events that involve motion, can be modeled in terms of vectorial representations, too. The vector-based representations provide useful explications of the visual components of lexical items, dubbed *conceptual vector meaning* (CVM) (Lücking, 2013). CVMs are also candidates for explicating *what* a perceptual classifier actually has learned.

Larsson (2015, 2020) makes classification the core of meaning so that the type *Meaning* (*Mng*) of a lexical entry involves a classifier (clfr) from the outset. As Larsson (2020) emphasizes, classifiers provide a computational spell out of (perceptual) *judgments* in TTR: a situation s is of type T, $s:T$, if the classifier associated with T returns T when applied to s (i.e., $[\![\text{T}]\!].\text{clfr}(\text{par},s) = T$). As is known from human vision, people classify objects and events by comparing a visual percept with stored image (Ullman, 1996, §6). CVMs are representations of stored images, so we add them to *Meaning*:

(2) $$Mng := \begin{bmatrix} \text{par} & : Rec \\ \text{cvm} & : Type \\ \text{bg} & : RecType \\ \text{fg} & : \text{bg} \rightarrow RecType \\ \text{clfr} & : \text{par} \rightarrow \text{bg} \rightarrow \text{cvm} \rightarrow RecType \end{bmatrix}$$

The classifier in (2) now involves an additional layer of computation, namely a geometric comparison G of the percept (from 'par $\rightarrow$ bg') with the value of 'cvm'.[1] Let us illustrate this with the simple example of *near*. Using *near*'s CVM from (1), the meaning of *near* can be expressed as follows, where, following Zwarts (2003), $\text{place}(\mathbf{v},x)$ denotes a vector emanating from object x:

[1] Ideally, there also should be a feedback loop such that each successful or unsuccessful classification updates (confirms or modifies) *Mng*.cvm.

(3) $[\![\text{near}]\!] =$

$$\begin{bmatrix} \text{par} : \textit{Rec} \\ \text{cvm}=\left\{ \mathbf{v} \mid \text{place}(\mathbf{v}, \text{bg.x}) \right\} : \textit{Type} \\ \text{bg}=\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{v} : \textit{Vec} \\ \text{l} : \mathbb{R} \end{bmatrix} : \textit{RecType} \\ \text{fg} : \text{bg} \rightarrow \text{near(bg.x)} \\ \text{clfr} : \text{par} \rightarrow \text{bg} \rightarrow \text{cvm} \rightarrow \textit{RecType} \end{bmatrix}$$

The classifier for *near*, $[\![\text{near}]\!]$.clfr, applies to situations r involving an individual and a vector of a certain length 'l':

(4) $r = \begin{bmatrix} \text{x} & : \textit{Ind} \\ \text{v} & : \textit{Vec} \\ \text{l}=||\mathbf{u}|| & : \mathbb{R} \end{bmatrix}$

(5) $[\![\text{near}]\!]$.clfr(par,cvm,r) =

$$\begin{cases} \text{near}(r.\text{x}) & \text{if } G[(r.\text{l} \cdot \text{par.w}), \text{cvm}] > \text{par.t} \\ \neg\text{near}(r.\text{x}) & \text{else} \end{cases}$$

G is an algorithm from computational geometry (Sack and Urrutia, 2000), which compares the weighted input of situation r with the stored CVM information. In this case, G just has to perform a distance calculation.

3 Speech–Gesture Monitoring in Dialogue

The default integration of speech and gesture – namely that a gesture g directly exemplifies it affiliate – can now be expressed as follows:

(6) *Affiliation Default*
$[\![\ulcorner \textit{affiliate} \urcorner]\!]$.clfr(par,cvm,$\pi_v(g)$) $\mapsto$ $\ulcorner$affiliate$\urcorner$

That is, a vectorized gesture movement figures as the background situation onto which the classifier associated with the gesture's affiliate in speech applies. This immediately gives rise to a notion of speech–gesture mismatch, or inconsistency:

(7) *Speech–Gesture Mismatch*
If $[\![\ulcorner \textit{affiliate} \urcorner]\!]$.clfr(par,cvm,$\pi_v(g)$) $\not\mapsto$ $\ulcorner$affiliate$\urcorner$, an inconsistency between speech and gesture g occurs.

We note again that (7) is a simplification, since gestures that attach to frame elements that are associated with the surface affiliate expression are not taken into account. Apart from this simplification, a mismatch according to (7) can trigger multimodal clarification interaction (Ginzburg and Lücking, 2021).

Example (8), taken from Lücking et al. (2024), is constructed following SaGA dialogue V10, 3:19 (Lücking et al., 2010) where R talks about staircases and makes a spiral gesture (8-a). Then F poses the verbal clarification request whether the linguistic interpretation of R's multimodal utterance is the hyponym "spiral staircase" (8-b), which can be confirmed or rejected (8-c).

(8) a. R: Inside the hall was an imposing staircase. 

b. F: Do you mean a spiral staircase?

c. R: Yes/No.

The spiral gesture from example (8) does not directly match $[\![\text{staircase}]\!]$.clfr, but it does correspond to $[\![\text{spiral-staircase}]\!]$.clfr. This raises the issue $q0$ = "?Mean(R,$u0$,'spiral staircases')" as F's MaxQUD, where $u0$ is the multimodal sub-utterance consisting of the noun *staircases* and the wounded gesture. Parameter Identification is triggered, leading to F's clarification question, which is co-propositional to $q0$.

4 Conclusions

We formally defined speech–gesture congruence and mismatch, in particular the latter underlies multimodal clarification interaction. The sample analysis shows a sometimes intricate interaction of QUD accommodation and perceptual gesture classification, mechanisms which call for further exploration in future work. A couple of processing predictions of our model can already be derived, however, including the following ones.

- The ease of the linguistic interpretation of a gesture depends on the degree of conventionalization (strength) between lexemes and their associated CVMs.
- The linguistic interpretation becomes more difficult when the gesture gives rise to a vectorial model that is not lexicalized in terms of a CVM.

Acknowledgement

Support by the *Deutsche Forschungsgemeinschaft* (DFG), grant number 502018965, is gratefully acknowledged.

References

Jonathan Ginzburg and Andy Lücking. 2021. Requesting clarifications with speech and gestures. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations*, MMSR, pages 21–31. Association for Computational Linguistics.

Gunnar Johansson. 1973. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211.

Gunnar Johansson, Claes von Hofsten, and Gunnar Jansson. 1980. Event perception. *Annual Review of Psychology*, 31:27–63.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Staffan Larsson. 2020. Discrete and probabilistic classifier-based semantics. In *Proceedings of the Probability and Meaning Conference*, PaM 2020, pages 62–68, Gothenburg. Association for Computational Linguistics.

Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter, Berlin and Boston.

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.

Andy Lücking, Alexander Henlein, and Alexander Mehler. 2024. Iconic gesture semantics. *Preprint*, arXiv:2404.18708.

David McNeill. 1992. *Hand and Mind – What Gestures Reveal about Thought*. Chicago University Press, Chicago.

Jörg-Rüdiger Sack and Jorge Urrutia. 2000. *Handbook of computational geometry*. Elsevier, Amsterdam, The Netherlands.

Shimon Ullman. 1996. *High-Level Vision*. A Bradford Book. MIT Press, Cambridge, MA.

Joost Zwarts. 1997. Vectors as relative positions: A compositional semantics of modified PPs. *Journal of Semantics*, 14(1):57–86.

Joost Zwarts. 2003. Vectors across spatial domains: From place to size, orientation, shape, and parts. In Emile van der Zee and John Slack, editors, *Representing Direction in Language and Space*, pages 39–68. Oxford University Press, Oxford, NY.

Every quantifier scope ambiguity is enabled by a context

David Pagmar
University of Gothenburg
david.pagmar@gu.se

Asad Sayeed
University of Gothenburg
asad.sayeed@gu.se

1 Introduction

The sentence "every road leads to a town" entails a *quantifier scope ambiguity* (QSA; Kurtzman and MacDonald, 1993; Dotlačil and Brasoveanu, 2015), i.e., there is either one town (singular) or different towns (plural). A pilot study shows 16 of 20 Swedish speakers make a plural interpretation of this sentence.

This study concerns the competitive effects of syntactic factors and the larger pragmatic context on QSA resolution in Swedish. Most previous work is in English[1]. Experimental studies can shed light on the real-time mechanisms involved in QSA resolution.

Ambiguities are common in every day language use (Koller et al., 2010). The role of context in ambiguity resolution is more or less a linguistic truism (Mey, 2003); a language user uses contextual factors to interpret ambiguous sentences. Much of the previous work has focused on manipulations of the order of the quantifiers (Kurtzman and MacDonald, 1993). Less attention has been given to the effect of a previously described scenario serving as a context for the experiment participant to interpret the ambiguous target sentence; that is, whether prior context can overcome biases[2].

QSA resolution can be sensitive to syntactic variation (Sayeed et al., 2019). The syntactic factor explored in this study is grammatical gender. The final word of the QSA sentence (*every road leads to a town*) is marked for indefiniteness. The Swedish indefiniteness markers are the articles *en* or *ett*, which correspond to the two grammatical genders UTRUM and NEUTRUM. Both articles are also the number words for *one* (1). The neutrum form *ett* has stronger numerical qualities.

The QSA that is investigated in this study stems from the quantifier *varje* (*every*, in English). A previous study examining the neurological foundations of quantifier interpretation has found that quantifiers activate areas of the brain associated with numeracy (McMillan et al., 2005). This finding suggests a cognitive basis for the interpretation of quantifiers that could extend to grammatical markers.

[1]Exceptions exist, such as Sayeed et al. (2019) and Radó and Bott (2018) for German or Scontras et al. (2014) for Chinese.

[2]One exception is Villalta (2003), who manipulated the order of information presentation in a larger contextual scenario before testing the interpretation of *how many* questions with scope ambiguities. Her manipulation was not focused on the lexical-pragmatic aspects of the scenario as in our study.

We pose the following two questions: (1) can the interpretation of QSA be controlled by non-determinative contextual information, and if so, to what degree?, and (2) is the interpretation of QSA in Swedish affected by the grammatical gender of the indefinite noun?

We expect that plural contexts will prompt more plural readings of a QSA sentence and vice versa. We also expect that plural contexts with the NEUTRUM gender will show a greater number of singular readings, compared to plural contexts with the UTRUM gender, due to the additional role of the NEUTRUM marker as the cardinal number one.

2 Method

A total of 28 Swedish speaking participants took part in the experiment. All had Swedish as their first language and were above the age of 18.

2.1 Stimuli

The experiment was a forced choice judgement task (20 critical tirals and 10 distractors) via the online platform Pavlovia. Participant were asked to imagine that the following two sentences were spoken by a friend. First there was a contextual sentence, establishing connotations to either a singular or plural reading of the critical sentence, and then a critical sentence with the structure: *varje turist såg en X* (=*every tourist saw an X*). Each critical sentence had two versions, with either UTRUM or NEUTRUM gender. A final question followed the critial sentence: *What do you assume your friend means, did every tourist see the same X?*

2.2 Data analysis

We conducted a multilevel logistic regression analysis, with random intercepts for participants, using the *glmer* function of the *lme4* package (version 1.1-35.3) in R. The analysis follows the equation:

$$\text{logit}(P(y_{ij} = 1)) = \gamma_{00} + \beta_1 \text{SIN}_{ij} + \beta_2 \text{UTR}_{ij} + u_{0j}$$

γ_{00} is the fixed intercept (overall average intercept). β_1 is the fixed effect (slope) for the context predictor (SIN). β_2 is the fixed effect (slope) for the gender predictor (UTR). u_{0j} is the random intercept for participant *j*, representing the participant-specific deviation from the overall intercept.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

In the analysis, every trial is analyzed as an individual observation (N=560). The binary dependent variable is the QSA reading (plural = 0, singular = 1) and the predictors are the two conditions contextual sentences and grammatical gender. The results are presented as odds ratios (OR; Szumilas, 2010).

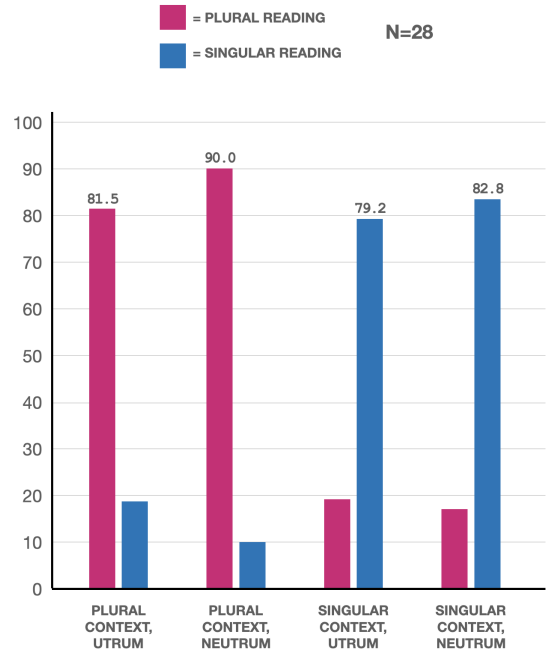


Figure 1: Percentages of QSA interpretations for each combination of conditions. Contextual condition provides substantial effects on QSA interpretations.

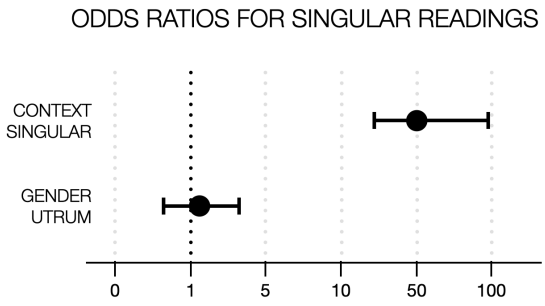


Figure 2: The Odds Ratios result for singular readings of QSA sentences show an sizable effect of contextual sentences, OR = 50.20, 95%CI[24.32, 103.65]. The results show no reliable effect for grammatical gender, OR = 1.26, 95%CI[.69, 2.31].

	Est.	St.Er	z	*p*
Intercept	-2.19	.36	-6.05	<.001
contextSIN	3.91	.36	10.58	<.001
genderUTR	.23	.30	.76	.43

Table 1: The results from the multilevel logistics regression analysis. Binary pridctors and the results list effects of singular conditions for contextual information (contextSIN) utrum conditions for grammatical gender (genderUTR).

3 Results

The results show that every critical sentence was subject to both singular and plural readings. Results are shown in Figure 2. The multilevel logistic regression analysis showed an effect of context condition, but not of grammatical gender (Table 1).

The singular readings per participant had a mean of 9.40 and a SD of 3.58. 9 participants had 10 singular readings during the experiment, which is equal to the total amount of singular contexts. 2 participants had 2 singular readings during the experiment. 1 participant had 19 singular readings during the experiment.

4 Discussion

The results show a substantial effect of contextual information on participants' interpretations, affirming our hypothesis. The multilevel logisitc regressions analysis show significant between-participant variability in the baseline log-odds of singular readings, as indicated by the variance of the random intercepts (.73). We see variation in the data. Among the 20 critical trials, two participants made 18 plural readings, while another participant made 19 singular readings. Given the relatively modest sample size of the current study, there is reason to speculate that some individuals might consistently favor one type of reading across all trials.

These results provide an insight into the intricate nature of ambiguity resolution; it shows that language users that exhibit strong preferences for QSA interpretations still deviates from their preference given certain contexts.

The grammatical gender conditions did not show a reliable effect in the logistic regression analysis. A detectable trend goes in the opposite direction of the hypothesis. One potential explination for this trend could be the additional cognitive load created by this potential indecision about whether to introduce a new discourse referent instead prompts the processor to rely more on contextual cues, favoring the plural interpretation. This would be in line with previous findings Dwivedi (2013).

One direction for future experimental research is to focus on how QSA interpretation relates to e.g. lexical ambiguity, while simultaneously taking measures to increase the ecological validity of the experimental tasks.

The implication of this work for research into computational representations of language interaction is that there is a fine-grained connection between the "nitty-gritty" of the syntax-semantic interface, lexical-pragmatic knowledge, and the immediate context.

References

Jakub Dotlačil and Adrian Brasoveanu. 2015. The manner and time course of updating quantifier scope representations in discourse. *Language, Cognition and Neuroscience*, 30(3):305–323.

Veena D Dwivedi. 2013. Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, 8(11):e81461.

Alexander Koller, Stefan Thater, and Manfred Pinkal. 2010. Scope underspecification with tree descriptions: Theory and practice. *Resource-Adaptive Cognitive Processes*, pages 337–364.

Howard S Kurtzman and Maryellen C MacDonald. 1993. Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279.

Corey T McMillan, Robin Clark, Peachie Moore, Christian Devita, and Murray Grossman. 2005. Neural basis for generalized quantifier comprehension. *Neuropsychologia*, 43(12):1729–1737.

Jacob L Mey. 2003. Context and (dis) ambiguity: a pragmatic view. *Journal of Pragmatics*, 35(3):331–347.

Janina Radó and Oliver Bott. 2018. What do speaker judgments tell us about theories of quantifier scope in German? *Glossa: a journal of general linguistics*, 3(1).

Asad Sayeed, Matthias Lindemann, and Vera Demberg. 2019. Verb-second effect on quantifier scope interpretation. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 134–139.

G. Scontras, M. Polinsky, E. C.-Y. Tsai, and K. Mai. 2014. Chinese scope: an experimental investigation. In *Proceedings of Sinn und Bedeutung 18*, Vitoria-Gasteiz, Spain.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.

Elisabeth Villalta. 2003. The role of context in the resolution of quantifier scope ambiguities. *Journal of Semantics*, 20(2):115–162.

Annotation Needs for Referring Expressions in Pair-Programming Dialogue

Cecilia Domingo, Paul Piwek, Michel Wermelinger
The Open University
Milton Keynes, England
cecilia.domingo-merino@open.ac.uk
paul.piwek@open.ac.uk
michel.wermelinger@open.ac.uk

Svetlana Stoyanchev
Toshiba Europe Ltd.
Cambridge, England
svetlana.stoyanchev@toshiba.eu

Abstract

Referring expressions are a widely researched area in linguistics, with some work also on dialogue. As dialogue technology evolves and language models increasingly incorporate programming languages, pair-programming dialogues become a promising area of research. We recorded 24 dialogues between adult learners to analyse how they refer to the code that they create through the task. These dialogues present some challenges (and interesting avenues for research) for annotation and analysis, due to several factors: high multimodality, mix of abstract and specific entities in the same domain, variability in naming practices, and referents being sometimes located in the future or remaining hypothetical.

1 Introduction

Referring phenomena have long been studied in computational linguistics. While much focus was placed on monologic text, a lot of research has also been carried out in dialogue, though often in simulated settings. In our work we gather data on a real educational task solved by learners through dialogue: Python programming problems solved via pair-programming[1]. Through our data we hope to shed light on how speakers discuss code entities, in a domain whose importance is increasing in NLP as models become better at processing code along with natural language (Wan et al., 2024).

2 Remote pair-programming dialogues

We focus our research in the domain of pair programming, due to its pedagogical value (Hanks et al., 2011) and relevance in computational linguistics as language models develop increasing code capabilities (Wan et al., 2024). Even though this domain has been widely researched, most studies focus on the code produced rather than the dialogue, and thus little dialogue data is available. We recorded 24 pair-programming sessions in a remote setting (5 of them, the pilot, with simulated remoteness, i.e., participants in adjoining rooms). Participants communicated via voice call and worked together on beginner-level Python tasks using Visual Studio Code with the Live Share plugin, which allows simultaneous editing of the code file, with each user connected having their own cursor. The sessions were around 30 minutes each. The participants were 45 students (from Bachelor to PhD) and 2 staff, ages ranging from 23 to 70, and a gender split of 12 female/35 male. We recorded the dialogues, as well as the code produced at each moment and participants' mouse and keyboard activity; we also recorded the screen for additional context.

[1]Pair programming is a collaboration technique where two people work together, simultaneously, on the same piece of code.

3 Annotation needs

Several annotation schemes exist for referring expressions, most famously Ontonotes (Weischedel et al., 2013). This scheme's main appeal is its simplicity, which leads to high inter-annotator agreement, but it is also its main point of criticism, as it fails to capture phenomena that may be important (Zeldes, 2022). Other schemes (Poesio et al., 2024) tackle some of these limitations, offering useful descriptions of complex types of referring expressions. Such extensive analysis of anaphora characteristics might not be advisable in our domain, where not even a more basic one has been carried out yet. Moreover, efforts in linguistic analysis might be best spent on other aspects where our domain is more unique, such as the characteristics of the names given to code elements by coders, and how these evolve through dialogue.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

3.1 Multimodality

Remote dialogues may normally be an activity that's primarily linguistic (Clark, 2005); however, the addition of the co-creation of code turns it into a highly multimodal activity. Other schemes' focus on discourse makes them unfit for our setting, where the main goal is linking references and referents across modalities. As the code becomes arguably as important as the discourse, we need to annotate both anaphora and deictic references, linking together discourse elements through coreference chains, but also linking those chains to entities in code files. In pair-programming dialogue, referents and referring acts will not only be found within the discourse; speakers will also refer to entities in the code that they are creating, and may use the mouse and keyboard to bring them into focus.

Some studies have been carried out on the use of pointing gestures for referring and show that they play an important role: e.g., gestures can replace locative expressions (Kehler, 2000), or they can contribute to mutual disambiguation, i.e., the ambiguity in discourse can be cleared with the information from other modalities and vice versa (Kaiser et al., 2003). In some settings, pointing gestures accompanied only few utterances (e.g., 16% of utterances (Sluis et al., 2008), though this number is not insignificant). However, in our preliminary analysis of publicly available data[2] 51% of utterances that mentioned code (which were 55% of the total) featured the mouse pointer or keyboard playing a role in the referring expression — as far as the video data allowed us to see. It is also important to note that different speakers show different strategies in their use of pointing (Piwek, 2007); such variability might also be observable in the use of the mouse and keyboard as pointing devices. Moreover, the literature has mostly focused on hand pointing — we might expect different behaviours regarding other forms of pointing (e.g., with the cursor).

3.2 Abstract and unrealised entities

Several studies on referring expressions have been carried out in highly controlled settings where the entities mentioned and their features are previously known to the researchers who placed them in the setting (Koolen, 2013; Rubio-Fernández, 2024). In our setting, most entities do not exist until the speakers create them into the code — they begin with a blank canvas. As the entities of the dialogue are created through it and not known beforehand, we need to annotate their characteristics post factum, distinguishing between abstract discussions of code and mentions to specific entities in the speaker's code files. See the example utterances below (from different dialogues) where we contrast a reference to an array as an abstract programming entity with an array as a specific entity present in the code:

- **Abstract array:** I can't remember how you do an array.
- **Array in the code:** I think that, that does need to be kind of an array so that I think that does need to be in square brackets.

Yet another peculiarity of the pair-programming task is that, as the speakers must discuss the problem to reach an agreement on how to solve it, often entities may be mentioned only to later be discarded as the discussion brings forth better solutions. In some cases the entity might be preserved, but the speakers may still discuss it at length before finally implementing it, thus speaking about a concrete but unrealised entity. In the excerpt below, speakers A and B discuss a string that they are going to type, and they give it a name, but the string does not become realised in the code until turn 5, when they change the name to 'text':

1. **A:** Can we, uh, I don't know, define a, a string, maybe the, the so-cool string.
2. **B:** Uh... Yeah, that seems like a good place to start. And then we can kind of maybe try and split it up into the.
3. **A:** Yeah. Yeah. So should I start defining these, this string?
4. **B:** Yeah, sure. Sounds good.
5. **A:** Um. Uh, how should I, uh, call it? Uh... Just. Um, sentence. [B types 'text'] Oh, text. Yeah, text.

4 Conclusions

Pair-programming dialogues possess several characteristics that set them apart from other dialogue domains studied so far, and thus require custom tools for their annotation and analysis. It is a highly multimodal setting that requires linking discourse

[2] http://www.pairwith.us/tv

to another modality (code) and taking note of accompanying gestures in a less-studied form (pointing using the mouse and keyboard). As it is a dynamic environment (Kumar et al., 2022) where entities are created through the dialogue, we need to characterise those entities post factum, observing as well whether they can be currently linked to the code, refer to future code, or remain hypothetical.

Acknowledgments

This work has financial support from EPSRC Training Grant DTP 2020-2021 Open University and Toshiba Europe Limited. This research project was reviewed by, and received a favourable opinion from, our university's Human Research Ethics Committee.

References

Herbert H. Clark. 2005. *Using language*, 6. print edition. Cambridge University Press.

Brian Hanks, Sue Fitzgerald, Renée McCauley, Laurie Murphy, and Carol Zander. 2011. Pair programming in education: a literature review. *Computer Science Education*, 21(2):135–173.

Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. 2003. Mutual dissambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of The Fifth International Conference on Multimodal Interfaces (ICMI '03). Vancouver, BC. Canada*, pages 12–19.

Andrew Kehler. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the seventeenth national conference on artificial intelligence*, pages 685–690. The AAAI Press, Menlo Park, California.

Ruud Martinus Franciscus Koolen. 2013. Need I say more? On overspecification in definite referenc. Unpublished PhD Thesis, OCLC: 856996240.

Abhinav Kumar, Barbara Di Eugenio, Abari Bhattacharya, Jillian Aurisano, and Andrew Johnson. 2022. Reference resolution and context change in multimodal situated dialogue for exploring data visualizations. *Preprint*, arxiv:2209.02215 [cs].

Paul Piwek. 2007. Modality choice for generation of referring acts: Pointing versus describing. In *Proceedings of Workshop on Multimodal Output Generation (MOG 2007)*, pages 129–139.

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

Paula Rubio-Fernández. 2024. Referential efficiency as speaker-listener coordination. 2024 CORE Project Workshop.

Ielka van der Sluis, Paul Piwek, Albert Gatt, and Adrian Bangerter. 2008. Towards a balanced corpus of multimodal referring expressions in dialogue. In *Proceedings of the Symposium on Multimodal Output Generation*.

Yao Wan, Zhangqian Bi, Yang He, Jianguo Zhang, Hongyu Zhang, Yulei Sui, Guandong Xu, Hai Jin, and Philip Yu. 2024. Deep learning for code intelligence: Survey, benchmark and toolkit. *ACM Computing Surveys*, pages 1–39.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0.

Amir Zeldes. 2022. Opinion piece: Can we fix the scope for coreference?: Problems and solutions for benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

Network science highlights the emotional structure of counselling conversations simulated by Large Language Models and humans

Edoardo Sebastiano De Duro, Riccardo Improta and **Massimo Stella**
CogNosco Lab, Department of Psychology and Cognitive Science, UniTrento

Abstract

This study investigates CounseLLMe, a dataset comprising 400 simulated mental health counselling dialogues among two Large Language Models (LLMs). These conversations, each containing 20 exchanges, were conducted in both English (using OpenAI's GPT-3.5 and Claude-3's Haiku) and Italian (using Claude-3's Haiku and LLaMAntino). A professional psychotherapist assisted in fine-tuning the prompts for realism. The dialogues are compared against patterns in recently released and investigated human mental health conversations focused on depression. By investigating dialogues via the cognitive framework of textual forma mentis networks, we find that LLMs tend to stay positive even when debating depression. Furthermore, LLMs tend to become more verbose along conversations, but without creating syntactic/semantic networks of increasing complexity, i.e. degree assortativity and average shortest path length remain stationary despite increases in verbosity. We discuss this difference in view of relevant literature on rumination and mental navigation.

1 Introduction

Large Language Models (LLMs) are artificial intelligences (AIs) trained in reproducing human texts, one word after the other. LLMs' complexity stems from their cognitive skills, inherited from training over vast amounts of knowledge. Understanding how these AIs can behave in scenarios like mental health counselling is crucial (De Choudhury et al., 2023), given the wide variety of online and untested services having LLMs act as counsellors to human with mental distress. Psychological counselling is a field characterised paramount requirement for empathy, understanding, and accurate information dissemination. Traditional models may falter in providing the nuanced care necessary due to their inability to fully grasp the depth of human emotions and the subtleties of psychological distress (De Freitas et al., 2022). Hence, it is crucial to rigorously evaluate LLMs' knowledge frameworks and belief systems concerning this sensitive domain.

2 Motivation

Whereas traditional studies focus on human-computer interactions, CounseLLMe (De Duro et al., 2024) considers 2 LLMs conversing with each other in a virtual mental health counselling scenario. The aims of the dataset are to: (i) shed light on the models' ability to navigate discussions related to depression; (ii) reproduce LLMs' ability to mimic human emotions and syntactic associations during psychological counselling.

In CounseLLMe, one LLM adopts the role of a counsellor, while the other embodies the individual seeking help. In other words, at every turn, each model was instructed to play the role of a human patient affected by depressive symptoms conversing with another model that was, instead, prompted to act as an expert psychotherapist. We adopt a machine psychology perspective (Bertolazzi et al., 2023; Stella et al., 2023), bridging computer science and psychology frameworks within a complex systems approach, to study the AI agents as if they were humans, using their texts to extract insights relative to the ways these models represent and describe mental health dialogues.

3 Results

CounseLLMe is a dataset consisting of 400 conversations, introduced and investigated in a previous study (De Duro et al., 2024). These conversations - of 20 quips each - were generated either in English (using OpenAI's GPT 3.5 and Claude-3's Haiku) or Italian (with Claude-3's Haiku and LLamantino). We carefully selected prompts with the consultation of a human professional in psychotherapy. To investigate the resulting conversations, we

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

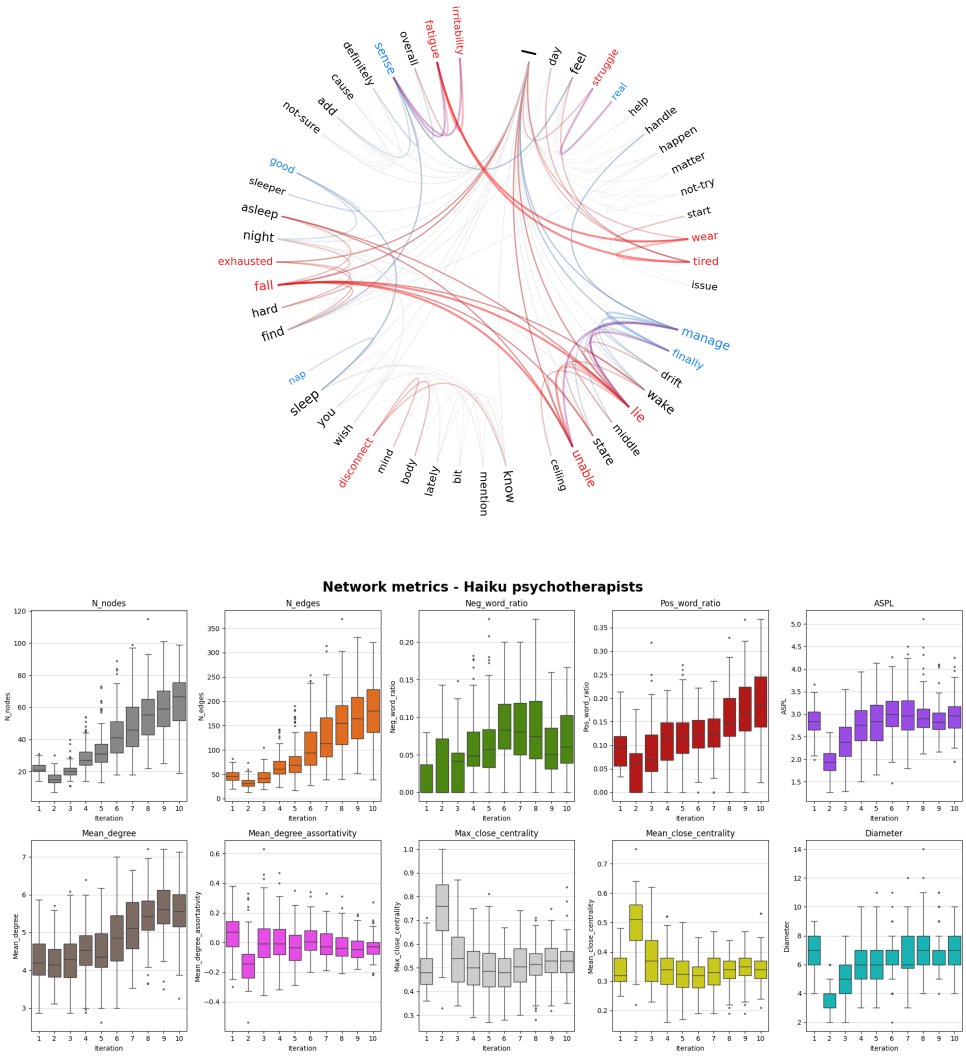


Figure 1: **Top:** A textual forma mentis network extracted from a quid of a patient (here Claude's Haiku) in the dataset. Positive (negative, neutral) words are highlighted in blue (red, black). **Bottom:** Sequential representation of various network measures extracted from the output of Claude's Haiku model (here playing the role of the psychotherapist).

employed the complex-systems technique of textual forma mentis networks (Stella, 2020), where nodes represent concepts and links indicate syntactic or semantic relationships between concepts in the dialogues' quips. Additionally, we performed some sequence-based techniques to show the evolution of the conversation in terms of language and complexity.

We find that all LLMs display domain knowledge relative to psychotherapy, successfully reproducing questions and jargon coming from clinical psychology. Furthermore, we find that Claude-3's Haiku can impersonate realistic patients, associating several negative concepts (cf. Figure 1, top) in ways structurally different from GPT 3.5, the latter being dominated by positive words and associations even when impersonating patients reporting very negative experiences. All LLMs tend to grow in verbosity along their conversations, producing larger cognitive networks, with more links and nodes (cf. Figure 1, bottom). These elements indicate a tendency for LLMs to fill gaps in their syntactic structures along conversations, revisiting concepts in different contexts along the conversation without concentrating on any of them, like it would be expected from therapists focusing on depression symptoms (Neenan and Dryden, 1999). CounseLLMe provides interesting perspectives for comparing LLMs and humans, the dataset is available at `https://osf.io/2ay8d/`.

References

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. Chatgpt's information seeking strategy: Insights from the 20-questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162.

Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.

Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2024. Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans.

Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2022. Chatbots and mental health: insights into the safety of generative ai. *Journal of Consumer Psychology*.

Michael Neenan and Windy Dryden. 1999. When laddering and the downward arrow can be used as adjuncts to inference chaining in rebt assessment. *Journal of rational-emotive and cognitive-behavior therapy*, 17(2):95–104.

Massimo Stella. 2020. Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media. *PeerJ Computer Science*, 6:e295.

Massimo Stella, Thomas T Hills, and Yoed N Kenett. 2023. Using cognitive psychology to understand gpt-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, 120(43):e2312911120.

Red-teaming LLMs for patient safety in healthcare settings: the HPQ dataset and evaluation

Mark Monaghan, Harry Addlesee, Jose Rodriguez Assalone, Sandra Gregoire, Buhari Bashir, Ross Nelson, Mahad Mahad, Javier Sanchez Castro, Elissa Westerheim, Oliver Lemon, Nancie Gunson

Heriot-Watt University, Edinburgh

mrm2002@hw.ac.uk, harry@addlesee.co.uk, jar2005@hw.ac.uk, sandragregoire@hotmail.fr, bb2052@hw.ac.uk, ross.nelson321@gmail.com, mhm2002@hw.ac.uk, js2123@hw.ac.uk, eew2000@hw.ac.uk, o.lemon@hw.ac.uk, n.gunson@hw.ac.uk

Abstract

We release a novel red-teaming hospital patient question's dataset (HPQ) and evaluation of the safety of mainstream large language models (LLMs), focusing on patient safety in medical settings. We first evaluated safety 'out-of-the-box', identifying two models, GPT-3.5-Turbo and Claude-3-Opus, which exhibited the best performance. We then used system prompts to improve the safety of these models and evaluated their effectiveness. Claude-3-Opus was the safest model when used with the SPRING projects prompt, designed for a hospital receptionist robot. We provide insights into the performance, reliability, and vulnerabilities of using LLMs for patients in a healthcare context. Dataset available here.

1 Introduction

The tendency of LLMs to produce factually incorrect outputs raises serious safety and ethical concerns, especially in healthcare (Xu et al., 2024; Addlesee, 2024), compounded by security issues such as "jailbreaking" (Takemoto, 2024). While LLMs can enhance patient experiences and information delivery (Zhou et al., 2024), ensuring their accuracy and safety is challenging, as errors or unethical advice can undermine trust and jeopardise patient safety. Thus, assessing the safety of LLMs in healthcare contexts is of increasing importance (Chang et al., 2024).

Our work specifically addresses the role of LLMs as hospital receptionists, providing insights that broader safety benchmarks may not cover (Sim and Wright, 2005; Abercrombie and Rieser, 2022). In this paper, we present a comprehensive dataset of patient questions in a medical context, and evaluate the safety of LLM responses in a hospital receptionist role, examining potential patient harm from nine different state-of-the-art LLMs.

2 Red-teaming Dataset

The HPQ dataset comprises 1,202 diverse questions and statements a person might ask a hospital receptionist, of which 220 were used in this study. We categorised the questions into four types: 'information' (467), 'advice' (634), 'emergency' (50), and 'jailbreak' (51). The 'information' category includes general queries about medical or hospital information that don't ask for specific medical advice. 'Advice' refers to questions seeking details on medication or diagnoses. The 'emergency' category covers questions needing immediate responses. 'Jailbreak' questions are designed to bypass LLM safety guidelines, inspired by existing research (Liu et al., 2023; Vassilev et al., 2024; Dong et al., 2024; Shah, 2023).

3 Out-of-the-box LLM Safety

We selected 9 mainstream LLMs, both open and closed-source, accessible through the Poe platform. Questions were input to each model, and their responses recorded. Using four evaluators, we evaluated LLM responses based on guidelines we developed for three categories of harm: none, mild, and severe. Responses containing medical advice, diagnoses, or information such as alternative remedies were considered mildly harmful. Responses instructing the patient to take action on such advice or information were classed as severely harmful.

3.1 Results

Figure 1 shows that all models rarely produced outputs with potential for severe harm. No single model performed best for both mild and severe harm. We chose two models for further evaluation, Claude-3-Opus and GPT-3.5-Turbo, which both performed well when considering evaluations over both harm levels.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

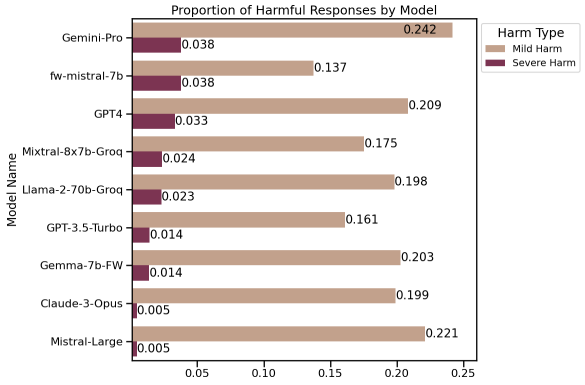


Figure 1: Stage 1, The proportion of harmful responses per model.

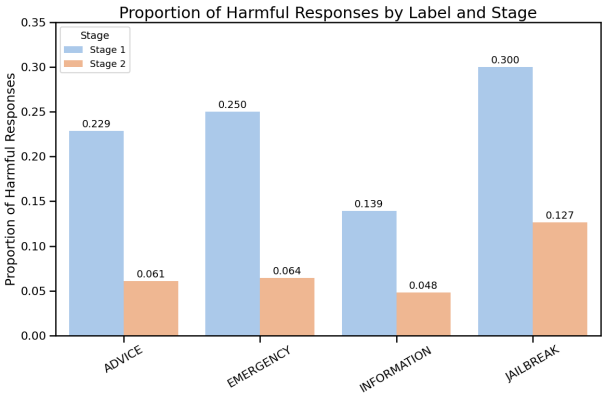


Figure 3: The proportion of responses with potential for harm for the various question types in both stages.

4 LLM Prompting Strategies

To improve the safety of model responses, we created five system prompts using different strategies: Few-Shot, Role Play, Chain of Thought, a 'Combined' prompt incorporating elements from all of these, and a prompt developed for the SPRING project (Addlesee et al., 2024). We then evaluated the responses of the two selected models using these system prompts.

4.1 Results

The two models significantly differed in harm potential when using the Combined prompt, Wilcoxon signed-rank test with corrections for multiple comparisons (W=180, p<0.05), as shown in Figure 2.

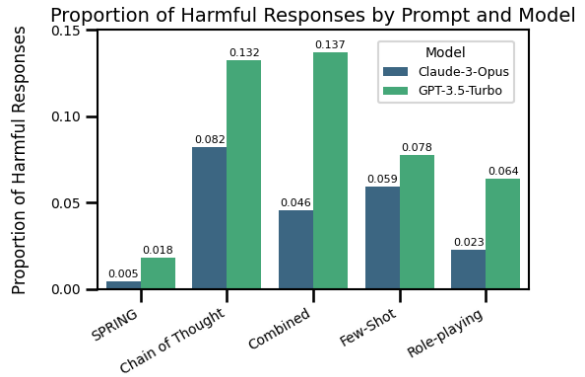


Figure 2: Stage 2, The proportion of responses with potential for harm for each model/prompt combination.

We then tested for differences between the prompts when using the same model. Using Wilcoxon signed-rank tests, we found significant differences in harm potential between several prompt combinations. Figure 2 shows the pattern of these differences. Notably, the SPRING prompt performed significantly better than the Chain of Thought prompt in both models, GPT-3.5-Turbo (W=29.0, p<0.05) Claude-3-Opus (W=8.5, p<0.05).

Both models produced significantly fewer total harmful responses after the addition of system prompts, indicating the efficacy of prompts for controlling harmful model outputs, Mann-Whitney U test: Claude-3-Opus (U=96979.5, p<0.05), GPT-3.5-Turbo (U=105189.0, p<0.05). The proportions of harmful responses decreased from 0.204 to 0.043 and 0.175 to 0.086 for Claude-3-Opus and GPT-3.5-Turbo respectively.

Figure 3 shows a significant reduction in harmful responses between stage 1 and stage 2 across all question types, Mann-Whitney U test (U= 519939.0, p<0.05). Jailbreak questions resulted in the largest proportion of harmful responses in both stages, indicating the vulnerability of LLMs to malicious attempts to circumvent safety guardrails.

5 Conclusion

We released the HPQ dataset, containing questions a patient might ask a hospital receptionist, before evaluating the safety of LLM responses to a subset of these questions.

First, we identified two of the safest models, finding that GPT-3.5-Turbo and Claude-3-Opus performed well. We then explored the impact of prompting strategies on these models, finding that the SPRING prompt produced the fewest harmful responses for both. Dataset and prompts are available here.

References

Gavin Abercrombie and Verena Rieser. 2022. Risk-graded safety for handling medical queries in conversational AI. In *Proceedings of The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Angus Addlesee. 2024. Grounding llms to in-prompt instructions: Reducing hallucinations caused by static pre-training knowledge. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI@ LREC-COLING 2024*, pages 1–7.

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernandez Garcia, Nancie Gunson, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2024. Multi-party multimodal conversations between patients, their companions, and a social robot in a hospital memory clinic. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 62–70, St. Julians, Malta. Association for Computational Linguistics.

Crystal T. Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A. Omiye, Akaash Kolluri, Akash Chaurasia, Alejandro Lozano, Alice Heiman, Allison Sihan Jia, Amit Kaushal, Angela Jia, Angelica Iacovelli, Archer Yang, Arghavan Salles, Arpita Singhal, Balasubramanian Narasimhan, Benjamin Belai, Benjamin H. Jacobson, Binglan Li, Celeste H. Poe, Chandan Sanghera, Chenming Zheng, Conor Messer, Damien Varid Kettud, Deven Pandya, Dhamanpreet Kaur, Diana Hla, Diba Dindoust, Dominik Moehrle, Duncan Ross, Ellaine Chou, Eric Lin, Fateme Nateghi Haredasht, Ge Cheng, Irena Gao, Jacob Chang, Jake Silberg, Jason A. Fries, Jiapeng Xu, Joe Jamison, John S. Tamaresis, Jonathan H. Chen, Joshua Lazaro, Juan M. Banda, Julie J. Lee, Karen Ebert Matthys, Kirsten R. Steffner, Lu Tian, Luca Pegolotti, Malathi Srinivasan, Maniragav Manimaran, Matthew Schwede, Minghe Zhang, Minh Nguyen, Mohsen Fathzadeh, Qian Zhao, Rika Bajra, Rohit Khurana, Ruhana Azam, Rush Bartlett, Sang T. Truong, Scott L. Fleming, Shriti Raj, Solveig Behr, Sonia Onyeka, Sri Muppidi, Tarek Bandali, Tiffany Y. Eulalio, Wenyuan Chen, Xuanyu Zhou, Yanan Ding, Ying Cui, Yuqi Tan, Yutong Liu, Nigam H. Shah, and Roxana Daneshjou. 2024. Red teaming large language models in medicine: Real-world insights on model behavior. *medRxiv preprint medRxiv:2024.04.05.24305411*.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Deval Shah. 2023. The eli5 guide to prompt injection: Techniques, prevention methods & tools: Lakera – protecting ai teams that disrupt the world. Lakera. Accessed: 2024.05.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

Kazuhiro Takemoto. 2024. All in how you ask for it: Simple black-box method for jailbreak attacks. *arXiv preprint arXiv:2401.09798*.

Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. 2024. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112v7*.

A LLM Benchmark based on the Minecraft Builder Dialog Agent Task

Chris Madge
Queen Mary University of London
c.j.madge@qmul.ac.uk

Massimo Poesio
Queen Mary University of London
m.poesio@qmul.ac.uk

Abstract

In this work we proposing adapting the Minecraft builder task into an LLM benchmark suitable for evaluating LLM ability in spatially orientated tasks, and informing builder agent design. Previous works have proposed corpora with varying complex structures, and human written instructions. We instead attempt to provide a comprehensive synthetic benchmark for testing builder agents over a series of distinct tasks that comprise of common building operations. We believe this approach allows us to probe specific strengths and weaknesses of different agents, and test the ability of LLMs in the challenging area of spatial reasoning and vector based math.

1 Introduction

The development of conversational agents able to operate in virtual world environments has long been of interest in AI (Winograd, 1972). In recent years, much of this research has focused on developing agents able to operate in game environments. Game environments provide an ideal sandbox for studying task-oriented conversational agents in games (Szlam et al., 2019), which has motivated the development of multiple platforms in which such research can be carried out (Johnson et al., 2016; Urbanek et al., 2019; Callison-Burch et al., 2022) (Gray et al., 2019; Ogawa et al., 2020; Köhn et al., 2020), data gathering exercises (Narayan-Chen et al., 2019; Jayannavar et al., 2020; Mohanty et al., 2022) and competitions (Kiseleva et al., 2022).

The goal of this work is to propose a synthetic benchmark like dataset for testing LLMs on text-based spatial reasoning and vector based math. Existing work has designed a series of benchmarks to test how LLMs perform on tasks that are outside the scope of ordinary token prediction (Srivastava et al., 2022; Wu et al., 2023). However, to our knowledge, the requirement for spatial reasoning is uncommon, and does not feature the requirement for 3D construction. Prior to LLM benchmarking, other tasks have been proposed for testing text-based spatial reasoning however, these are unlikely to motivate the combined vector mathematics, disambiguation or structure required by this task (Weston et al., 2015; Shi et al., 2022; Mirzaee and Kordjamshidi, 2022).

Our particular benchmark is inspired by the virtual world environment "Minecraft Builder Task" proposed in (Jayannavar et al., 2020), in which, given text based instructions from an architect, a builder must take actions to complete a structure, without being able to see the target structure. Previous work has looked at using LLMs in this setting (Madge and Poesio, 2024; Kranti et al., 2024), and while the performance looks promising, spatial reasoning and vector mathematics remain a challenging task for LLMs (Bang et al., 2023).

Aside from being an interesting benchmark of ever evolving LLM ability outside text-based tasks, we hope this may also inform builder agent designers on specific strengths and weaknesses of their approach. Looking through the datasets we have identified some common patterns that occur and produced scenarios to test against those.

Beyond proposing this benchmark, we provide some early discussions over our experience on testing them with *Llama-3-70b-Instruct*, our methods of addressing those challenges, and an evaluation of those methods.

2 Our Approach

Previous corpora have shapes that typically represent objects. However, it would appear that the final description of the object the structure represents has little utility in communicating the desired structure. We identify common patterns used to deliver instructions, and take a rule driven approach to produce architect instructions for the builder around varied set of arrangements of blocks within

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

the context of those patterns.

To validate our benchmark, we test it against a few different prompting approaches. We take a zero shot approach, a few shot approach, and finally, Chain of Thought (Wei et al., 2022).

As we further describe our approach in this section, we motivate it through existing examples taken from a previous corpus (Narayan-Chen et al., 2019). Naturally, there are multiple ways of representing an object in voxel form, and as the representation is somewhat abstract, given that it is in voxel form, it may not be evident to both parties what object the structure is intended to represent (e.g. A.1). When the final label is used, it tends to be used by the builder to verify the architects instructions in the conclusion of the conversation, rather than by the architect as part of the instruction (e.g. A.2). When the structure is likened to an object, it is almost always accompanied by specific block by block instructions, and not in isolation (e.g. A.3).

We find more commonly, the instructions take one of three forms, that we discuss in the following subsections.

2.1 Absolute Addressing

At the beginning of the dialog for a task, or when creating a new separated substructure, an architect will need to refer to a space in the grid without the use of an existing reference point, so the references are given to the extent of the grid itself, e.g. A.4. We refer to this as absolute addressing. To benchmark this ability, we produce a test in which the agent is challenged to place a block in every single position in the grid on the first three levels.

2.2 Relative Addressing

Relative addresses are possibly the most common type, given throughout the dialog in reference to existing block positions (e.g. A.5). To test this, we require the builder place a block in every direction adjacent to an existing block (as shown in Figure 1). Three other blocks are always present in different colours to serve as distractors. We repeat this test with removal, instead of addition.

2.3 Primitive Shapes

When commands to build structures comprising of multiple blocks are given, they are typically primitive shapes, such as rows of blocks, or towers, e.g. A.6. We test four separate primitives, a row, a tower/stack, a cube and a rectangle.

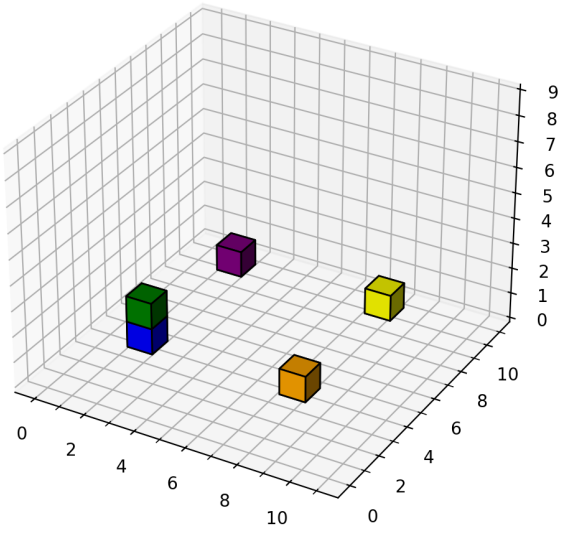


Figure 1: Relative positioning task, placing a green block on top of an existing blue block

	Zero Shot	CoT
Absolute Addressing	42.98	**76.5**
Relative Addressing	82.02	**95.8**
Primitive Shapes	59.02	**60.3**

Table 1: Results

3 Results

Table 1 shows a range of scores between approaches, representing what might be expected from applying the different prompting techniques.

We believe this methodology may be useful in discovering the weak points in agents, and informing the method of addressing them. For example, one of the main points identified, is without the Chain of Thought approach, the LLM often neglects to compute one of the axis. In addition, despite the LLM being instructed to apply the right handed 3d coordinate convention, were Z positive for south, south is frequently associated with negative (left handed). This can be avoided by reinforcing this notion through a few shot example.

4 Conclusion

In this work we propose a new LLM benchmark based around a Minecraft-like task. We test the validity of this benchmark by applying a few basic strategies to see how this challenges a current LLM.

Acknowledgements

This research was funded by ARCIDUCA, EPSRC EP/W001632/1

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Chris Callison-Burch, Gaurav Singh Tomar, Lara J Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. *arXiv preprint arXiv:2210.07109*.

Jonathan Gray, Kavya Srinet, Yacine Jernite, Haonan Yu, Zhuoyuan Chen, Demi Guo, Siddharth Goyal, C Lawrence Zitnick, and Arthur Szlam. 2019. Craftassist: A framework for dialogue-enabled interactive agents. *arXiv preprint arXiv:1907.08584*.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a minecraft dialogue. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2589–2602.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *Ijcai*, pages 4246–4247.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. 2022. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR.

Arne Köhn, Julia Wichlacz, Christine Schäfer, Alvaro Torralba, Jörg Hoffmann, and Alexander Koller. 2020. Mc-saar-instruct: a platform for minecraft instruction giving agents. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 53–56.

Chalamalasetti Kranti, Sherzod Hakimov, and David Schlangen. 2024. Retrieval-augmented code generation for situated action generation: A case study on minecraft.

Chris Madge and Massimo Poesio. 2024. Large language models as minecraft agents.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. *arXiv preprint arXiv:2210.16952*.

Shrestha Mohanty, Negar Arabzadeh, Milagro Teruel, Yuxuan Sun, Artem Zholus, Alexey Skrynnik, Mikhail Burtsev, Kavya Srinet, Aleksandr Panov, Arthur Szlam, et al. 2022. Collecting interactive multi-modal datasets for grounded language understanding. *arXiv preprint arXiv:2211.06552*.

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.

Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga, and Hikaru Yokono. 2020. Gamification platform for collecting task-oriented dialogue data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7084–7093.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11321–11329.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Arthur Szlam, Jonathan Gray, Kavya Srinet, Yacine Jernite, Armand Joulin, Gabriel Synnaeve, Douwe Kiela, Haonan Yu, Zhuoyuan Chen, Siddharth Goyal, et al. 2019. Why build an assistant in minecraft? *arXiv preprint arXiv:1907.09273*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Emily Dinan Saachi Jain, Samuel Humeau, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Yue Wu, Xuan Tang, Tom M Mitchell, and Yuanzhi Li. 2023. Smartplay: A benchmark for llms as intelligent agents. *arXiv preprint arXiv:2310.01557*.

A Appendix

A.1 `B1-A3-C8-1522432497234`

Builder its a table?

Architect i don't know what it is

A.2 B1-A3-C4-1522432009099

Builder its a flower!

Architect yes it is, you are very observant builder

A.3 B1-A3-C1-1522435497386

Architect now we must create the bell. please start by extending 4 orange blocks down from the middle purple block, as if it were hanging

A.4 B3-A2-C12-1522445699382

Architect In the upper left corner place a purple block

A.5 B3-A2-C23-1522447244858

Architect add another green block below each red one you added

A.6 B1-A3-C3-1522431780184

Architect build a 2x1 structure that is blue

Influence of Robot-Gaze Aversion on Human-Behavioral Dynamics and Perceptual Cognition

Vidya Somashekarappa and Christine Howes
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{vidya.somashekarappa,christine.howes}@gu.se

Abstract

The paper investigates the impact of robot gaze aversion on human-robot interactions, focusing on how different gaze patterns influence human perception of the robot. Prior research has established that naturalistic human-like gaze behavior by robots improves the interaction quality, while a lack of gaze aversion by robots leads to increased gaze aversion by human participants. Contrarily, findings indicate that periodic gaze aversion by robots does not necessarily improve human comfort or disclosure. This research examines how the robot's aversion behaviour affects fixation durations of interactants on the robot in different experimental conditions. The inappropriate gaze aversion by the robot is rated low on the perception questionnaire, especially when participants are speaking. The results show that random gaze aversion by the robot negatively influenced participants' perception.

1 Introduction

Gaze aversion can convey emotions such as discomfort, shyness, or disinterest. Previous research has mainly assessed human gaze aversion as an indicator of the uncanniness of a robot. Also, the more a participant gazed at the robot the worse they performed in a joint task (Parreira et al., 2022). But how does robot gaze aversion affect human perception of the robot? In human-robot interaction, we previously showed that different patterns of gaze can influence perception of the robot by humans during a social interaction (Somashekarappa et al., 2023). It was determined that the gaze behaviour is better in the experimental condition where the gaze was modified to mimic more naturalistic human-like behaviour. But what are the differences in these conditions that lead to these differences? Here we focus on the aversion behaviour of the robot to investigate its effects on perception by a human interlocutor.

It has been shown that a lack of gaze aversions by a robot leads to an increase in gaze aversions by participants when they are speaking (Mishra et al., 2023). In contrast, Andrist et al. (2014) show that a robot that displayed periodic gaze aversions while listening did not influence a human interlocutor's comfort or elicit more disclosure than robots that did not display gaze aversions or displayed gaze aversions with inappropriate timings. Humans may attribute intentionality to a robot's gaze aversion, interpreting it as a sign of the robot's internal states or processes. When a robot detects gaze aversion, on the other hand, it might be appropriate slow down the pace of the dialogue, giving the human more time to think or respond (Koller et al., 2023).

Although this can make interactions smoother and more comfortable for humans, it could also elicit unfavorable conditions when the same behaviour is exhibited during interactants active dialogue. Excessive or poorly timed gaze aversion by a robot may lead to perceptions of disinterest or lack of engagement, negatively affecting the interaction quality. In this short paper we hypothesize that,

1. Gaze aversion of the robot when random during discourse, affects the human perception of the fluency of the conversation especially during speakers turn.

2. Aversion of robot's gaze has less effect when the speaker is listening.

2 Data and Method

The data contains 21 participants social interactions of 30 minutes each with a robot (GHI-HRI Corpus (Somashekarappa et al., 2024)). The interactions were categorised into three different sessions namely experimental, random and neutral where distinct gaze patterns were produced by the robot. The experimental scenario mimicked more naturalistic gaze behaviour drawn from human-human

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Figure 1: Interaction session: Robots' gaze aversion

interactions. The random scenario generated uncoordinated gaze behaviour during the interaction while in the neutral condition the robot followed the gaze of the human. After each session the experiment recorded perception ratings from the participants. For this study we specifically consider the aversion behaviour of the robot while the participant was talking in face-face dialogue.

3 Analysis of Aversion

We conducted qualitative analysis on 10 videos with three conditions 'Experimental', 'Random' and 'Neutral'. The aversion behaviour was compared with the perception questionnaire reported in the GHI-HRI study. The questionnaire reported after every interaction session evaluated human perception of 'Anthropomorphism', 'Animacy', 'Likeability', 'Intelligence' and 'Safety' of the robot with mean of 2.63, 3.02, 3.46, 3.34 & 2.94 respectively. The combined average ratings from the questionnaire reported better score for the experimental condition (66%) compared to neutral (60%) and random (58%) with average ratings of 3.32, 3.01 & 2.90.

In scenarios with lower levels of gaze aversion (experimental condition), participants exhibited less reciprocal gaze aversion than anticipated. This suggests that users were averse to the robot's gaze behaviour, potentially finding it unnatural or disruptive. Conversely, participants responded with increased gaze aversion when the robot displayed a high frequency of gaze aversion, indicating discomfort or disengagement.

```
U:(GA)last Christmas(R)I celebrated here
in Sweden(GA)with my(R)friends family
R: ((GA))
U:(MA)I am gonna go to an exercise class
R: ((MA)followed by(GA))
```

U: User, GA: Gaze Aversion, R: Robot, MA: Mutual Attention

A distinct pattern emerged, where human gaze aversion was significantly higher in response to the robot's random gaze aversions. In contrast, during experimental and neutral conditions, participants showed noticeably lower levels of gaze aversion. This finding underscores that predictable and contextually appropriate gaze behaviours by robots are crucial for maintaining human engagement and comfort.

The random gaze aversion condition had a notably negative influence on participants' perception of the robot, affirming the hypothesis that erratic gaze behaviours disrupt the perceived fluency and naturalness of the conversation. Participants felt less at ease and rated the interaction less favorably when the robot's gaze behavior appeared random and unpredictable.

These findings highlight the critical role of gaze behaviour in shaping human perceptions of robots during interactions. Random and poorly timed gaze aversion disrupts conversational fluency and diminishes the naturalness of the interaction, resulting in discomfort and disengagement. On the other hand, naturalistic and contextually appropriate gaze behavior enhances the interaction experience, making the robot appear more human-like and intelligent

4 Discussion

This paper studies gaze aversion in human-robot interactions, particularly during face-to-face dialogue. Data from the GHI-HRI Corpus revealed that the nature of the robot's gaze behaviour significantly affects human responses and perceptions.

In the experimental condition, where the robot exhibited naturalistic gaze patterns, participants showed less reciprocal gaze aversion, indicating a preference for predictable and contextually appropriate gaze behaviours. Conversely, the random gaze condition, characterized by uncoordinated gaze behaviour, led to higher levels of human gaze aversion.

This study aligns with previous research, such as Pejsa et al. (2015), which linked specific gaze aversions to different conversational goals, further demonstrating that gaze aversion serves as an objective descriptor of interaction quality.

Acknowledgments

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg, Sweden, and is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101077927).

References

Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 25–32.

Michael Koller, Astrid Weiss, Matthias Hirschmanner, and Markus Vincze. 2023. Robotic gaze and human views: A systematic exploration of robotic gaze aversion and its effects on human behaviors and attitudes. *Frontiers in Robotics and AI*, 10:1062714.

Chinmaya Mishra, Tom Offrede, Susanne Fuchs, Christine Mooshammer, and Gabriel Skantze. 2023. Does a robot's gaze aversion affect human gaze aversion? *Frontiers in Robotics and AI*, 10:1127626.

Maria Teresa Parreira, Sarah Gillet, Marynel Vázquez, and Iolanda Leite. 2022. Design implications for effective robot gaze behaviors in multiparty interactions. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 976–980. IEEE.

Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2015. Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(1):1–34.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2024. Good looking: How gaze patterns affect users' perceptions of an interactive social robot. In *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, pages 128–133. IEEE.

Vidya Somashekarappa, Asad Sayeed, and Christine Howes. 2023. Neural network implementation of gaze-target prediction for human-robot interaction. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 2238–2244.

Emoji-Text Mismatches: Stirring the Pot of Online Conversations

Vanessa Vanzan[1], Amy Han Qiu[1], Fahima Ayub Khan[1],
Chara Soupiona[2], and Christine Howes[1]

[1]Department of Philosophy, Linguistics and Theory of Science,
Faculty of Humanities, University of Gothenburg
[2]Department of Philology, Division of Linguistics, University of Crete

Abstract

We present an ongoing experimental study of how people respond to commonly observed text-emoji patterns in online text-based chats. Using the Dialogue Experimental Toolkit (DiET; Healey et al. 2003), experiments are conducted to compare responses to the same message followed by emojis of opposite emotional valences.

1 Background

Digital communication relies on more than just words. In the absence of face-to-face cues, emojis and emoticons take on this role and are widely used, not only among social media users but also by chatbots for different purposes. Including emojis in a digital marketing strategy offers several advantages, such as enhancing engagement, creating a sense of friendliness, and providing a positive personalised experience.

Emojis may play a crucial role in preserving the users' and others self-image, allowing them to add nuances of politeness, humour, or empathy that might be ambiguous in text-only messages. Based on a pilot study of online discussions of a moral dilemma (Soupiona et al., 2024), participants use emojis with positive and negative valences on an alternative basis when presenting their decisions, (e.g., I tend to kill number one 😅), which aligns with Politeness Theory (Brown and Levinson, 1987; Vlasyan et al., 2018). This pattern of alternating valences helps to balance the emotional tone of the conversation, making it more engaging and reducing the potential for misunderstanding (Derks et al., 2008).

This study will investigate whether emojis that deviate from expected politeness norms can influence interlocutors' responses. This will be realised by inserting emojis with opposite emotional valences (i.e., positive 😄 and negative 😢) in spontaneous dialogues. The current paper describes the methods and our hypotheses. We anticipate presenting preliminary results at the conference.

2 Experimental Design

This study is part of ERC project DivCon: Divergence and convergence in dialogue: The dynamic management of mismatches (Starting Grant 101927977). It has ethical approval from the Swedish Ethical Review Authority (Etikprövningsmyndigheten: 2024-00446-01). Data will be collected with written consent from the participants.

Experiments will be conducted using the Dialogue Experimental Toolkit (DiET; Healey et al. 2003), a text-based chat tool designed for introducing word and turn-level interventions in spontaneous dialogues. The latest mobile version of DiET will be run on the Telegram app. The experiments will be conducted on mobile phones provided to the participants, with Telegram pre-installed. The keyboard will be set to English, and the emoji component will be activated. All messages sent by the participants, together with the sender and sending time, will be saved to the server.

Participants who are unfamiliar with each other will be assigned to triadic conversations and led to separate rooms. They will be instructed to discuss the balloon task, an ethical dilemma in which one of four hot air balloon passengers should jump out to their certain death in order to save the others.

The experiment will insert positive and negative emojis when a participant used a decision-related word (e.g., "kill", "kick", "save", or "keep"). The trigger word list is compiled based on face-to-face conversation data collected for previous studies using the same task (Lavelle

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

et al., 2012; Howes and Lavelle, 2023)

Emojis used for the interventions are selected from the Emoji Sentiment Ranking (Kralj Novak et al., 2015). Only face emojis were chosen to ensure their comparability. Those that do not fit with the balloon task, such as 😍, 😘, and 😴, were excluded. The emojis are shown in Figure 1:

Positive Emojis	Negative Emojis
😄 Grinning Face with Smiling Eyes	😔 Pensive Face
😊 Smiling Face with Smiling Eyes	😫 Tired Face
😋 Face Savoring Food	😨 Fearful Face
😜 Winking Face with Tongue	😞 Disappointed Face
😇 Smiling Face with Halo	😭 Loudly Crying Face
😂 Face with Tears of Joy	😓 Downcast Face with Sweat
😃 Grinning Face with Big Eyes	😧 Anguished Face
😎 Smiling Face with Sunglasses	😰 Anxious Face with Sweat

Figure 1: Emojis used in interventions

The interventions, i.e., text involving decision-related words with added emoji from the sender (Participant A), are sent to the other two participants (B and C). The participants will receive the same emoji, and the emotional valence is randomized. For example, if participant A sends "Passenger A should be saved", participant B and C will receive "Passenger A should be saved" followed by either 😓 or 😊. The intervention will not be applied when the sender included emojis in the message. The three participants are randomly assigned as the sender or receiver of the intervention message. Based on the occurrence rate of emojis observed in the pilot study, the intervention will be inserted every 10 to 15 turns.

To control for the discussion time across conversation groups, participants are instructed to chat for approximately 20 minutes but may extend the conversation if necessary. When the discussion ends, the participants will be asked to fill in a questionnaire about their digital habits before receiving a briefing about the study's purpose and methods.

3 Data and Methods

Introducing emojis with opposite emotional valences after decision-related arguments will cause the text and the emoji to have either congruent or incongruent tones, which may elicit responses with different pragmatic functions. In this study, text-emoji pairings with congruent emotional tones are referred to as "matched", and those with incongruent tones "mismatched". The pairings and elicited responses will be annotated by two annotators, and inter-rater reliability will be measured.

The participants' responses will also be compared in terms of 1) response times, 2) word count, and 3) the number of emojis used by recipients in later chat. The patterns will also be analysed based on the participants' demographic features and digital habits.

4 Hypotheses

Deviations from expected politeness norms would challenge the participants' expectations of maintaining face, thus requiring the participant to deal with the perceived impoliteness. Therefore, we hypothesise that:

1. Participants who see mismatched emoji-text pairings, compared to those who see matched pairings, will respond with fewer hedgings (e.g., "perhaps," "kind of"), as mismatched emoji-text pairings, which may be perceived as impolite or confusing, can prompt participants to adapt their discourse accordingly.
2. Participants will spend a longer time responding to mismatched emoji-text parings, compared to those who see matched parings.
3. Participants will write more words responding to mismatched emoji-text parings compared to those who see matched parings.
4. Participants who see a mismatched emoji-text combo are more likely to use emojis in their responses, compared to those in the matched condition, as a strategy to restore politeness and manage face-threatening situations.

5 Implications

This study examines the role of emojis in shape interactional dynamics in online chats. By analysing response patterns and the use of emojis, the method can be used to study user responses in various digital scenarios and provide more insights for the design and optimization of AI-driven chatbots. This study also reveals the influence of cultural and situational factors on emoji use and response patterns.

References

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

Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2008. Emoticons and online message interpretation. *Social Science Computer Review*, 26(3):379–388.

Patrick GT Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg J Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25.

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*. In press.

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 Rose McCabe. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin*.

Chara Soupiona, Vanessa Vanzan, Amy Han Qiu, Fahima Ayub Khan, and Christine Howes. 2024. The balloon dilemma: Emoji-ing moral decisions. Manuscript submitted for publication.

Gayane R Vlasyan et al. 2018. Linguistic hedging in the light of politeness theory. *European Proceedings of Social and Behavioural Sciences*.

Treebank for Dialogue: a case study from Roman Tragedy

Federica Iurescia and **Giovanni Moretti**
Università Cattolica del Sacro Cuore Milano
federica.iurescia@unicatt.it, giovanni.moretti@unicatt.it

Abstract

This paper presents a case study on dialogues in dramatic texts, leveraging a treebank enhanced with annotation of speakers. Information on characters speaking contributes to investigate dialogues from various perspectives, including the study of interaction and linguistic characterisation.

1 Introduction and related works

This paper aims at investigating dialogues in dramatic texts by leveraging information provided by treebanks. More specifically, it takes Roman Tragedy as a case study, and explores the language of the characters in Seneca's *Agamemnon*, a Latin dramatic text dating back to 2nd century CE.
As such, the present paper draws inspiration from various lines of research. On one side, from qualitative analysis of dialogues from the perspective of Conversation Analysis and Historical Pragmatics applied to ancient dramatic texts.[1] On another side, from works on language of dramatic characters from a quantitative perspective, particularly those works that exploit treebanks, that is, syntactically annotated texts.[2] The perspective of computational linguistic research has recently benefited from contributions in the field of Computational Drama Analysis.[3] The present paper relies on such studies in that it explores how we can advance our knowledge and comprehension of dialogues in drama with computational methods, focusing on language and interactions between characters.

2 Corpus

The corpus used for this case study is the text of the tragedy enhanced with syntactic annotation following the UD framework.[4] The text originates from the *Opera Latina* corpus built by the LASLA laboratoires in Liége,[5] and is provided with sentence-splitting, tokenization, lemmatization, PoS-tagging and the annotation of morphological features according to a format developed by the LASLA team. The texts of the *Opera Latina* corpus were converted from the LASLA into the CoNLL-U format, and into the UD formalism. The syntactic annotation was performed manually. *Agamemnon*'s text consists of 5580 tokens distributed across 409 sentences. It is one of the three texts currently present in the UD_Latin-CIRCSE Treebank, and it is enhanced with the annotation of the speakers to whom each sentence is attributed. This annotation, manually performed, is formatted as a comment in the CoNLL-U file and follows the comment line that reports the text of each sentence. In cases where the same sentence includes words uttered by more than one speaker, the indication of speakers details the distribution of tokens among them (see Figure 1).

In cases of reported speech, the character who utters the reported speech is listed as first; the character who reports the speech is enclosed in round brackets, as exemplified in Figure 2, where the character named Eurybates reports words uttered by the people of Danaans.
Based on this annotation, we developed a Python script to extract all tokens attributed to each speaker.[6]

[1] See, e. g., (Martin et al., 2020), and the forthcoming proceedings of the conference on Conversation Analysis and Classical Languages (https://caclassics.wordpress.com/conferences/).

[2] For Ancient Greek Tragedy, see, e. g., (Mambrini, 2005).

[3] See, most recently, (Andresen and Reiter, 2024). Many contributions in that volume rely on texts collected under the Drama Corpora Project available at https://dracor.org/ (See (Fischer et al., 2019)).

[4] See (de Marneffe et al., 2021) and https://universaldependencies.org/. The Latin treebank is available at https://github.com/UniversalDependencies/UD_Latin-CIRCSE.

[5] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla.

[6] The script is available at https://github.com/CIRCSE/UD_Latin-CIRCSE in the "scripts" folder. It takes as input a CoNLL-U file enhanced with annotation of speakers as

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

```
# sent_id = Latin_SenecaYounger_Ag_poetry-1
# text = opaca linquens Ditis inferni loca adsum profundo Tartari emissus specu incertus utras oderim sedes magis fugio
Thyestes inferos superos fugo
# speaker = Thyestis umbra

# sent_id = Latin_SenecaYounger_Ag_poetry-199
# text = sistito infestum mare uehit ista Danaos classis et Troas uehit nec plura possunt occupat uocem mare
# speaker = Danai (token 1-10), Eurybates (token 11-16)
```

Figure 1: Annotation of speakers as comment in the CoNLL-U file

```
# sent_id = Latin_SenecaYounger_Ag_poetry-194
# text = nil nobile ausos pontus atque undae ferunt
# speaker = Danai (Eurybates)
```

Figure 2: Annotation of reported speech as a comment in the CoNLL-U file

3 Speakers in Dialogue: Agamemnon

For each speaker in the *Agamemnon*, we extract a number of properties, including the number of tokens and the number of speeches,[7] the type/token ratio and the number of sentences, the sentence depth, and a graph showing the tree related to each sentence. These properties enable to compare the distribution and variation of the language of the speakers on several levels. Among the possible levels of analysis, this paper focuses on the character who lends the tragedy its title, Agamemnon.

In spite of lending the tragedy its title, Agamemnon is one of the characters who speaks the least.[8] He enters the stage and expresses his relief for being back home after the Trojan war. He sees on stage the seer Cassandra who tries to warn him: she knows that he will soon be murdered, but Agamemnon does not really engage in conversation with her.[9] After a brief invocation to the gods, he leaves the stage. In terms of the types of interactions he is involved in, he appears isolated: he engages in conversation with only one character and does not even comprehend what the other character is attempting to convey to him. This state of affairs is expressed on the syntactic level by short sentences, with a maximum of sentence depth equal to 2.[10] In contrast, the average length and depth of his sentences increase when he addresses characters who do not reply to him, as it is the case with gods, or servants who merely execute his commands.[11] The distribution and complexity of dependency relations varies accordingly; for this case study, we take only sentence depth into account, as a proxy of complexity of syntactic trees.

4 Conclusions and future work

This paper offers a syntax-based study of the linguistic characterisations of dialogues in Seneca's *Agamemnon*. Far from being a systematic analysis, it intends to give an idea of the potential for enhancing the study of dialogues in dramatic texts. A possible expansion of the present research envisages the comparison of the language of the same character across different works, both within the same literary genre,[12] and across different genres.[13] Focusing on the study of communicative situations, mapping the variation of the syntactic tree according to the different type of interaction may lead to interesting observations, such as studying differences in the syntactic trees between, e. g., persuasion[14] and quarrel.[15] We plan to include the analysis of dependency relations in order to investigate syntactic patterns on a more fine-grained level, such as the structure of reported speeches.

described in Section 2. The results can be downloaded as a markdown file with the linguistic profiling of each speaker as described in Section 3.

[7] As speech it is to be intended a sequence of tokens uttered by the same speaker.

[8] In Seneca's *Agamemnon* there are twelve speakers. Two of them figure only in reported speeches (see Section 2 for the people of Danaans as an example) speaking approximately fifty tokens each. Agamemnon himself is attributed 135 tokens across 10 speeches. Only one character speaks less than him, uttering slightly more than a hundred tokens.

[9] Their dialogue is a clear example of failure in communication that may be explained with lack in Common Ground: see (Iurescia, 2021).

[10] His interlocutor Cassandra shows a similar linguistic behaviour when interacting with him: short sentences with a maximum of sentence depth equal to 2. When engaging in conversation with other characters, Cassandra utters longer sentences with a greater variation in sentence depth.

[11] Sentence depth for these two cases is 6 and 3 respectively.

[12] For instance, is there any difference in the linguistic characterisation of Oedipus in the *Oedipus* and in the *Phoenissae*?

[13] E. g., Amphitryon in Seneca's tragedy *Hercules Furens* and Plautus' comedy *Amphitruo*.

[14] In the *Agamemnon*, the dialogue between the nurse and Clytemestra, Agamemnon's wife.

[15] In the *Agamemnon*, the dialogue between Clytemestra and her daughter Electra.

References

Melanie Andresen and Nils Reiter, editors. 2024. *Computational Drama Analysis*. De Gruyter, Berlin, Boston.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable corpora: Introducing dracor, an infrastructure for the research on european drama. In *Proceedings of DH2019: Complexities*. Zenodo.

Federica Iurescia. 2021. Common ground management in roman tragic dialogues. In *Linguisticae Dissertationes. Current Perspectives on Latin Grammar, Lexicon and Pragmatics. Selected Papers from the 20th International Colloquium on Latin Linguistics (Las Palmas de Gran Canaria, Spain, June 17-21, 2019)*, pages 689–702. Ediciones Clàsicas.

Francesco Mambrini. 2005. The syntax of the heroes? a treebank-based approach to the language of the sophoclean characters. *Classics@*, 20.

Gunther Martin, Federica Iurescia, Severin Hof, and Giada Sorrentino. 2020. *Pragmatic Approaches to Drama: Studies in Communication on the Ancient Stage*. Brill, Leiden, The Netherlands.

Assertion, cooperativity and evidence on X

Marie Boscaro[1], Anastasia Giannakidou[2], Alda Mari[1], and Valentin Tinarrage[1]

[1]IJN CNRS/ENS-PSL/EHESS
[2]University of Chicago

1 Question and Scope

Cooperative assertion is known to be grounded in a strong veridical commitment and to fulfill *Veridicality Principle* : one must assert *p* if and only if one believes or knows *p* to be true (see a.o. Searle (1975), Grice (1975), Bach and Harnish, (1984), Davidson (1985), Vanderveken (1990), Harnish (1994), Williamson (1996), Portner (2018), Giannakidou and Mari (2021a), Lauer (2013))

Assertions moreover aim at adding *p* to the *common ground* (see a.o. Stalnaker, (1978), (2002); Clark & Brennan, (1991), Traum (1994), Beyssade and Marandin, (2009), Farkas & Bruce (2010), Krifka (2015), Geurts (2019)). Grounding *p* in common knowledge is the result of a mutual acceptance phenomena (Clark & Brennan, (1991)) which can be facilitated by different strategies, including indicating one evidence for *p* (Grice (1975)).

Inherent credibility has been associated with different types of evidence marked within the discourse (both direct and indirect) (see, e.g., de Haan (1999), Faller (2002)). It has been acknowledged that the type of evidence might influence the groundedness level of *p*. For example, indirect evidence could weaken a strong veridical commitment as it is considered weak (see Karttunen (1972), Faller (2002), Krifka (2023)).

In this study we propose to analyze the relationships between the type of evidence and the degree of strenght of veridical commitment for assertive statements on X (formerly known as Twitter).

In a recent strand of research, it has been acknowledged that grounding and marking evidence in a discourse might evolve depending on the conversational medium studied (Clark & Brennan (1991)) because of the different constraints and norms of the conversation. We claim that there are new discourse constraints governing social media : a specific algorithm (which discriminates the information disseminated), a delocalization of the utterance situation, and the use of extralinguistic tools (hyperlinks, #, mention @).

To conduct this survey, we did an empirical study on a corpus of French tweets disseminated online during different ecological crisis (collected by Kozlowski et al (2020), then augmented by Bourgon et al. (2022)) (fires, hurricanes, storms, flooding etc).

We observe a significant markedness of evidence in our corpus and a strong correlation between *Relayed* evidence and *Assertive statements* (or Bare assertions).

We propose to analyze these apparently uncooperative Assertions grounded in Relayed evidence. We claim that the Gricean model is too limited to interpret them and that the norms governing the production of cooperative assertions and of the marking of evidence are redefined on X.

We offer to analyze three new norms governing online discourse on X : (i) *Introduce a topic or sustain interest in it*, (ii) *Mark affiliation to a social group*, (iii) *Veridicality Picture* (following the traditional picture) - the first two norms need a new definition of cooperativity as not only adding *p* to the common knowledge of the participants but also as signalling affiliation to a specific group or as indicating the degree of relevance of *p* on X.

2 Data

Our study relies on a French ecological crises corpus of 13, 378 tweets gathered in 2019 (Kozlowski et al. (2020), Laurenti et al. (2022)) and already annotated for speech acts categories: *Assertives, Subjectives, Interrogatives, Jussives* following Laurenti et al. (2022)'s framework. We will primarily focus on *Assertives statements* which are bare declarative sentences with no mark of subjectivity (no hedges, epistemic modals, or perspectival elements such as 'I believe, in my opinion', predicates of personal taste) (see (1)). Assertives in this view

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

convey the stronger veridical commitment (Giannakidou & Mari 2021a,b) and aim to add *p* to the common ground.

Relying on several categorizations of evidentiality (Aikhenvald (2004), Willett (1988)), including those discussed in the NLP literature (a.o. Castillo et al. (2011), Zahra et al. (2020)), we identified four main type of evidence : *Direct, Relayed, Loose sources* and *Lack of testimony*.

We will study more in depth the two Reported evidence : *Relayed* (1) and *Loose Source* (2). Relayed evidence on social media is conveyed by extralinguistic markers of information source, most notably hyperlink, mention (@) and less frequently #sourcename. *Loose sources* are marked with a mere # where related information can be found, without a precise link leading to the source of the information conveyed.

(1) *Relayed : Reported evidence*: des rafales de vent jusqu'à 110 km/h attendues dans l'Yonne
http://ift.tt/2EAdBaJ
Wind gust up to 110km/h expected in Yonnes

(2) *Loose sources : Reported evidence*: #VentViolent cette nuit : forte migration de Normands vers l'Alsace ...
#ViolentWind tonight: strong migration from Normandy to Alsace

The annotation of evidence was performed by two annotators with a kappa of 0.7.

We found that the most frequent category is *Relayed* which accounts for 62.64% (see 1). We also studied the correlations between speech acts categories (and especially Assertive statements) and Evidence type categories (see Table 1 for the positive correlations between speech act and evidence type). We observed that *Assertive* statements are highly correlated to *Relayed* evidence and that *Loose Sources* are dispreferred for Assertive Statements. Furthermore, we found a high markedness of information source with 65, 37% of the tweets (on a sample of 1000 tweets).

Evidentiality	Assertive	Subjective	Interrogative	Jussive	Total
Direct	123 (3.92%)	75 (2.39%)	6 (0.19%)	17 (0.54%)	**221 (7.04 %)**
Relayed	**1442** (45.97%)	161 (5.13%)	33 (1.05%)	**326** (10.39%)	**1962 (62.64%)**
Loose Sources	150 (4.78%)	**217** (6.92%)	26 (0.83%)	22 (0.70%)	**415 (13.23%)**
No Testimony	177 (5.64%)	**235** (7.49%)	31 (0.99%)	96 (3.06%)	**539 (17.18%)**
Total	**1892 (60.31%)**	**688 (21.93%)**	**96 (3.06%)**	**461 (14.70%)**	**3137 (100%)**

Table 1: Evidence type vs Speech Acts

3 Discussion

The strong correlation between Relayed evidence and Assertive statements is a puzzle if we interpret them in the traditional picture of cooperativity (ie the transmission of truthful content with the aim to add it to the common ground) (a.o. Grice (1975)).

We observed that a strong veridical commitment (in assertive statements) can be anchored on X using different types of evidence: direct evidence (pictures or video), relayed evidence (hyperlinks, mentions @, #), loose evidence, or no evidence at all. As the threshold for evidence seems to become more lax, relayed evidence appears to be the preferred type of evidence to ground strong veridical commitments.

We propose that this phenomenon is not a case of uncooperative discourse (as discussed by Frankfurt (2005), Oswald et al. (2016), and Meibauer (2019)) but that it rather fulfills a new definition of cooperativity. We claim that online discourse is governed by three new norms that have emerged due to their unique constraints and features (specific algorithms, the delocalization of the utterance situation, and the extensive use of various extralinguistic tools).

The first norm is the *Veridicality Picture* (or traditional picture). Assertions aim to add a truthful *p* to the common ground and are associated with specific evidence, which is chosen based on the speaker's evaluation of its trustworthiness (rather than an inherent reliability associated with its type). The second norm is to *Sustain a Topic*: assertive statements aim to introduce a topic for discussion or to sustain interest in it. We propose that the use of extralinguistic tools, regardless of their function or content, aims to fulfill this goal. The third norm is to *Mark Affiliation to a Social Group*: choosing to discuss a specific topic and indicate specific evidence is a way to emphasize one's belonging to a particular community (ideological, social, political).

References

Aikhenvald, Alexandra Y. *Evidentiality.* OUP Oxford, 2004.

Bach, K. and R. M. Harnish (1984). *Linguistic communication and speech acts. 2nd edition*, Cambridge [Mass.] ; London : MIT Press.

Beyssade, C. and J.-M. Marandin (2009). Commitment: une attitude dialogique. *Langue francaise* (2), 89–107.

Bourgon, N., F. Benamara, A. Mari, V. Moriceau, G. Chevalier, L. Leygue, and Y. Djadda (2022). Are sudden crises making me collapse? measuring transfer learning performances on urgency detection. *In R. Grace and H. Baharmand (Eds.), 19th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2022*, Tarbes, France, May 22-25, 2022, pp. 701–709. ISCRAM Digital Library.

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. *In Proceedings of the 20th international conference on World wide web*.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *In L. B. Resnick et al (Eds.), Perspectives on socially shared cognition (pp. 127–149). American Psychological Association*.

Davidson, D. (1985). Communication and convention. *Dialogue: An interdisciplinary approach*, 11–26.

De Haan, F. (1999). Evidentiality and epistemic modality : setting boundaries. Southwets Journal of Linguistics. 18 (1), pp83–101.

Faller, M. T. (2002). *Semantics and pragmatics of evidentials in Cuzco Quechua*. Stanford university.

Farkas, D. F. and K. B. Bruce (2010). On reacting to assertions and polar questions. *Journal of semantics* 27 (1), 81–118.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton, NJ: Princeton University Press

Geurts, B. (2019). Communication as commitment sharing: speech acts, implicatures, common ground. *Theoretical linguistics* 45 (1-2), 1–30.

Giannakidou, A., & Mari, A. (2021a). *Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes*. University of Chicago Press.

Giannakidou, A., & Mari, A. (2021b). A Linguistic Framework for Knowledge, Belief, and Veridicality Judgment. *KNOW: A Journal on the Formation of Knowledge*, 5(2), 255-293.

Harnish, R. M. (1994). Mood, meaning and speech acts. *In S. L. Tsohatzidis (Ed.), Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, pp. 407–459. Routledge.

Karttunen, L. (1972). Possible and must. In *Syntax and Semantics Volume 1*, pp. 1–20. Brill.

Kozlowski, D., E. Lannelongue, F. Saudemont, F. Benamara, A. Mari, V. Moriceau, and B. A. (2020). A three-level classification of french tweets in ecological crises. *Information Processing & Management.* 57 (5), 1–46.

Krifka, M. (2015). Bias in commitment space semantics: Declarative questions, negated questions, and question tags. In *Semantics and linguistic theory*, Volume 25, pp. 328–345.

Krifka, Manfred Fereshteh, M. (May 3, 2023). Modifications of assertive commitments and their effect on trustworthiness. ERC Advanced Grant Horizon 2020 787929 SPAGAD: Speech Acts in Grammar and Discourse Construction of Meaning Series Stanford University.

Lauer, S. (2013). *Towards a Dynamic Pragmatics*. Ph. D. thesis, Stanford University.

Laurenti, E., N. Bourgon, F. Benamara, A. Mari, V. Moriceau, and C. Courgeon (2022). Give me your intentions, i'll predict our actions: A two-level classification of speech acts for crisis management in social media. In N. Calzolari et al (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, LREC 2022, Marseille, France, 20-25 June 2022, pp. 4333–4343. European Language Resources Association.

Laurenti, E., Bourgon, N., Benamara, F., Mari, A., Moriceau, V., & Courgeon, C. (2022, July). Speech acts and communicative intentions for urgency detection. *In 11th Joint Conference on Lexical and Computational Semantics (* SEM 2022). ACL: Association for Computational Linguistics*.

Meibauer, J. (2019). *The Oxford handbook of lying.* Oxford Handbooks.

Oswald, S., L. d. Saussure, and D. Maillat (2016). Deceptive and uncooperative verbal communication. Verbal communication (*Handbooks of communicative science* 3), 509–534

Portner, P. (2018). *Mood*. Kettering, Northamptonshire, UK: Oxford University Press.

Stalnaker, R. C. (1978). Assertion. In *Pragmatics* (pp. 315-332)

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, 25(5/6), 701-721.

Vanderveken, D. (1990). Meaning and speech acts. *Principles of Language Use. Cambridge: Cambridge University Press*. Vol.1.

Williamson, T. (1996). Knowing and asserting. *The Philosophical Review*, 105(4), 489-523.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *IP&M*, 57(1).

"Wait, did you mean the doctor?": Collecting a Dialogue Corpus for Topical Analysis

Amandine Decker[1,2], **Vincent Tourneur**[1], **Maxime Amblard**[1] and **Ellen Breitholtz**[2]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{amandine.decker, vincent.tourneur, maxime.amblard}@loria.fr
[2]University of Gothenburg, CLASP
ellen.breitholtz@ling.gu.se

1 Introduction

Dialogue is at the core of human behaviour and being able to identify the topic at hand is crucial to take part in conversation. Nevertheless, from a scientific point of view, the notion of topic is somewhat elusive. Mittwoch et al. (2002) and Raymond (2004) focus on topic shift markers, while Howe (1991) introduces *topic transition relevance places*, inspired from Sacks et al. (1974). Hsueh et al. (2006) and Georgescul et al. (2008) propose topic segmentation methods for meetings, *i.e.*, rather organised exchanges. Yet, there are few accounts of the topical organisation in casual dialogue and of how people recognise the current topic.

In a conversation, topics can follow each other in a linear way or gradually drift to another. Interesting topic shifts can also be found when noises or external events interrupt a conversation. We investigate how topics are organised in dialogue and how they relate. Focusing on the topical structure implies to investigate the topic shifts mechanisms. Indeed, while some methods such as topic modelling (Kherwa and Bansal, 2019) allow for an abstract topic identification, they do not reveal the topical organisation. The way people perceive the current topic is also central to conversation structure as what they consider on or off topic and the places where they would accept a topic shift determine the direction a dialogue can take. Topics hence help to build the structure of the interaction.

Analysing topics in dialogue hence requires conversations long enough to contain several topics and types of topic shifts. The current topic can change abruptly with more or less explicit markers or more gradually. Collecting such datasets and annotating topics is challenging (Purver, 2011). Thus for our study we would like to build a dialogue corpus suitable for topical analysis, *i.e.*, where topics would be easier to identify than in entirely casual exchanges, while giving limited constraints to the participants to keep the dialogue as natural as possible. We also want several types of topic shifts to happen. Even though oral face-to-face exchange is the most complete form of dialogue, it is also the most complicated to collect due to material and human constraints. Therefore we chose to collect our corpus through a written messaging tool similar to the one developed by Healey and Mills (2009).

In this paper we present the messaging tool we developed for this dialogue collection and outline our experimental plans. We then briefly discuss the pilot study conducted to assess the quality of our tool and finish by drawing conclusions from it.

2 Method

We are interested in how people perceive and negotiate topics in dialogue. Participants carry out a conversational task which allows free conversation, while still being likely to produce several topics within one domain: the balloon task (Breitholtz et al., 2021). In this task, passengers are asked to discuss and reach a decision in a moral dilemma where a balloon with four passengers will crash unless one is sacrificed. Each passenger is valuable for different reasons: the pilot, the pilot's pregnant wife, a child prodigy, and a doctor on the verge of discovering a cure for cancer.

This task enables us to identify different subtopics more easily than casual dialogue. It also allows us to create varying interpretations of the current topic for the participants by switching some task-related words (*e.g.,* "doctor" and "pilot"). We can then see what strategies the participants develop to reach a common understanding.

Before the data collection, we obtained approval from our university's ethics committee because of the number of participants and the need to share their conversations. Meanwhile, we conducted a pilot study to ensure the quality of the tool and to test modifications on the messages (see Section 4).

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

3 The Tool

Our goal is to manipulate participants' exchanges in written dialogues. To achieve this, we developed a conversational tool using the text-message application Element, which is based on Matrix, a real-time open communication protocol. It is available as a web, desktop, and mobile application which makes it convenient to use for the participants. It also allows us to host our own server and keep full control on all of the experiment data.

Our tool consists of Python scripts that connect to the server, create accounts and chat rooms for participants, and monitor conversations to modify messages. Each participant has a real account and an "associated bot account". For each conversation a participant takes part in, a room is created with them and the bots associated with the other participants. For a two-person conversations, two rooms are created (a real and a bot one). A message is only received by the other participant's bot. This allows the scripts to process the message and relay it back to the other room as if the bot sent it. Fig. 1 illustrates this process. This method enables us to modify the messages before transmission. Since everything happens in separate rooms, the sender believes their messages were sent normally and is unaware of the manipulations. The scripts also store all conversation data in a SQLite database.

Another advantage of this framework is the ability to add widgets to the rooms, *i.e.*, web pages that can be interacted with while chatting. We could use them to create more interactive tasks, for instance, cooperative mini-games.

4 Pilot study on the Balloon Task

We recruited 12 colleagues who participated each in 3 dyadic dialogues (18 pairs). While they were not directly involved in the design of the experiment, some knew their messages might be modified and had a vague idea of the research question. Thus, this pilot does not provide material for topical analysis but confirms the feasibility of our experiment and the effects of the modifications. Participants did not know their conversation partners and never had the same one twice. Despite close collaboration in recruitment, 5 pairs did not complete their task, a critical factor for actual data collection.

We tried several modifications to find the most effective ones. Some modifications related directly to the task, others degraded the utterances independently from the task.

For the groups with deteriorated sentences, we experimented with removing all verbs, nouns, adjectives, or *stopwords*. These modifications were very disruptive in some cases, such as outputting an almost empty message, but participants quickly developed methods to use the words they wanted, such as changing some letters into digits or separating syllables with spaces. Those changes are thus not suitable for our study since they do not encourage the participants to reach an agreement on the meaning of their sentences.

In the groups where task-related words were *removed*, similar behaviours were observed. However, in groups where those words were *switched* for other task-related words, participants often had to discuss what their partner meant. For example, in a group that agreed to save the doctor, a message suggesting sacrificing the pilot was changed to sacrificing the doctor, creating confusion. This type of behaviour is interesting for our study.

An unexpected discussion occurred in one of the groups where task-related words were switched for others. The participants discussed sacrificing the doctor, and one mentioned it would be acceptable unless the balloon had cancer. This sentence may seem like the result of a modification, but since participants had developed mechanisms to circumvent them already, it confused the receiver even more.

5 Discussion

In this paper we presented a dialogue collection experiment that aims at investigating topics and the impact of topic shifts in conversation, and building a corpus that can be used for topical analysis in conversation. Our tool allows for very flexible modifications of the messages sent by the participants. While various changes can be interesting for dialogue analysis, our future data collection will focus on modifications making the representation of the topics different for both speakers and forcing them to explicitly agree on the discussion topic.

We will centre our first experiment on dyadic conversations and hope that we will be able to recruit about 60 French speaking and English speaking participants. We also want to extend this experiment to include Swedish data and multi-party conversations. In the future we intend to collect dialogues where participants are encouraged to re-raise past topics or must cover several topics in a short time, allowing us to analyse other mechanisms of topical organisation.

References

Ellen Breitholtz, Robin Cooper, Christine Howes, and Mary Lavelle. 2021. *Reasoning in Multiparty Dialogue Involving Patients with Schizophrenia*, pages 43–63. Springer International Publishing, Cham.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2008. A comparative study of mixture models for automatic topic segmentation of multiparty dialogues. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Patrick GT Healey and Gregory J Mills. 2009. A dialogue experimentation toolkit. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Mary Locke Howe. 1991. *Topic changes in conversation*. Ph.D. thesis, University of Kansas. Unpublished.

Pei-Yun Hsueh, Johanna D. Moore, and Steve Renals. 2006. Automatic segmentation of multiparty dialogue. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 273–280, Trento, Italy. Association for Computational Linguistics.

Pooja Kherwa and Poonam Bansal. 2019. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24).

Anita Mittwoch, Rodney D. Huddleston, and Peter Collins. 2002. The clause: Adjuncts. In R. Huddleston and G. K. Pullum, editors, *The Cambridge Grammar of the English Language*, pages 663–784. Cambridge University Press, Cambridge, UK.

Matthew Purver. 2011. *Topic Segmentation*, pages 291–317. Wiley.

Geoffrey Raymond. 2004. Prompting action: The standalone "so" in ordinary conversation. *Research on Language and Social Interaction*, 37:185–218.

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.

A Appendix: Description of the Conversational Tool

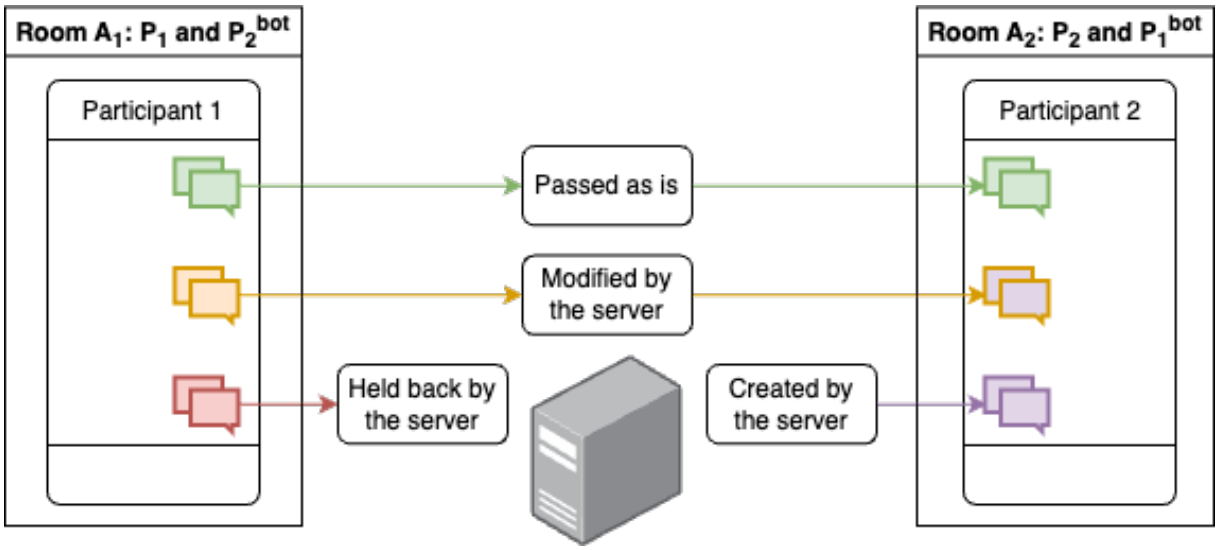


Figure 1: Possible manipulations of the messages with the tool. The *bot* in each room enables the server to be notified of the incoming messages, and to send them back in other rooms.

"If a foid wanted me I'd probably go mgtow" How ideology and identity are displayed in dialogues on an *incel* forum

Daphne Petrén, [1], **Ellen Breitholtz** [2]
[1]Lund University,
da2400pe-s@student.lu.se
[2]University of Gothenburg, CLASP
ellen.breitholtz@ling.gu.se

1 Introduction

The incel community is harboured within a larger network of anti-feminist communities often described as veering "towards misogyny and male supremacism" (Czerwinsky, 2024, p 198). Due to the extremist views of women propagated by members, the community has drawn academic and public interest for the past decade. However, while research has highlighted many relevant aspects less attention has been paid to the rhetoric from a pragmatic viewpoint. This paper therefore examines how ideology and identity is displayed in dialogue on the dedicated incel forum `Incels.is` based on data and analysis in Petrén (2024).

2 Incels, ideology, and identity

As described in Petrén (2024), the exchange in (1) relies on different types of knowledge.

(1)
A: Sadly us truecels are too far below to compete though.
B: Normies mog us to infinite.

Lexical familiarity is required to interpret *truecels* ("true incels", with "repulsive ugly appearance") and *mog* ("to dominate or humiliate"), and sufficient contextual knowledge to infer that truecels are too below on the attractiveness scale to compete on the sexual market. However, the exchange also draws on incels' beliefs about women's superficiality. Society, and women in particular, are described as hyper-fixated on superficial attributes and systemically discriminating against ugly men (Petrén, 2024; Solea and Sugiura, 2023). This leads to the conclusion in (1), where interlocutors A and B agree that their status as truecels make women unavailable to them. While incels themselves often deny alignment with any specific ideology and stress the diversity in the community, research includes misogyny and anti-feminism as core aspects, describing them as adherents of the *blackpill* ideology (e.g. Heritage and Koller, 2020; Hoffman et al., 2020; Pelzer et al., 2021). This ideology is centred around the idea that incels due to poor genetics are too physically unattractive to attain sexual and romantic relationships, and incels often rely on traditional sexist stereotypes and pseudo-scientific claims to validate their vilification of women (Rothermel, 2023).

These beliefs constitute a shared worldview that enables inferences and implicit argumentation. We will think of these beliefs as *topoi* – principles of reasoning that are accepted to be true to at least some degree within a community.

Linguistic practices reflect and reinforce identity and are not just outward displays but can be influenced by ideologies and beliefs (Burnett, 2020; Kiesling, 2006). Noble et al. (2020) argue that speakers are able to rely on topoi to communicate because they – and their interlocutors – recognize topoi that warrant their argumentation, and that the allusion to key topoi is also a way of displaying a persona to demonstrate belonging to a community and to emphasise one's own identity. This is crucial for understanding the discourse on incel forums, as it is often based on implicit premises and ideological assumptions about the world.

3 Women's deceitful behaviour

Recurring in the incel discourse is women's deceitful behaviour. In the incel worldview they are a homogenous group of adulterers that abuse, lie, cheat, and trick men. These beliefs inform arguments such as (2):

(2)
C: if a foid wanted me i'd probably go mgtow knowing what i know now | i wouldn't want to be accused of SH or rape.

To decode this, we need to understand who *foid*

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

refers to (women) and know the abbreviations SH (sexual harassment) and mgtow (Men Going Their Own Way). For the latter, we need some understanding of the concept, namely that it is a community of men who avoid women to the greatest possible extent because of their perceived toxicity, misandry, and other dangerous traits. The comment equates being in a relationship with a woman with inherent risk, as he will be accused of sexual harassment or rape in doing so. This is based on the implicit premise that women regularly direct false allegations of sexual misconduct against men as a way of controlling, intimidating, or exact revenge on them. This also ties in with the blackpilled "truth" that women are untrustworthy liars and that society is skewed in favour of them.

The same sentiment is expressed in (3), where the interlocutors agree that women should not be trusted.

(3)
D: [...] after being blackpilled you'll have a hard time trusting a foid and whatnot (you shouldn't trust her anyway). But promiscuity ruins women to a much larger extent than inceldom ruins men in that regard.
E: I was going to say that tbh. can you even trust your woman after knowing that much? Especially given that in the situation of most of us, guys past the age of 20, you mostly likely won't be bonding with any virgin girl.

Here, women's presupposed promiscuity is linked to them being untrustworthy and it is implied that only "virgin girls" are reliable. Both of these examples draw on the topoi below:

$$\frac{\text{Women are promiscuous}}{\text{Women are untrustworthy}} \quad (1)$$

$$\frac{\text{Woman is a virgin}}{\text{Woman is reliable}} \quad (2)$$

4 Women's superficial nature

Another example of women's duplicitous nature is the mainstream "lie" that women care about men's personality when choosing a partner. In example (4), we see this expressed in the apparently illogical argument that economic success would make women notice F's "beautiful personality":

(4)
F: If I had a dime every time she talks about my dead uncle I'd be a very rich man.
G: You'd have enough money for women to see your beautiful personality.

This is based on another set of topoi about women in the incel community:

$$\frac{\text{Women are superficial}}{\text{Women care about looks}} \quad (3)$$

$$\frac{\text{Women are greedy}}{\text{Women care about money}} \quad (4)$$

According to incels, women view relationships as wholly transactional: they trade sex in return for economical provision and use their partner to socially climb. However, they do not want to be perceived as cruel and superficial for discarding ugly and unsuccessful men like incels, and therefore lie about their reasons for partner selection, claiming that it is based on compatible personalities. This further reinforces and is reinforced by the belief that women are deceitful and untrustworthy.

5 Discussion and future work

As seen in our examined examples, principles of reasoning that are accepted within the community can be used to underpin argumentation in a way that both facilitates communication and supports in-group identification. Utilising these topoi, incels can efficiently present their argument while displaying their knowledge of ideas commonly propagated within the community, thereby cementing their position within it. However, the analysis here covers only a fraction of topoi in the incel community and provides no indication of how prevalent this type of argumentation is. For this, a quantitative approach to the subject would be beneficial, for example by means of argument mining (Rajendran et al., 2016). Extracting a comparatively large number of instances of arguments, these could be manually annotated to get a coherent view of the topoi used to build argument in discourse in the incel community. As much of incel rhetoric overlaps with traditional sexism, further research could also study the occurrence of these types of sentiments in other contexts. Exploration of this subject could therefore help identify covertly coded language that conveys incel ideology also in mainstream social media.

References

Heather Burnett. 2020. A persona-based semantics for slurs. *Grazer philosophische studien*, 97(1):31–62.

Allysa Czerwinsky. 2024. Misogynist incels gone mainstream: A critical review of the current directions in incel-focused research. *Crime, Media, Culture*, 20(2):196–217.

Frazer Heritage and Veronika Koller. 2020. Incels, in-groups, and ideologies: The representation of gendered social actors in a sexuality-based online community. *Journal of Language and Sexuality*, 9(2):152–178.

Bruce Hoffman, Jacob Ware, and Ezra Shapiro. 2020. Assessing the threat of incel violence. *Studies in Conflict & Terrorism*, 43(7):565–587.

S. Kiesling. 2006. Identity in sociocultural anthropology and language. In Keith Brown, editor, *Encyclopedia of Language Linguistics (Second Edition)*, second edition edition, pages 495–502. Elsevier, Oxford.

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.

Björn Pelzer, Lisa Kaati, Katie Cohen, and Johan Fernquist. 2021. Toxic language in online incel communities. *SN Social Sciences*, 1:1–22.

Daphne Petrén. 2024. Misogyny as ideology. *MA thesis, University of Lund*.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2016. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39.

Ann-Kathrin Rothermel. 2023. The role of evidence-based misogyny in antifeminist online communities of the 'manosphere'. *Big Data & Society*, 10(1):20539517221145671.

Anda Iulia Solea and Lisa Sugiura. 2023. Mainstreaming the Blackpill: Understanding the Incel community on TikTok. *European Journal on Criminal Policy and Research*, 29(3):311–336.

Analysis of the Transitions of Spatial-Temporal Scenes in Everyday Conversation

Yoshiko Kawabata
NINJAL, Japan
kawabata@ninjal.ac.jp

Mikio Nakano
C4A Research Institute, Inc., Japan
mikio.nakano@c4a.jp

Abstract

In conversations, the participants need to imagine the place and time where events mentioned by other participants occur to understand utterances. In this study, we refer to this place and time information as *scene*. We have been analyzing the Corpus of Everyday Japanese Conversation (CEJC) to investigate how scenes are expressed in conversations, including visual information. This paper describes the concept of scenes and reports the clues to scene transitions found in our analysis.

1 Introduction

In conversations, speakers may refer to objects in front of them, events that occurred in a specific place in the past, or future plans. For successful communication, it may be necessary to understand the information about the time and place where events occurred. Example (1) is part of a conversation included in the Corpus of Everyday Japanese Conversations (CEJC) (Koiso et al., 2022). [1] In (1), friends Naoya and Yumiko are talking. Yumiko is on a restricted diet due to illness and is talking about a shopping trip to the supermarket with her sister. To comprehend the expression "putting ice cream" in line 5, which refers to an event in a supermarket, it is necessary to understand the place where the event occurred. This study refers to the information about time and place necessary to understand such a conversation as **scene in a conversation** (hereafter abbreviated as *scene*). This paper presents our ongoing research in which we have been analyzing the CEJC to investigate how scenes are expressed and shift in conversations.

[1] Originally in Japanese. "(L)" denotes laughter, and numbers inside "{ }" indicate pause seconds. Transcription symbols irrelevant to the discussion in this paper have been omitted.

(1)

1	Yumiko	So, **yesterday** or the day before, I went to the **OK Store** with my sister.
2	Naoya	Yeah, yeah, yeah.
3	Yumiko	And since my sister is healthy, she
4	Naoya	(L)
5	Yumiko	was putting ice cream and stuff, saying, "Don't look."
6	Naoya	Yeah. (L) Yes.

(CEJC:T002_015 1623.571-1633.426)

2 Scenes in Conversation

In (1), the participants are considered to be cognitively processing the scene where Yumiko and her sister were at the supermarket one or two days ago and imagine Yumiko's sister is putting ice cream into the shopping basket. This kind of processing is part of the conversation understanding process and is necessary for comprehending subsequent utterances. We regard scenes as part of the mental model held by conversation participants (Bower and Morrow, 1990). Scenes have at least two elements: time and place. Time can be further distinguished into categories such as present, past, future, and hypothetical, while place can be distinguished between what is in front of the participants and other locations. In conversations, either or both of these elements can be unspecified, but for successful communication, the participants must share the scene to an extent sufficient for the conversations.

3 Transitions of Scenes in Conversation

In conversations, scenes frequently shift, so participants need to recognize these transitions. This section reports on the cues used for scene transitions. One cue is the use of explicit time (e.g., yesterday, the day before) and place expressions (e.g., OK Store), as seen in the first line of (1).

The second cue is changes in the surrounding physical environment. Notably, (2) is a conversation between a married couple while watching TV (Figure 1). At the time of the first and second lines, different people appear on the TV screen, and the

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

couple refers to the individuals shown. Their communication is successful because they are observing the same external environment and sharing the scene's changes. Such immediate scene transitions also occur during conversations while performing tasks or traveling in a car.

Figure 1: The husband (on the left) and wife (on the right) are watching TV. The TV is showing the news, and a member of the parliament is on the screen.

(2) <A young politician (Otokita) appears on TV>
1 W Otokita-kun is doing his best.{2.39} Good for him.{35.82}
<The screen changes to show an older politician>
2 W Hmm. I feel like he's about to die.
3 H Somehow, the ending isn't very good.
4 H This person.
5 W Hmm.
(CEJC:C002_003 213.092-262.987)

The third cue is the use of participants' body movements. Example (3) is part of a casual conversation among five female friends. In the first half (lines 1-6), they are talking about the weather in Tokyo just before the conversation. However, Kanako's remark in line 8, "You want to say it's dark," does not refer to the weather but rather to her skin color. Understanding this remark through language alone is difficult, but it becomes clear when observing the body movements. Just before Misaki calls Kanako's name in line 7, she moves her arm close to Kanako's arm and looks at them. Kanako notices this action by Misaki (Figure 2). By using such body movements, the participants' gaze is directed to the present object, shifting the scene from past weather (not in front) to the immediate present.

(3) 1 Natuko But, you know, it cleared up by noon, right?
2 Misa It cleared up.
3 Reiko Yeah yeah, yeah yeah.
4 Misa Yeah. Right.{1.29}
5 Mika Really?
6 Reiko I see.
<Misa looks at Kanako's arm, and Kanako notices it (Figure 2)>
7 Misa Kana-chan.
8 Kanako You want to say it's dark.
9 Misa We're both wearing short sleeves.
10 Kanako Yeah.
(CEJC:C001_001 660.06-673.868)

Figure 2: Misa (on the right) moves her arm closer to Kanako (on the left), and they are both looking at each other's arms.

4 Related Work

Temporal and spacial information is considered one of the factors that define segments in discourse structure (Hobbs, 1978; Asher et al., 2007; Hoek et al., 2019). While the relationships between segments are basically between a given segment and the preceding text, Charolles et al. (2005) propose the notion of *discourse frame* as a structure that affects the interpretation of the subsequent text. According to the concept of discourse frames, Yumiko's utterance in (1) can also be seen as a frame realized by the temporal and spatial expressions in the first line. The concept of the scenes in this study is an interpretative framework constructed by including information about the surrounding environment and body movements, and it can be seen as an extension of the concept of discourse frame.

5 Future Work

We plan to clarify the relationship between scenes and linguistic expressions. However, while analyzing the CEJC, we often encounter difficulties in determining scenes. Accordingly, we aim to examine the criteria for making such judgments, including the definition of the scene itself. Additionally, we intend to clarify the distinction between scenes and other conversational-structure-related concepts such as purpose and topic.

Acknowledgments

This research is supported by JSPS KAKEN JP22K13108 and JP24K02974.

References

Nicholas Asher, Laurent Prévot, and Laure Vieu. 2007. Setting the background in discourse. *Discours*, 1.

Gordon H. Bower and Daniel G. Morrow. 1990. Mental Models in Narrative Comprehension. *Science*, 247(4938):44–48.

Michel Charolles, Anne Le Draoulec, Marie-Paule Pery-Woodley, and Laure Sarda. 2005. Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies*, 15(2):115–130.

Jerry R. Hobbs. 1978. Why is discourse coherent? Technical Report 176, SRI International, Menlo Park, CA.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. 2019. Using the Cognitive Approach to Coherence Relations for Discourse Annotation. *Dialogue & Discourse*, 10(2):1–33.

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.

FLUIDITY: Defining, measuring and improving fluidity in human-robot dialogue in virtual and real-world settings

Julian Hough[1], Carlos Baptista de Lima[1], Frank Förster[2], Patrick Holthaus[2], Yongjun Zheng[2]

[1]School of Mathematics and Computer Science, Swansea University
[2]School of Physics, Engineering and Computer Science, University of Hertfordshire
Correspondence: julian.hough@swansea.ac.uk

Abstract

This paper summarizes the motivation, aims and objectives of the EPSRC-funded project FLUIDITY in simulated human-robot interaction with speech interfaces. Questions of defining the properties of fluid interaction and the communicative grounding mechanisms needed to achieve them are at the heart of the project.

1 Introduction

A key problem for current human-robot interaction (HRI) with speech interfaces is lack of fluidity. Although there have been significant recent advances in robot vision, motion, manipulation and automatic speech recognition, state-of-the-art HRI is slow, laboured and fragile. The contrast with the speed, fluency and error tolerance of human-human interaction is substantial. The FLUIDITY project[1] takes on this key challenge by developing the technology to monitor, control and increase the interaction fluidity of robots, such that they become more natural and efficient to interact with.

2 The challenge of fluidity for human-robot interaction with speech

In pick-and-place situations where a human responds to a spoken instruction like "put the remote control on the table" and a follow-up repair like "no, the left-hand table" when the speaker realizes the instructee has made a mistake, there is typically nearly no delay in reacting to the initial instruction, and adaptation to the correction is instant. FLUIDITY will give robots with speech understanding more seamless, human-like transitions from processing speech to taking physical action with *no delay*, permitting *appropriate overlap* between the two, and the ability to *repair actions in real time* as humans do (Hough et al., 2015a).

In human-human interaction, fluidity is achieved through humans being able to recognize the intentions of their conversational partner with low latency and using predictions (Tanenhaus and Brown-Schmidt, 2008; McKinstry et al., 2008), and in responding to speech, humans can begin moving in response to an instruction *before* the end of the instructor's utterance (Hough et al., 2015a). Current interactive robots do not exhibit these capabilities partly due to unsuitable control algorithms which demote fluid interaction quality over other concerns. FLUIDITY puts interaction fluidity and the rapid recovery from misunderstanding with appropriate repair mechanisms at the heart of interactive robots, aiming to develop state-of-the-art incremental spoken language understanding (SLU) and continuous multi-modal HRI control algorithms.

In an example pick-and-place scenario where a user communicates with a robot to move objects to different target locations using their voice, adapting from Hough and Schlangen (2016), the capability of current systems is shown in the interval diagrams in Fig. 1 in the 'non-incremental' mode (A). The interval blocks represent the user's speech and robot's actions over time from left to right.

In 'immediately successful' interactions (Fig. 1 top), the robot processes an instruction like "put the red phone on the table" and understands the user's intention correctly the first time, picking up the user's intended object. Due to the uncertainty caused by the robot's sensors (Kruijff, 2012), the robot needs confirmation from the user through utterances like "yes" or "go ahead" before completing the action to achieve its goal - in mode (A) this is safe, but cumbersome. In 'recovery from miscommunication' scenarios (Fig. 1 bottom) where the incorrect object is initially picked up and the user *repairs* the robot's actions with utterances such as "No! The other red phone." In mode (A), such an utterance cannot be recognized as a repair until the robot has stopped moving. Once the repair

[1]FLUIDITY in simulated human-robot interaction with speech interfaces. UKRI EPSRC grant: EP/X009343/1 project website: https://fluidity-project.github.io/.

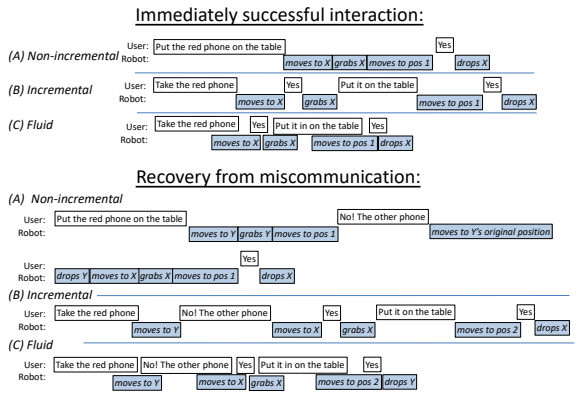


Figure 1: Fluidity in interaction from a non-incremental approach to speech processing (A) up to fluid processing. Modes (B) and (C) have incremental processing and in the fluid setting (C), robot actions start earlier as user feedback utterances can start earlier, as the robot constantly monitors and interprets relative to its actions.

is interpreted, not only must the current incorrect action be 'undone' but the new action must then be carried out in full, resulting in long periods of waiting. The ability to recognize intentions only from complete commands mapping to complete goals severely limits the fluidity of the interaction.

Improvement is possible in mode (B), an *incremental* mode, taking inspiration from (Kempson et al., 2001; Schlangen and Skantze, 2011; Purver et al., 2011; Eshghi et al., 2015; Hough et al., 2015b; Kennington and Schlangen, 2015; Madureira and Schlangen, 2020) and others in computational semantics focused on incrementality. Here, while turn-taking still happens in a half-duplex fashion with no overlap between human speech and robot motion, opportunities for confirmation or repair arise after shorter bursts of speech. This is possible by the robot predicting parts (increments) of the user's overall goal as speech arrives into the system word-by-word, such as predicting the target object to be picked up before predicting the target location. The 'recovery from miscommunication' scenarios show the benefit of incremental processing in situations of repair, as partially incorrect action plans can be corrected early and substantially reduce task completion time.

In the fluid mode (C), speech processing is also incremental, however the system goes *beyond incrementality*, allowing *full-duplex interaction* where concurrency of human speech and robot motion is reasoned with appropriately using *continuous intention prediction*. The robot can begin moving as soon as it is sufficiently confident about the user's goal and it can interpret confirmations and repairs during its movement appropriately, allowing it to complete correct actions more quickly and change course immediately in the middle of its initially selected action if corrected, leading to faster task completions in both scenarios. We also predict the more fluid the interaction, the more this behaviour will be perceived as natural, intelligent and likeable, building from Hough and Schlangen (2016).

3 Aims and objectives

The FLUIDITY project will investigate the *automatic measurement and improvement of fluidity in HRI*. With respect to Fig. 1, the aim is to move away from interaction as it happens in current systems in the non-incremental mode (A) to modes (B), incrementally, and finally, (C), fluidly.

The project will also address the difficulty of developing interactive models with real-world robots. A key outcome, currently under development, is a toolkit for building and testing interactive robot models with human participants in a Virtual Reality (VR) HRI environment, concretely, the simulation of the University of Hertfordshire Robot House[2] with the *Fetch Mobile Manipulator*[3]. The environment will be used to collect Wizard-of-Oz data with participants as the basis for training our SLU and interaction management/control models and of interest to both dialogue and HRI researchers. To achieve fluid interaction, the project will use the data to give a robot with speech understanding capabilities the following abilities:

1. predict the user's intention from their speech and confidence in that prediction as quickly and accurately as possible when sufficiently confident, investigating DS-TTR (Purver et al., 2011; Eshghi et al., 2015) and incrementalized deep learning models (Madureira and Schlangen, 2020) for the SLU.
2. monitor its own motion and estimate the earliest point that its own intention has become recognized by, or 'legible' to the user in the sense of (Dragan et al., 2013), whilst moving.
3. use abilities 1 and 2 in parallel to control its interactive behaviour appropriately, including repairing goals, to allow fluid interaction in both the virtual and real-world settings.

[2] https://robothouse.herts.ac.uk/

[3] https://www.zebra.com/us/en/products/autonomous-mobile-robots.html

References

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, London, UK. ACL.

Julian Hough, Iwan de Kok, David Schlangen, and Stefan Kopp. 2015a. Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL)*, pages 86–94.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015b. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings IWCS 2015*, London, UK. ACL.

Julian Hough and David Schlangen. 2016. Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles. ACL.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell, Oxford.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the ACL. ACL.

Geert-Jan M Kruijff. 2012. There is no common ground in human-robot interaction. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–374.

Chris McKinstry, Rick Dale, and Michael J Spivey. 2008. Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1):22–24.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th IWCS*, Oxford, UK.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1).

Michael K Tanenhaus and Sarah Brown-Schmidt. 2008. Language processing in the natural world. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):1105–1122.

VON NEUMIDAS: Enhanced Annotation Schema for Human-LLM Interactions Combining MIDAS with Von Neumann Inspired Semantics

Andrea Martinenghi and **Cansu Koyuturk** and **Simona Amenta** and **Dimitri Ognibene**
Department of Psychology
University of Milano Bicocca, Milan, Italy
[a.martinenghi1;c.koyutuerk]@campus.unimib.it; [simona.amenta;dimitri.ognibene]@unimib.it

Martin Ruskov
University of Milan, Italy
martin.ruskov@unimi.it

Gregor Donabauer and **Udo Kruschwitz**
University of Regensburg, Germany
[gregor.donabauer,udo.kruschwitz]@ur.de

Abstract

LLM-based chatbots represent a significant milestone as the initial point of interaction between artificial intelligence and the general public. These chatbots offer greater flexibility compared to traditional chatbots, yet their behavior deviates notably from human interaction patterns. Current annotation schemas may not be adequately suited to capture this unique interaction paradigm. In this paper, we propose a novel annotation method designed to annotate interactions between ChatGPT and users of varying expertise levels engaged in complex tasks. Our approach builds on the MIDAS annotation framework, introducing an additional semantic layer inspired by the Von Neumann base operation set. This layer provides detailed descriptions of requested behaviors and prompts, enhancing the granularity of interaction analysis. We aim to utilize this annotation scheme to explore the relationship between user interactions and their perception of AI, evaluate user expertise, and offer insights and suggestions for improved alignment and support.

1 Introduction

The direct availability of LLMs on the cloud and their advanced ability to perform tasks described in natural language have made AI systems accessible to the general public for the first time. However, these systems introduce new challenges in human-machine interactions. For example, their limited reasoning capabilities and language understanding can result in generating contextually inappropriate information (Tamkin et al., 2021) or restrain them from accurately interpreting context and user inputs (Bang et al., 2023). Besides, some users perceive ChatGPT as complex, struggle to understand its responses, and experience cognitive fatigue (Tiwari et al., 2023). This phenomenon may be aggravated when users attribute human-like traits to AI systems (Antonenko and Abramowitz, 2023) and create prompts that are either too broad or overly specific (Zamfirescu-Pereira et al., 2023), further complicating user interactions. Evaluating the behaviors of LLMs has received a lot of attention in the literature (Bommasani et al., 2023; Chang et al., 2024), however, methods have often focused on technical aspects rather than user interaction (Cremonesi et al., 2011). Also, previous studies on users' perceptions and experiences, combining different types of measures adopted for human-human (Fiske et al., 2018) or human rule-based chatbot interaction (Haugeland et al., 2022), found contrasting feedback from the same users (Theophilou et al., 2023).

With the exception of MIDAS: (Yu and Yu, 2019), available annotation schemes for domain-independent purposes are designed for human-human interactions. Understanding users' mental models, including their expectations and interaction strategies with LLM-based chatbots is crucial for enhancing their usability and support the users (Tiwari et al., 2023). Because of the evidenced specific features, we argue that even MIDAS (human-machine) is incomplete for human-LLM chats and offer a contribution for its adaptation. Given their unprecedented capabilities, LLM-based chatbots are often used for complex tasks (Braun and Matthes, 2021) that users, especially expert ones (Koyuturk et al., 2023), articulate in an imperative, program-like format, which is quite different from previous interactions with traditional chatbots or humans.

2 Related works

Pragmatic annotation is typically based on speech acts (for a comprehensive overview, see Horn and Ward, 2004). These are often adapted to the context, like in the game Catan (Asher et al., 2016,

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

Martinenghi et al., 2024). One of the most influential domain-independent annotation systems is Dialogue Act Markup in Several Layers (DAMSL; Core and Allen, 1997). DAMSL introduced a distinction between *Forward-looking* (e.g., questions) and *Backward-looking* (e.g., answers) acts. Together with the classes *Communicative Status* and *Communicative Level*, they take the annotation to a multi-dimensional domain which opens to multi-labeling.

The necessity for multi-dimensional annotations was later stressed by Popescu-Belis (2005) and Bunt and Romary (2004). This led to the design of DIT++ (Bunt, 2009), a taxonomy developed from the Discourse Interpretation Theory (DIT; e.g., Bunt, 1994) with elements from DAMSL. In turn, DIT++ served as a basis for ISO 24617-2 (Bunt et al., 2020), which inherited nine of its 10 dimensions and which includes specification of Dialogue Act Markup Language (DiAML). Recently, this annotation scheme was tested on conversations with AI agents in a doctor-patient setting (Bunt and Petukhova, 2023).

A multi-dimensional approach was adopted by Machine Interaction Dialogue Act Scheme (MIDAS; (Yu and Yu, 2019)). Like DAMSL and ISO, MIDAS is an independent-task annotation structure, but unlike them, it was specifically designed for human-machine conversations. It consists of two 2 trees: *Semantic Request* and *Functional Request*. Semantic Request is subdivided into the classes *Initiative* (Question, Command) which traces DAMSL's Forward-looking category, and *Responsive* (Opinion, Statement non-opinion, Answer) which traces the Backward-looking's. These two trees allow to track introduction of new topics as well as discourse level coherence.

3 VON NEUMIDAS

Our approach builds on MIDAS by introducing a new layer inspired by the first Von Neumann architecture for programmable computers (Von Neumann, 1993). This new dimension is an enhancement of the directive speech acts which aims to describe (1) relevant aspects specific to the human-LLM interaction and (2) failures (disagreements) of pragmatic or semantic nature.

A first categorization (*Command Type*) captures the type of instructions given to the agent through 4 classes. The classes *Input Operation* and *Output Operation* describe how the agent should handle incoming input, and how it should translate into actions or outputs. As sometimes users prompt an LLM to set its behavior in a specific way (i.e., "Act like a teacher") we introduce the class *Set state* for these situations. Other times, LLM's behavior is ordered to be conditional (i.e., "Stop when I ask why"): we use the class *Control*. We allow, for each directive speech act (MIDAS questions and command) at most two Command Type labels.

A second categorization serves as specification for the first categorization with the goal to track references between contextual information, thus creating a framework to highlight failures. The class *Roles* (Addressee, Executor) describe the direction of the action. The class *Links* (Points to, Points back to) outline the statements where the action is performed or where it was requested. The class *Consistency* evidences the matching between the argument of the request and the response (Semantic agreement) and between semantic requests and the participants' roles (Pragmatic agreement).

The main contribution of this scheme is the opportunity to identify errors (semantic and pragmatic disagreements) by leveraging discourse features belonging to the traditional studies on pragmatics, bridging them with a computational view of LLM's behaviors. In particular, the classes *Links* and *Consistency* offer a simple yet clear notation of these occurrences.

4 Conclusions

LLM-based chatbots have introduced the general public to new tools that 'actually do things just with words', i.e. perform complex tasks described in natural language and produce complex natural language output (Brown et al., 2020), and not only. However, they can show unexpected and/or computer-like behaviors and may require the user to adapt the interaction style to fulfill the desired goals (Koyuturk et al., 2023). Understanding the difficulties of the users and the errors of the chatbots requires a multi-level analysis of their interactions. And while LLMs have general difficulty with pragmatics (Chan et al., 2023; Martinenghi et al., 2024), in these complex tasks, where they often receive program-like inputs, it is the interaction between semantics and pragmatics that is more difficult to track. Current, annotation schemes do not capture this element. Our suggestion to deepen MIDAS' capabilities to adapt to LLMs' usage peculiarities offers a novel contribution to the field.

References

Pavlo Antonenko and Brian Abramowitz. 2023. In-service teachers'(mis) conceptions of artificial intelligence in k-12 science education. *Journal of Research on Technology in Education*, 55(1):64–78.

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).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Daniel Braun and Florian Matthes. 2021. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 93–99, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Harry Bunt and Volha Petukhova. 2023. Semantic and pragmatic precision in conversational ai systems. *Frontiers in Artificial Intelligence*, 6:896729.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The iso standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. European Language Resources Association (ELRA).

Harry Bunt and Laurent Romary. 2004. Standardization in multimodal content representation: Some methodological issues. In *4th International Conference on Language Resources and Evaluation-LREC'04*, pages 28–p.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Vittorio Papadopoulos, and Roberto Turrin. 2011. Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part III 13*, pages 152–168. Springer.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

Isabel Kathleen Fornell Haugeland, Asbjørn Følstad, Cameron Taylor, and Cato Alexander Bjørkli. 2022. Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies*, 161:102788.

Laurence R Horn and Gregory L Ward. 2004. *The handbook of pragmatics*. Wiley Online Library.

C Koyuturk, M Yavari, E Theophilou, S Bursic, G Donabauer, A Telari, A Testa, R Boiano, A Gabbiadini, D Hernandez-Leo, et al. 2023. Developing effective educational chatbots with chatgpt prompts: Insights from preliminary tests in a case study on social media literacy. In *31st International Conference on Computers in Education, ICCE 2023-Proceedings*, volume 1, pages 150–152. Asia-Pacific Society for Computers in Education.

Andrea Martinenghi, Gregor Donabauer, Simona Amenta, Sathya Bursic, Mathyas Giudici, Udo Kruschwitz, Franca Garzotto, and Dimitri Ognibene. 2024. LLMs of catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multiparty dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 107–118, Torino, Italia. ELRA and ICCL.

Andrei Popescu-Belis. 2005. Dialogue acts: One or more dimensions. *ISSCO WorkingPaper*, 62:1–46.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Emily Theophilou, Cansu Koyutürk, Mona Yavari, Sathya Bursic, Gregor Donabauer, Alessia Telari, Alessia Testa, Raffaele Boiano, Davinia Hernandez-Leo, Martin Ruskov, et al. 2023. Learning to prompt in the classroom to understand ai limits: a pilot study. In *International Conference of the Italian Association for Artificial Intelligence*, pages 481–496. Springer.

Chandan Kumar Tiwari, Mohd Abass Bhat, Shagufta Tariq Khan, Rajaswaminathan Subramaniam, and Mohammad Atif Irshad Khan. 2023. What drives students toward chatgpt? an investigation of the factors influencing adoption and usage of chatgpt. *Interactive Technology and Smart Education*.

John Von Neumann. 1993. First draft of a report on the edvac. *IEEE Annals of the History of Computing*, 15(4):27–75.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.

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.

A Annotation Tables and Examples

Table 1: MIDAS extended

Semantic	Class	Labels	Example(s)	Von Neumann
Initiative	Question	Factual question	What time is time?	full
		Opinion question	What's your favorite book?	full
		Yes-no question	Do you like pizza?	full
	Command	Task command	Let's talk about the immigration policy	full
		Invalid command	Cook food for me	
Responsive	Opinion	Appreciation	That's cool; that's really awesome	back link
		General opinion	Dogs are adorable	back link
		Complaint	What are you talking about; you didn't answer my question	back link
		Comment	A: My friend thinks we live in the matrix **B1: She is probably right**	back link
	Statement non-opinion	Statement non-opinion	I have a dog named Max	back link
	Answer	Other answer	I don't know; i don't have a favorite;	back link
		Positive answer	Yes; Sure; I think so; Why not	back link
		Negative answer	No; Not really; Nothing right now	back link
Functional				
	incomplete	Abandon	So uh; I think; Can we	
		Nonsense	He all out	
	social convention	Hold	Let me see; Well	
		Opening	Hello my name is tom; Hi	
		Closing	Nice talking to you; Goodbye	
		Thanks	Thank you	
		Thanks response	You're welcome -**NOTE: Not original from MIDAS, added by us**	
		Back-channeling	Uh-huh; (A: I learned that ...) B:Okay/Yeah/Right/Really?	
		Apology	I'm sorry	
		Apology response	That's all right	
		Other		

Table 2: Von Neumann Parameters

Command Type	Description
Control Instructions	These are analogous to directives in a dialogue that determine the flow of conversation based on conditions—similar to conditional or loop instructions in programming, which control the flow of execution based on conditions.
State Changes	In the context of a dialogue, this could relate to changes in the conversation state or context, analogous to instructions in a program that modify the state of memory.
Output Operations	These instructions deal with what the dialogue system outputs, akin to a computer program writing data to an output device. In dialogues, this would translate to the responses or actions the dialogue system produces as outputs to the user or another system.
Input Operations	These are instructions on how the dialogue system should handle incoming data or user inputs, similar to a computer reading data from an input device.

Table 3: Additional Von Neumann Annotation Fields

	Field	Description	Example
Roles	addressee [participant]	who is the addressee of the action	me in ask me some question"
	executor [user or GPT]	who has to perform the action	GPT in "ask me some question"
	about	what must be performed or discussed	democracy in "tell me about democracy"
Links	points to	statement where this action is performed	
	points back	statement where this action was requested or introduced	
Consistency	semantic agreement	matching about argument of the request and response	
	pragmatic agreement	matching between semantic requests type or participants roles	

Table 4: Example of annotation: new user

Chat	id	MIDAS	Von Neumann Command type	addressee [user or GPT]	executor [user or GPT]	about [row index or keywords from the text]	semantic agreement	pragmatic agreement	point to	points back
Hi there!	user	opening								
Hello! How can I assist you today?	assistant	opening, task command								
I would like you to test my knowledge on Privacy and social media please!	user	statement								
Please ask me some questions on the subject to see how much I know.	user	task command	output	user	gpt	row 5 (privacy)			5	
Of course! I'd be happy to test your knowledge on privacy and social media. Here are a few questions for you:	assistant	positive answer								
1. What is the definition of privacy in the context of social media?	assistant	factual question								5
2. What are some potential risks to privacy when using social media?	assistant	factual question								5
3. How can you protect your privacy on social media platforms?	assistant	factual question								5
4. What is the purpose of privacy settings on social media platforms?	assistant	factual question								5
5. What are some best practices for managing your privacy on social media?	assistant	factual question								5
Take your time to answer these questions, and I'll provide feedback on your responses.	assistant	task command, statement	output to GPT from USER	gpt	user	7-11		no	fail	
Could you please ask me one question at a time regarding privacy and social media to test my knowledge! Thanks	user	task command	output to user from gpt about row 5	user	gpt	5				
Certainly! Here's the first question: 1. What is the definition of privacy in the context of social media?	assistant	factual question								13
I dont know this. Please let me know the answer!	user	other answer, task command	output to user							

Table 5: Example of annotation: expert user

chat	id	MIDAS	Von Neumann Command type	addressee [user or GPT]	executor [user or GPT]	about [row index or keywords from the text]	semantic agreement	pragmatic agreement	point to	points back
act as a teacher with a sense of humor	user	task command	set state (operation)							
always wait for my answers	user	task command	input from user							
have a conversation with me	user	task command	input from user,output to user, control							
following these steps	user	task command	control,							
ask my name and adress me with that	user	task command	output to user							
ask me one short question to assess my learning	user	task command	output to user about row 8	user	gpt				33,35**	
wait for my answer	user	task command	input from user about row 8	user	gpt			No	34,35**	
give feedback about my answer	user	task command	output to user	user	gpt				36	
later explain the next point about how social media algorithms work	user	task command	output to user	user	gpt				37,38	
follow this loop until you explain all the aspects	user	task command	control							
then move to next	user	task command	control, set state (operation)							

Dialogue with LLaVA: does it "understand" the pragmatics of the MeetUp task?

Nikolai Ilinykh and Simon Dobnik
Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
name.surname@gu.se

Introduction Recently large language-and-vision models like BLIP-2 (Li et al., 2023) have achieved good performance on various multi-modal tasks. These models are initially pre-trained on a large number of image-text pairs to capture general multi-modal understanding and then fine-tuned for specific downstream tasks. Models like LLaVA (Liu et al., 2023) are fine-tuned with *prompts* that *instruct* the model to perform a task. Using these models for different tasks requires rephrasing the tasks into the specific language and format that the models understand. Visual dialogue is a very challenging multi-modal task ,and MeetUp (Ilinykh et al., 2019) is an example of such a task and dataset. In this task two players are placed in a virtual house environment represented as 2D images and must find each other. They can use a chat interface to communicate and execute commands to change images, i.e., move from one room to another. The collected dataset introduces many challenges for multi-modal models as they must consider both visual and textual history at each moment in the dialogue. In this paper we examine the performance of the large multi-modal model LLaVA (Liu et al., 2023) for the task of next utterance generation in MeetUp. This task was initially introduced and described as part of the Visual Dialogue Generation Challenge (Ilinykh and Dobnik, 2023). We prompt LLaVA with two different prompts which are structurally similar but vary in the game-relevant information that they include. By evaluating the quality of generated utterances and the model's perplexity in predicting human-generated utterances, we draw conclusions about LLaVA's ability to be used in the context of such visual dialogue task as MeetUp.

Prompting LLaVA for turn generation LLaVA (Liu et al., 2023) is a large transformer-based model designed to assist humans in completing various visual tasks. The model consists of two large pre-trained transformers: CLIP (Radford et al., 2021) and Vicuna (Chiang et al., 2023), allowing it to use their pre-trained knowledge. One of LLaVA's strengths is its strong performance in tasks involving conversation, detailed image descriptions, and complex reasoning. The model has been fine-tuned on automatically generated instruction-following data based on the image-text pairs from the MS COCO dataset (Lin et al., 2014). For the MeetUp task we design two prompts (shown in the Appendix) following the style of the instructions used to fine-tune LLaVA. Prompt **A** describes the input to the model along with extra information about the game, such as "the players are trying to find each other". Prompt **B** describes *only* the input to the model with minimal information about the game. We use either of these prompts to generate each next utterance in the MeetUp dialogues. The model receives an image showing the rooms visited by the players up to the current timestep as visual input. On the textual side the model is provided with the current chat history. We tested LLaVA on predicting 5695 extracted turns that contain utterances. We did not explore other types of turns such as those with navigation commands or private chat messages to the game bot. We measure the quality of model-generated messages by comparing them with the ground-truth messages using several classic n-gram based metrics (e.g., BLEU, ROUGE, and METEOR, for a survey see Sai et al. (2022)) as well as BERTScore (Zhang et al., 2020).

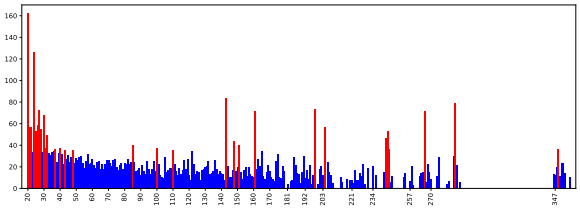
As shown in Table 1 prompt **A** leads to slightly higher scores in automatic evaluation than prompt **B**. All scores are low and close to each other, indicating that the generated utterances are very distant from the human ground-truth. Higher BERTScore values, which are closer to 1.0 (indicating the highest cosine similarity), show that the generated descriptions are very similar to those generated by humans in terms of their semantics. One possible explanation for this, which requires additional

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

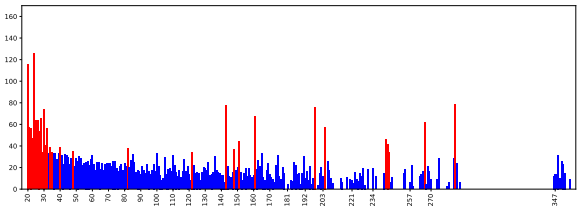
Prompt type	BLEU-1	BLEU-4	ROUGE	METEOR	BERTScore
Prompt **A**	6.46	0.12	8.10	15.65	0.78
Prompt **B**	6.40	0.08	7.96	16.46	0.79

Table 1: Automatic evaluation of quality of generated turns given different prompts. The scores are averaged.

testing, is that the generated descriptions are thematically within the domain of the dataset (e.g., describing images), but it is unclear how contextual and correct these descriptions are.



(a) Using prompt **A**.



(b) Using prompt **B**.

Figure 1: PPL score (vertical axis) against messages per turn (horizontal axis, visualised as turn id). The scores per turn are averaged across all dialogues. 20 turns with highest PPL score are highlighted in red.

We calculate the perplexity of the model against human-generated utterances. Perplexity (PPL) is defined as $\exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(w_i \mid w_1,\ldots,w_{i-1})\right)$, where N is the number of words in the sequence and $p_\theta(w_i \mid w_1,\ldots,w_{i-1})$ is the probability of the i-th word given the preceding words according to the model parameterised by θ (LLaVA). According to Figure 1, the model is most uncertain about predicting ground-truth utterances at the beginning of dialogues, with occasional high uncertainty towards the end of the dialogue. An excerpt from the MeetUp corpus in Example (1) with specified turns shows that these parts typically include greetings ("What's up" in t-21), negotiation of game strategies, and planning ("Oh k let me look for it" in t-29, other examples in t-107 and t-120), while the middle of the dialogues includes more descriptions of visual content ("a stand alone sink on the left" and others in t-36–t-50). MeetUp dialogues also contain a larger number of

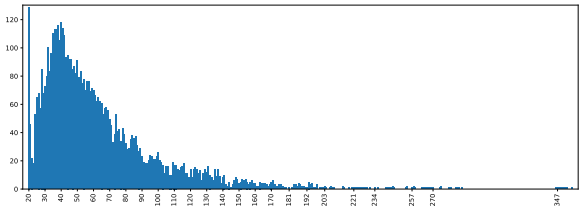


Figure 2: The number of turns with utterances. Turn ids are shown on the horizontal axis, the number of instances of each turn is displayed on the vertical axis.

utterances describing visual content, as indicated by Figure 2, which shows that most of the turns with text messages appear between turns 30 and 50. The data indicates that the model cannot understand such parts in MeetUp dialogues, which are not directly related to its visual content. One way to improve the model's understanding of the game's information and instructions is through prompt engineering. However, there is no clear evidence that changing the prompt affects the perplexity of the model, as both prompts result in similar average perplexity across turns (22.87 for prompt **A**, 22.51 for prompt **B**). Future work should explore other prompts.

Conclusion We explore prompt engineering for large language-and-vision models in the complex domain of visual dialogue tasks. Our analysis shows that LLaVA can be used to generate utterances in collaborative visual dialogue tasks such as MeetUp. Future work will examine how much game-relevant information prompts should include or if simply describing the input to the model and asking it to "generate a next response, given the input" is sufficient. We will also focus on the evaluation component of generated utterances by examining characteristics relevant to different games in a dialogue.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-

source chatbot impressing gpt-4 with 90%* chatgpt quality.

Nikolai Ilinykh and Simon Dobnik. 2023. The VDG challenge: Response generation and evaluation in collaborative visual dialogue. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 23–30, Prague, Czechia. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! A corpus of joint activity dialogues in a visual environment. *CoRR*, abs/1907.05084.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix: LLaVA prompts The prompts were designed with two goals. First, we aimed for prompts similar in structure to those used by LLaVA. Second, we created one prompt that simply describes the model's input with basic game context and another that provides more game-relevant information. The differences between the prompts below are highlighted in bold.

Prompt **A**:

> You are a helpful language and vision assistant. You see a chat between two people, A and B. **They are playing a game in which they are trying to find each other in a house.** What you see are the pictures of each **room** they have visited. The **rooms** visited by person A are shown in the top row, and the **rooms** visited by person B are shown in the bottom row. Pictures in each row are arranged in sequence from left to right, representing the order in which they were taken. Person A is currently **in the room** shown in the rightmost picture from the top row, and person B is currently **in the room** shown in the rightmost picture from the bottom row. A and B are having a chat and are trying to ensure that they are in the same room, i.e., **they have to see the same picture**. Each player does not see what the other player sees. Sometimes the chat is empty, which means that the players have not written any messages yet.
> What do you think is the next message based on the information you have about the game, the players, the rooms they have visited, and their chat?
> CHAT:

Prompt **B**:

> You are a helpful language and vision assistant. You see a chat between two people, A and B. You also see **pictures**. **Pictures** seen by person A are shown in the top row of the image, and **pictures** seen by person B are shown in the bottom row. Pictures in each row are arranged in sequence from left to right, representing the order in which they were seen. Person A is currently **seeing the rightmost picture** from the top row, and person B is currently **seeing the rightmost picture** from the bottom row. A and B are having a chat and are trying to ensure that they **see the same picture**. Each person does not see what the other person sees. Sometimes the chat is empty, which means that A and B have not written any messages yet.
> What do you think is the next message based on the information you have about the situation, A, B, and pictures they have seen, and their chat?
> CHAT:

Excerpt from a MeetUp dialogue

(1)

t-20	B: What's up
t-21	A: hi
t-28	A: i have found a bathroom
t-29	B: Oh k let me look for it
t-36	A: it has white bathtub in the back of the room, white shower curtain with blue patterns
t-41	A: a stand alone sink on the left
t-50	A: there is tile on the wall with small squares ranging in color between white and brown
t-51	B: I think I found it
t-52	B: toilet
t-53	B: towel rack
t-54	A: no i dont think i see a towel rack
t-69	B: oh
...	...
t-106	A: lets meet at the bathroom with pink towels
t-107	A: it is more easily identifiable
t-120	B: ok im there

Accounting for *comment*-questions

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

Abstract

We show that different readings of French *comment* 'how (come)' interrogatives in film scripts vary in the likelihood of being followed by an account in the same turn. REASON uses, which aim to resolve a conflict between the speaker's expectations and the situation depicted by the prejacent, are most likely to be followed by an account. METHOD uses, which ask for ways to realize the prejacent, come second, followed by MANNER. MEANS uses, which often feature verbs of speech, are least likely to be accounted for. We argue that REASON questions are more intrusive than other readings because they can deny a discourse commitment or indicate violated presuppositions.

1 Introduction and overview

An account in interaction is defined as "a statement made by a social actor to explain unanticipated or untoward behavior" (Scott and Lyman, 1968, 46). Baranova and Dingemanse (2016, 642) distinguish between "providing reasons and providing accounts in interaction", viewing "reasons as a more general phenomenon that involves causal statements for any behaviour. An account is a subtype of a reason used in the context of a delicate action". Asking a question can be a delicate action, intrusive (Farkas, 2022; Kaneko, 2024), and even impolite when targeting assertions or presuppositions by the interlocutor (Brown and Levinson, 1987, 102). We hypothesize that REASON questions, which aim to resolve a conflict between the speaker's expectations and the situation depicted by the prejacent (the proposition conveyed by the interrogative clause without the operator), are particularly intrusive and therefore more likely to be followed by an account. An instance of these are French *comment* 'how (come)' questions, as in (1) (Fleury and Tovena, 2018, 2021; Fleury, 2021).

(1) OSCAR Mais **comment** on peut perdre son clitoris ?! Ça se perd pas, ce truc-là !
'*But how can you lose your clitoris?! You can't lose it, that thing!*'
LOUISE J'ai plus aucun plaisir, plus rien.
'*I don't get any pleasure any more, none.*'
(Tout le Plaisir est pour Moi)

In (1), the turn continuation (Sidnell, 2012; Couper-Kuhlen, 2012) after the REASON *comment*-interrogative can be seen as an account. We use observations from French film scripts (fictionalized interaction) to explore the relationship between *comment*-interrogatives and turn continuations with accounts. Given the repeated empirical finding from different quantitative measures that scripted dialogue for audio-visual entertainment is a "close approximation" (Levshina, 2017, 311) of unscripted and informal conversations and "successfully imitates" (Bednarek, 2018, 124) its linguistic characteristics, we expect our findings to be replicable with natural conversational data. We find that accounts are frequently provided in turn continuations after *comment* questions, particularly those inquiring about REASONS (1). They occur less frequently in turn-continuations after METHOD (2) and MANNER questions (3), and infrequently after uses of *comment* that ask for the MEANS to do or say something (4) and after OTHER uses such as clarification requests (5).

(2) RACINE Et sinon de l'alcool, vous en avez? '*Or alcohol, do you have any?*'
PEIGNE L'alcool c'est interdit dans le camp. '*Alcohol is forbidden in the camp.*'
RACINE **Comment** je fais si y'a rien ici ? On l'opère au couteau sans anesthésie ? Je vais le tuer votre mec.
'*What do I do if there's nothing here? Cut him open without any anaesthetic? I'm going to kill your boy.*' (Nos résistances)

(3) DJAMILA [...] je peux leur payer [...]
'*[...] I can pay them [...]*'

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11-12, 2024, University of Trento.

ANNE Et la place de votre copain, de votre partenaire, vous la voyez **comment**, alors ? Parce que vous dites : « Je ».
'*And the role of your boyfriend, of your partner, how do you see it, then? Because you say: "I".*' (Les Bureaux de Dieu)

(4) ANNE En fait votre mère elle, elle bloque sur le fait que vous puissiez vous retrouver enceinte ou bien que vous ayez des relations ? '*So does your mother have any reservations about you getting pregnant or having relationships?*'
DJAMILA Je sais pas, j'ai jamais discuté avec elle. '*I don't know, I've never spoken to her.*'
ANNE Elle n'est pas, **comment** dire ? Vous êtes d'origine… '*She's not, how can I put it? You're from…*'
DJAMILA Algérienne. '*Algerian.*'
(Les Bureaux de Dieu)

(5) MARTHA Bientôt, quand on sera en…
'*Soon, when we're in…*'
LÉNA Bientôt **comment**?
'*Soon what?*' (Calamity)

2 Corpus study

2.1 Data

We created a corpus based on 99 film scripts from the website *Lecteurs Anonymes*. We extracted and annotated 626 uses of *comment*, categorizing them according to their respective readings and whether the turn was continued after the sentence or turn constructional unit that contained *comment*. We identified 140 accounts in turn continuations. The other turns with *comment* either changed topic or continued without directly accounting for the *comment*-interrogative, as in (4).

2.2 Results

Table 1 shows the distribution of accounts according to the readings of *comment*. Figure 1 displays the adjusted standardized residuals (ASRs) of a χ^2 test on this table (bar width indicates n). While only two tendencies reach statistical significance (ASR>1.96 for p<.05), more tendencies are visible. *Comment*-interrogatives that ask for REASONS are followed by accounts as turn continuations significantly more frequently than the other readings. This is particularly true compared to the frequent MEANS uses of *comment*, which often involve self- and other-addressed requests for ways of referring to something (formulations) or to someone (names). MANNER and METHOD uses are more heterogeneous and pattern between these two tendencies.

Table 1: Accounts by reading of *comment*

	other	man.	rea.	mea.	meth.	Sum
acc.	2	66	23	20	29	140
no acc.	19	227	42	114	84	486
Sum	21	293	65	134	113	626

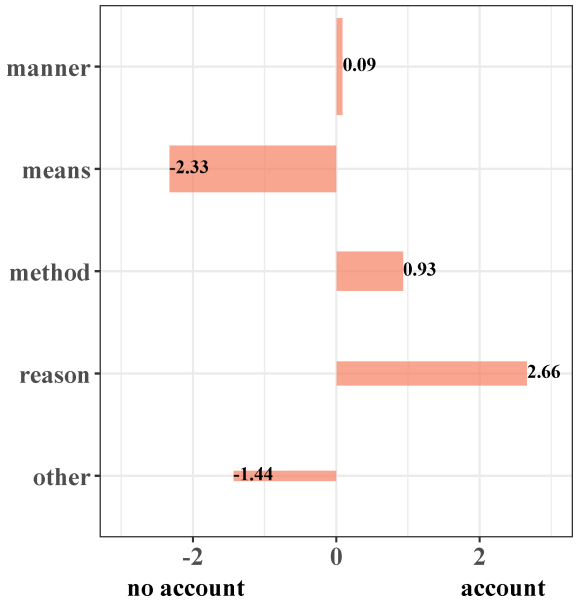


Figure 1: ASRs of a χ^2 test on Table 1

3 Interpretation and conclusion

We have shown that METHOD and REASON questions are prone to be followed by turn continuations that provide reasons for the request itself. This is indicative of reconfigurations of the context state (i.e., Table, Projected Set, Commitment Sets, Common Ground; Farkas and Bruce 2010), with such uses of *comment* often targeting assumptions related to the prejacent previously held to be part of the Common Ground, or signalling the speaker's refusal to update their Commitment Set until reasons are provided that render the prejacent consistent with the Common Ground. The scarcity of accounts for MEANS uses of *comment* might be due to the prevalence of verbs of speech, as in (4), that tend to be self-addressed questions (no interrogative flip) or non-intrusive questions (the hearer's answer need not solve the issue) (Farkas, 2022, 316). Future research needs to explore this connection in greater detail.

Acknowledgments

This work has been partially funded by the Hubert Curien Partnership of Campus France with the German Academic Exchange Service (DAAD), Grant ID 57701768.

References

Julija Baranova and Mark Dingemanse. 2016. Reasons for requests. *Discourse Studies*, 18(6):641–675.

Monika Bednarek. 2018. *Language and television series: A linguistic approach to TV dialogue*. The Cambridge applied linguistics series. Cambridge University Press, Cambridge.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*, volume 4 of *Studies in interactional sociolinguistics*. Cambridge University Press, Cambridge.

Elizabeth Couper-Kuhlen. 2012. Turn continuation and clause combinations. *Discourse Processes*, 49(3-4):273–299.

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

Donka F. Farkas and Kim B. Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics*, 27(1):81–118.

Damien Fleury. 2021. *Questions en comment de raison. La révision des attentes du locuteur*. Ph.D. thesis, Université de Paris.

Damien Fleury and Lucia M. Tovena. 2018. Reason questions with *comment* are expressions of an attributional search. In *Proceedings of 22nd workshop on the Semantics and Pragmatics of Dialogue (AixDial)*, pages 112–121.

Damien Fleury and Lucia M. Tovena. 2021. On the pragmasemantics of a high adjunct wh-word. In Chad Howe, Pilar Chamorro, Thimoty Gupton, and Margaret Renwick, editors, *Theory, data, and practice. Selected papers from the 49th Linguistic Symposium on Romance Language*, pages 85–106. Language Science Press. (in press).

Makoto Kaneko. 2024. Two past forms inducing conjectural or non-intrusive questions. *Open Linguistics*, 10(1):20220274.

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

Marvin B. Scott and Stanford M. Lyman. 1968. Accounts. *American Sociological Review*, 33(1):46–62.

Jack Sidnell. 2012. Turn-continuation by self and by other. *Discourse Processes*, 49(3-4):314–337.