SemDial 2023

MariLogue

Proceedings of the 27th Workshop On the Semantics and Pragmatics of Dialogue

Held at
University of Maribor
Faculty of Electrical Engineering and Computer Science / The Internet
August 16–17 2023

Andy Lücking, Chiara Mazzocconi & Darinka Verdonik (eds.)

University of Maribor

Faculty of Electrical Engineering
and Computer Science

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

MariLogue Website

https://mezzanine.um.si/en/conference/semdial-2023-marilogue/

MariLogue Endorsements

Preface

The University of Maribor, and the Faculty of Electrical Engineering and Computer Science (FERI) in particular, was honoured to host the 27th SemDial workshop. Maribor is the second-largest city of Slovenia, a small country situated between the Alps, the Adriatic Sea and the Pannonian Plain. If you come to Slovenia, you can visit these three very diverse landscapes in just one day. As diverse as Slovenia's landscape is, so is its language. With only 2 million of speakers, Slovene is known for a record-breaking number of dialects – it is not uncommon that Slovenes from different regions to have difficulties understanding each other when speaking their local dialects. For a linguist interested in spoken language it is a real treasure and a pleasure to explore this linguistic diversity, and for a speech technologist it is a real challenge to address it properly. FERI has been dedicated to the development of language technologies and the creation of language resources since the 1990s. Together with the Slovene CLARIN.SI consortium, we are pleased today to provide open language corpora and databases for the Slovene language created by our researchers. With the SemDial workshop, we were privileged to host researchers whose work brings inspiration and added enthusiasm to the study of language in all its dimensions and facets.

This year we received 25 full paper submissions, 12 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 15 submissions from a combination of recommended pre-accepted re-submissions of long papers and a further call for research in progress and short papers – 13 of these poster abstracts were presented. All accepted full papers and 12 poster abstracts (one has been withdrawn) are included in this volume.

We would like to extend our thanks to our Programme Committee members for their very detailed and helpful reviews.

MariLogue features three keynote presentations by Philippe Blache, Director of Research at the CNRS, *Laboratoire Parole et Langage*, Director of the *Institute of Language Communication and the brain*; Liesbeth Degand, Professor at Université Catholique de Louvain; Marko Robnik Šikonja, Professor at the University of Ljubljana, Faculty of Computer and Information Science, member of the *Laboratory for Cognitive Modeling*. We are honoured to have them in this year's SemDial and we thank them for their participation. Their contributions cover important and different areas of the study of dialogue. Abstracts of their contributions are included in this volume.

We would also like to thank our local organizers at Maribor University for chairing and bringing SemDial to such a special setting. We also thank Julian Hough and Ellen Breitholtz for core administrative support building up to the event, and to Julian Hough (again), Casey Kennington and Brielen Madureira for their support on converting the proceedings to the SemDial anthology format. Thanks to everyone who helped with all aspects of the organization.

And last, but not least, a special thank you to the authors and conference participants, whose contributions and participation make this an exciting SemDial – and proved once more why dialogue is the crux of the matter in the melting pot of linguistics-related disciplines and perspectives.

Andy Lücking, Chiara Mazzocconi and Darinka Verdonik

Maribor

September 2023

Programme Committee

Andy Lücking (chair)	Goethe University Frankfurt
Chiara Mazzocconi (chair)	Aix-Marseille University
Darinka Verdonik (chair)	University of Maribor
Jedediah Allen	Bilkent University
Maxime Amblard	University of Lorraine, LORIA
Špela Antloga	University of Maribor
Alex Berman	University of Gothenburg
Philippe Blache	CNRS & Université de Provence
Maria Boritchev	Polish Academy of Sciences
Ellen Breitholtz	University of Gothenburg
Harry Bunt	Tilburg University
Heather Burnett	CNRS & Université Paris Cité
Robin Cooper	University of Gothenburg
Simon Dobnik	University of Gothenburg
Kaja Dobrovoljc	University of Ljubljana, Jožef Stefan Institute
Kallirroi Georgila	University of Southern California
Jonathan Ginzburg	Université Paris Cité
Julian Hough	Swansea University
Christine Howes	University of Gothenburg
Julie Hunter	Lingora
Amy Isard	University of Hamburg
Casey Kennington	Boise State University
Pierre Lison	Norwegian Computing Center
Andy Lücking	Goethe University Frankfurt
Vladislav Maraev	University of Gothenburg
Chiara Mazzocconi	Aix-Marseille Université
Elin McCready	Dept. of English, Aoyama Gakuin University
Robert Mirski	The John Paul II Catholic University of Lublin
Bill Noble	University of Gothenburg
Valeria de Paiva	Samsung Research America and University of Birmingham
Massimo Poesio	Queen Mary University of London
Matthew Purver	Queen Mary University of London / Jožef Stefan Institute
Kyle Rawlins	Johns Hopkins University
Hannes Rieser	Bielefeld University
Robert Ross	Technological University Dublin
David Schlangen	University of Potsdam
Gabriel Skantze	KTH Royal Institute of Technology
Matthew Stone	Rutgers, State University of New Jersey
Lucas Thorpe	Bogazici University
David Traum	Institute for Creative Technologies, USC
Shu-Chuan Tseng	Institute of Linguistics, Academia Sinica

Darinka Verdonik	University of Maribor
Grégoire Winterstein	Université du Québec à Montréal

Local Organizing Committee

Darinka Verdonik (chair)	University of Maribor
Špela Antloga	University of Maribor
Andreja Bizjak	University of Maribor
Simona Majhenič	University of Maribor
Aleksandra Zögling Markuš	University of Maribor

Table of Contents

Poster Abstracts

Invited Talks

Keynote 1

Deep and compositional, or shallow and direct: two routes and one loop for a new approach to language understanding in conversation

Philippe Blache
Laboratoire Parole and Langage (LPL-CNRS), Aix-en-Provence, France
Institute of Language, Communication and the Brain (ILCB), Marseille, France
philippe.blache@univ-amu.fr

Abstract

Language understanding is a complex task, integrating different sources of information, from sounds and gestures to context. However, in spite of its complexity, this process is extremely fast and robust, performed in real-time during conversations. Many studies have shown that this robustness and efficiency are made possible by different mechanisms: the ability to predict, the possibility of directly accessing entire pieces of meaning and the possibility to perform a "good-enough" processing, sufficient to access the meaning. These mechanisms, by substituting to the classical incremental and compositional architecture, facilitate access to the meaning. However, existing models do not explain precisely when these facilitation mechanisms are triggered and whether they inhibit or on the contrary work in parallel with the standard ones.

I propose in this presentation a new model integrating both facilitation and standard mechanisms by revisiting the different stages of the processing: segmentation of the input, access to the corresponding meaning in long-term memory and integration to the interpretation under construction. This architecture is based on different features: unique representation of linguistic objects (independently from their granularities), control of memory access (in particular thanks to search space reduction) and multiple-level prediction. This neuro-cognitive model provides a new framework explaining how deep and shallow mechanisms of language processing can cohabit. It is also a good candidate for explaining different effects of mismatch observed at the brain level.

Keynote 2

Discourse markers in interaction: distribution and functions

Liesbeth Degand
Université Catholique de Louvain,
Louvain-la-Neuve, Belgium
liesbeth.degand@uclouvain.be

Abstract

In this presentation I give an overview of the use of Discourse Markers (N= 1872) in a multi-genre corpus of spoken French. Contrasting their use in dialogic and monologic contexts, I will show how their distribution and function is impacted by their context of use. The focus will be on the relationship between frequency, polysemy and polyfunctionality of the markers, and how their (syntactic) form influences their functional distribution.

Keynote 3

Challenges in explaining machine learning models for text

Marko Robnik Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si

Abstract

The area of Explainable Artificial Intelligence has developed many approaches for the explanation of machine learning models. The most successful methods are based on counterfactuals, prototypes, and perturbation of inputs. Unfortunately, none of these approaches works well to explain large language models, which currently dominate natural language processing. The presentation will focus on challenges in using the most successful explanation methods for text classification, such as the interpretation of feature attributions, explanation of longer textual units, on-manifold vs. off-manifold explanations, unstable and uncertain explanations, and inadequate and unsystematic evaluation of explanation techniques. We will present possible solutions and outline a framework for more general explanation approaches.

Full Papers

Lexical entrainment on target words during task-oriented interaction in children with and without autism spectrum disorder

Joanna Kruyt
Institute of Informatics,
Slovak Academy of Sciences/
in Bratislava, Slovakia
FIIT, Slovak Technical University
in Bratislava, Slovakia
joanna.kruyt@savba.sk

Katarína Polónyiová and **Daniela Ostatníková**
Academic Research Center for Autism,
Institute of Physiology, Faculty of Medicine,
Comenius University in Bratislava, Slovakia

Štefan Beňuš
Institute of Informatics, Slovak Academy of Sciences in Bratislava, Slovakia/
Constantine the Philosopher University in Nitra, Slovakia

Abstract

One widely observed strategy that interlocutors use to facilitate mutual understanding during dialogue is the repetition of each other's words, or lexical entrainment. Despite being well-researched, the underlying mechanisms of the phenomenon are debated. Specifically, the role of social factors and theory of mind are contested. This study aimed to investigate the role of theory of mind and neurotype on lexical entrainment. We recruited children with and without autism spectrum disorder, asked them to complete a collaborative task with an adult, and measured how often they entrained to the experimenter on "dispreferred" terms. We administered tests to measure IQ, executive functioning, and theory of mind for each child. Our results suggest that neither neurotype (i.e. autistic or neurotypical) nor theory of mind score predict entrainment, but that increased executive functioning difficulty predicts lower entrainment. Additionally, gender seems to influence entrainment. Theoretical implications of these results are discussed.

1 Introduction

During dialogue, two interlocutors need to collaborate to ensure that they understand one another. Mutual understanding can be achieved through several strategies. One of these strategies is the tendency of interlocutors to behave more similarly over time. This tendency is often referred to as entrainment, though other terms such as alignment, convergence, or synchrony are also used. This paper focuses on entrainment at the lexical level, i.e. on similarity in word choice.

Though entrainment has been widely observed, the exact psychological mechanisms underlying the phenomenon are debated. Specifically the role of higher-order cognition, in particular mentalising or "theory of mind" (ToM), is a topic of discussion. Some theories of entrainment postulate that it is an automatic process that occurs through priming mechanisms (interactive alignment hypothesis, Pickering and Garrod (2004, 2013)), while another theory is based on the idea of audience design: interlocutors tailor their utterances to whomever they are talking to, and take into account their "common ground" or mutually shared knowledge, which required perspective-taking and ToM skills (common ground/audience design account, Clark and Marshall (1978); Clark and Murphy (1982)). Yet another theory hypothesises that entrainment occurs because interlocutors aim to emphasise or minimise social differences between themselves and the person they are interacting with (Communication Accommodation Theory, Giles et al. (1991)).

In other words, the role of social and higher-order cognitive factors in entrainment is unclear. One way to elucidate the role of these external and internal factors, is by investigating entrainment in a group of people that exhibits both social and cognitive differences compared to the general population. Autism spectrum disorder (ASD) is often said to involve both of these: individuals with ASD report struggling with friendships and romantic relationships more than their neurotypical (NT) peers (e.g. Bossaert et al., 2015; Taheri et al., 2016), and ASD is associated with differences in ToM processing (e.g. Baron-Cohen, 2000; Baron-Cohen et al., 1985; Tager-Flusberg, 2007). Investigating entrainment in individuals with ASD can further inform us about the relationship between ToM and entrainment. This study aims to compare lexical entrainment in children with ASD and their NT peers, to characterise any potential between-group

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

differences and to examine the role of ToM in entrainment.

2 Previous work

Entrainment in individuals with ASD has been investigated on several levels including syntax and lexical choice. Research suggests that individuals with ASD show similar levels of syntactic entrainment to individuals without ASD, both in experimental settings (Allen et al., 2011; Slocombe et al., 2013) and more naturalistic conversations (Hopkins et al., 2016). In terms of lexical entrainment, with which the present study is concerned, results from existing studies appear somewhat less consistent.

Lexical entrainment in individuals with ASD has been investigated using different methodologies: some studies focus on entrainment on target words, while others focus on overall lexical entrainment. Entrainment on target words is typically measured in a collaborative card-placing task during which an experimenter uses uncommon or "dispreferred" words to describe objects. Whether individual with ASD also adopt this dispreferred term is taken as a measure of entrainment to the experimenter. When such paradigms are used to measure lexical entrainment in individuals with ASD, results typically suggest that individuals with and without ASD do not show different entrainment patterns (e.g. Slocombe et al., 2013; Branigan et al., 2016; Hopkins et al., 2016). Importantly, conversations during such tasks are usually highly constrained, with predictable turn-taking and short turns. Du Bois et al. (2014) even refer to the speech during such structured tasks as "serial monologue" (p. 436) rather than dialogue, highlighting how such structured tasks do not resemble naturalistic, interactive conversation.

Rather than looking at entrainment on target words, some studies investigate overall lexical entrainment. Overall lexical entrainment in individuals with ASD is typically measured during more unstructured naturalistic conversations, where the proportion of shared vocabulary between participants is measured (e.g. Stabile and Eigsti, 2022; Patel et al., 2022; Fusaroli et al., 2023). The majority of studies that measured lexical entrainment on a more global level rather than on target words, typically during less restricted conversations, report significant between-group differences, with individuals with ASD exhibiting lower degrees of lexical entrainment.

The present study aims to combine approaches taken in previous experiments: we will measure entrainment on target words, but will record more naturalistic, task-oriented conversations, with less predictable turn-taking than traditional studies in which entrainment on target words is measured. We hypothesise that a less structured and less predictable task will increase the cognitive load of participants, as they have to spend cognitive resources on the task, as well as on predicting turn-taking and other communicative and social processes. We hypothesise that the cognitive load during a semi-structured conversation is higher for people with ASD than NT people due to their differences in social processing. Since increased cognitive load is associated with reduced entrainment (Abel and Babel, 2017), we expect any between-group differences in entrainment to be more salient during semi-naturalistic conversation than in a highly structured one. In line with existing research, we expect to find less entrainment on the lexical level in our participants with ASD than in their NT peers.

3 Methods

3.1 Participants

For this experiment we collaborated with the Academic Research Center for Autism (ARCA) in Bratislava, Slovakia. With their help, we recruited two groups of children who were native Slovak speakers and had normal to corrected sight and hearing: one group of children with (suspected) ASD (diagnosis was later confirmed through standardised diagnostic testing) and one group of NT children who did not have suspected ASD or other developmental disorders.

In total, we recruited 67 children (14F, 62M), of whom 41 were diagnosed with ASD (7F, 34F) and 35 were NT (7F, 28M). The mean age of all recruited children was 9.21 (±1.86). For further details on the demographic information and various test scores of both groups, see Table 1. Note that we did not include data from each child in the analyses due to some technical issues with our audio recordings.

All children suspected of having ASD underwent a comprehensive diagnostic procedure, consisting of the Autism Diagnostic Observation Schedule (ADOS, Lord et al. (2008)) and the Autism Diagnostic Interview (Revised, ADI-R). Furthermore, the Woodcock-Johnson test was administered to assess the intelligence quotient (IQ) of each child,

Table 1: Summary of demographic information and test scores for both groups of participants.

	ASD		NT		t-test
	mean (std)	range	mean (std)	range	p
Age	9.10 (1.71)	6.14 - 12.30	9.34 (2.04)	6.18 - 12.97	>0.05
IQ	96.80 (16.81)	52 - 131	105.8 (14.84)	67 - 134	<0.05
BRIEF	67.32 (9.61)	47 - 85	58.77 (12.81)	36 - 83	<0.01
ToM	8.29 (2.14)	2 - 14	12.17 (1.74)	8 - 15	<0.001

while the Comic strip task (Cornish et al., 2010) was employed to measure Theory of Mind (ToM) abilities. Additionally, information regarding the executive functions (EF) of each child was gathered through the Behavior Rating Inventory of Executive Function (BRIEF) questionnaire (Gioia et al., 2000), which was completed by the parent(s)/caregiver(s) of the children. All materials used in the study were in Slovak and all tests were administered by a trained clinician at PhD-level.

3.1.1 Maps task

To elicit semi-naturalistic, task-oriented conversation, we used the Maps task. During this task, an experimenter and participant are both given maps that differ slightly (see Figure 1). One map contains a pre-drawn route, and the goal of the task is for the "instruction giver" to explain this route to the "instruction follower", so that the instruction follower can replicate the route on their own map as closely as possible.

We edited the original maps to change the original landmarks to different objects. Our maps contained two types of objects: control objects, which had one clearly preferred lexical label (e.g. *orech* (walnut) in Figure 1), and target objects which had both a "preferred" and "dispreferred" label (e.g. for the picture of the orange in in Figure 1, the preferred term was *pomaranč* (orange) and the dispreferred term was *mandarínka* (mandarin)). In the present experiment, we aimed to see whether children would entrain to the adult experimenter on dispreferred terms for target objects.

We selected our target objects based on an online norming study, in which we distributed a survey that contained coloured pictures and asked children to answer the following two questions: "What is the first word you would use to describe this picture?" and "What other word would you use to describe this picture?". Based on these answers, we selected our control objects (words that had one clearly preferred term and no common dispreferred terms) and target objects (words that had a preferred term and a commonly provided less-preferred or dispreferred term).

To minimise any discomfort or distress for the children with ASD, we decided that the experimenter who completed the Maps task with them would be somebody they were familiar with. The experimenter was the clinician who administered the other tests. The experimenter was thus aware of whether the child received an ASD diagnosis or not, which could introduce experimenter bias. To mitigate the influence of such potential bias, we provided the experimenter with training and detailed instructions on how to act prior to the task. Importantly, the experimenter was instructed to always use the dispreferred word, and as a reminder, her maps had written labels indicating with which word she should use to describe each object (see Figure 1).

The Maps task consisted of different trials, and different target objects were used in these different trials (see Table 2). The maps in the first and last "real" (i.e. non-practice) trial contained all 8 target items. The maps in trial 2 and 3 contained half of the target objects (the same 4 target objects each). This allows for the comparison of entrainment on dispreferred terms that were repeated more often and more recently to entrainment on dispreferred terms that were mentioned less often and a longer time ago. The maps used in trials 1 and 4 were counterbalanced, as were the maps used in trials 2 and 3.

In a typical Maps task, roles of participants (i.e. instruction giver or follower) switch between every trial. Since many children with ASD struggle with executive functioning, and such constant rule-switching might be challenging for them, we decided to only have one role switch in the Maps task: during the first few trials, the child was the instruction follower, and during the last few trials, the child was the instruction giver. Each part started with a practice trial (see Table 2), so we could ensure that the children understood the task and got used to their roles.

Figure 1: Example of maps used in the Maps task (not true to size). These specific maps were used for Trial 2, during which the experimenter is the instruction giver and only half of the target objects are on the map (see Table 2). The objects that are on both maps are target objects. Objects that are only present on one map are control objects that only have one preferred term. The target objects on the experimenter's map are labeled with their dispreferred terms as a reminder for her to use only the dispreferred terms.

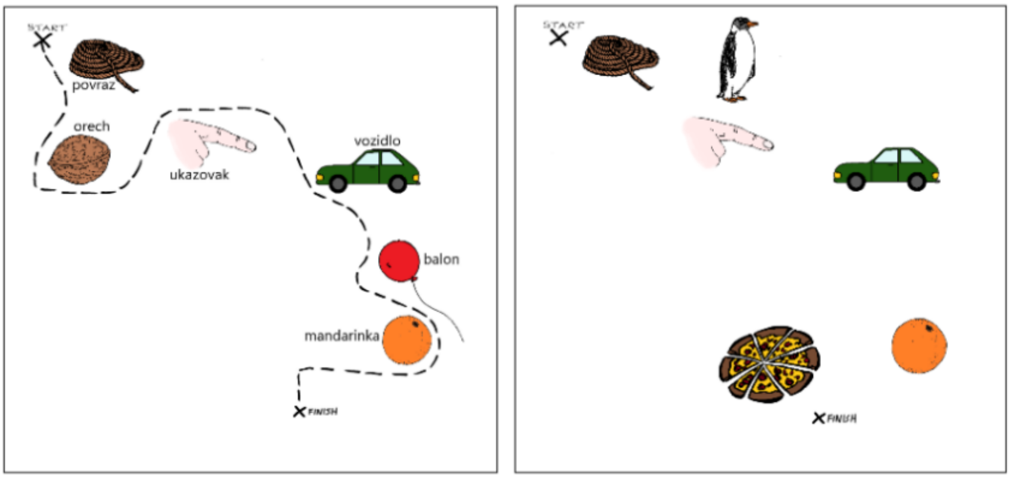


Table 2: Overview of Maps task trial structure.

Map	Instruction giver	Objects
Practice 1	Experimenter	Control only
Trial 1	Experimenter	All target objects
Trial 2	Experimenter	Half of the target objects
Practice 2	Child	Control only
Trial 3	Child	Half of the target objects
Trial 4	Child	All target objects

The Maps task was recorded with various microphones. All audio recordings were orthographically transcribed using Transcriber by experienced annotators who were native Slovak speakers. The subsequent transcriptions were transformed to Praat TextGrid format for further analysis.

3.2 Analysis

The Maps task elicits more naturalistic dialogue than many other tasks that have been used to research entrainment in children with ASD, and though this was an important goal of this project, it must be noted that it comes at a cost: the dialogue is unpredictable, and it is nearly impossible to predict how often each interlocutor will say a particular word. This applies to the child, but also to the experimenter. One could argue that if the experimenter says a dispreferred word more often, a child may be more likely to be primed to repeat it. For this reason, we included the number of times an experimenter said a particular word as a predictor of how many times the child used that same word.

To calculate lexical entrainment, we used the aforementioned TextGrids. One file was excluded because of technical issues, so 75 interactions were analysed. We used the Slovak *simplemma* lemmatiser to lemmatise all of each speaker's utterances and then counted how many times speakers said preferred and dispreferred words: per target word, we counted how many times the child used the preferred word and the dispreferred word, and how often the experimenter used the dispreferred word (since she had been instructed to always use the dispreferred word and never the preferred one). We then calculated the difference between the number of times the child said the preferred word and the dispreferred word, such that the number would be negative if the child dis-entrained, i.e. used the preferred word more often than the dispreferred. In the linear mixed effects model formula below, this value is represented as "diff child".

The linear mixed effects model formula we used to measure lexical entrainment was as follows, where "dispref exp" represents the number of times the experimenter used the dispreferred word:

diff child $\sim$ dispref exp + group + ToM + gender + trials + age + BRIEF + IQ + (1 | participant) + (1 | target item)

Group (i.e. ASD or NT), gender (M or F), ToM score, trials (i.e. whether the target stimuli was repeated in every trial or only in the first and last trials), age (in years), and BRIEF score are included as fixed effects, while participant and target stimulus are included as random effects. No interaction effects were included as this seemed to lower the AIC of the model and thus indicated that the addition of these interaction terms did not lead to a better fit. The lmerTest R package (Kuznetsova et al., 2017), which provides p-values via Satterthwaite's degrees of freedom method, was used to assess significance of effects.

4 Results

The intercept of the model used to predict lexical entrainment, which corresponds to the values of group = ASD, trials = target item was repeated only in first and last trials, ToM score = 0, gender = F, number of times experimenter said dispreferred word = 0, age = 0, IQ = 0, and BRIEF score = 0, is presented in the first row of Table 3.

Table 3: Effects in the LMEM constructed for the lexical entrainment analysis. Significant effects are indicated with an asterisk.

effect	beta	std	t	df	p
intercept	1.29	1.78	0.73	75.44	0.470
dispref adult	0.19	0.03	5.40	533.08	<0.001*
group	-0.56	0.41	-1.37	68.66	0.174
ToM	0.06	0.08	0.73	68.68	0.471
gender	0.78	0.37	2.10	68.76	0.039*
trials	-0.57	0.79	-0.74	5.99	0.489
age	0.06	0.08	0.75	69.26	0.454
BRIEF	-0.03	0.01	-2.05	68.81	0.044*
IQ	-0.01	0.01	-0.92	68.77	0.359

Within this model, effects that were found to be non-significant and negative were group, target item repetition, and IQ score. Effects that were found to be non-significant but positive were those of ToM score, and age (see Table 3).

Several effects were found to be significant (see Table 3). The effect of the number of times the experimenter said the dispreferred word (dispref exp) was found to be significant and positive, suggesting that more repetitions of a term by the experimenter led to higher lexical entrainment on that term. Additionally, general BRIEF score was found to have a significant and negative effect on children's lexical entrainment, suggesting that increased issues with executive functioning (reflected in a higher BRIEF score) was associated with lower degrees of lexical entrainment. Finally, gender was found to be significant, suggesting that boys showed more lexical entrainment than girls (beta is positive and the intercept is for gender = F).

5 Discussion

Our two recruited groups did not differ significantly in age, though significant between-group differences existed for IQ, ToM, and BRIEF scores (see Table 1). The latter two are in line with existing research that suggests that children with ASD perform less well on ToM tests than NT children (e.g. Baron-Cohen, 2000; Baron-Cohen et al., 1985; Tager-Flusberg, 2007), and typically struggle more with executive functioning than NT children as well (see Demetriou et al. (2018) for a meta-analysis). We tried to match our groups as closely on possible on age and approximate IQ, but this is a difficult task. The group of NT children we recruited had a significantly higher mean IQ score than our group of children with ASD. Though this is not ideal for a between-group comparison, we added IQ as a fixed effect in our model and did not find that it predicted entrainment.

The present study aimed to assess lexical entrainment on target words, but during less constrained conversations than most existing studies. The results of the our analysis suggest that group (i.e. NT or ASD), ToM score, target item repetition, IQ score, and age do not significantly predict the degree to which a child entrained to the experimenter on a dispreferred term. On the contrary, the number of times the experimenter repeated a word and a child's BRIEF score both significantly predicted the child's lexical entrainment behaviour, such that more repetitions by the experimenter predicted higher entrainment, while a higher BRIEF score and thus more problems with executive functioning predicted lower lexical entrainment.

Our results are consistent with the majority of existing research that did not show between-group differences or significant effects of ToM ability in lexical entrainment on target words. Importantly, the only existing studies that indicated decreased entrainment in individuals with ASD during more unstructured, unpredictable dialogue assessed lexical entrainment in general, rather than on specific target words. In other words, these studies measured the proportion of shared vocabulary between interlocutors, rather than entrainment on specific lexical terms. It is possible that these different approaches to quantifying lexical entrainment may in reality measure two different conversational mechanisms or processes on different levels. Further research can elucidate whether measuring lexical entrainment in these different ways produces results that reflect the same underlying process, or whether lexical entrainment on a "global" versus more "local" scale perhaps rely on different mechanisms.

Based on the absence of significant effects regarding group membership or ToM scores, our findings do not appear to support the common ground/audience design account (Clark and Marshall, 1978; Clark and Murphy, 1982) of entrainment. Interestingly, our results suggest that certain social factors, such as gender, do significantly predict entrainment. This could be taken as support for the Communication Accommodation Theory by Giles et al. (1991).

Our results suggested that boys show more lexical entrainment than girls. A possible explanation for this is the observation that girls with ASD tend to use more compensatory strategies to "fly under the radar", or blend in in social settings. This behaviour is referred to as *camouflaging* or *masking* (Dean et al., 2017). It has been hypothesised that such strategies may also be used in language production (Parish-Morris et al., 2017) in interaction. It is possible that girls with ASD in this study were more likely to use the preferred word for an object because in everyday circumstances, this word would be used more commonly, and using a dispreferred word might make them stand out.

To see whether there was a difference in lexical entrainment between girls and boys between groups, we plotted the difference in preferred and dispreferred lexical item use by group and gender (see Figure 2). This figure shows that girls with ASD indeed show slightly less entrainment than boys and girls without ASD, though this difference is not significant. It is possible that camouflaging in girls with ASD and a general tendency towards social conformity that likely also exists in NT girls explains why girls show significantly less entrainment on dispreferred terms than boys.

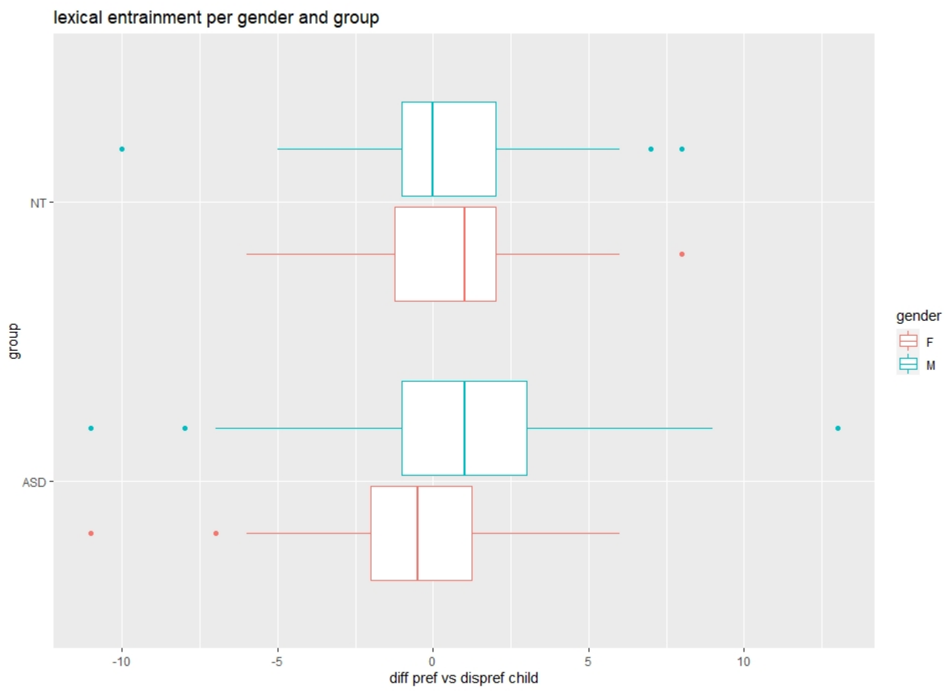
One could argue that our findings support Pickering and Garrod's interactive alignment hypothesis 2004; 2013: there is no difference in entrainment between groups, and no effect of ToM, suggesting that higher-order cognition is not required for entrainment. Additionally, the finding that the number of times the experimenter says a dispreferred word significantly predicts increased lexical entrainment in a child, supports the idea that priming underlies entrainment. However, Pickering and Garrod's theory 2004; 2013 does not explain why gender would affect entrainment, or why executive functioning significantly predicts the degree to which children lexically entrained to the experimenter.

Interestingly, few of the previous studies on entrainment in individuals with ASD included measures of executive functioning. Hopkins et al. (2016) investigated the effect of conflict inhibition on lexical entrainment and found no significant effects. Our results, which suggest that decreased executive functioning relates to lower lexical entrainment, are thus not in line with this previous study. A possible explanation of this is that in the Hopkins et al. (2016) study, a measure of one specific executive function, namely conflict inhibition, was included, and that this was measured with a specific test, whereas we asked participants' parents to fill out the BRIEF questionnaire as an indication of their executive functioning in everyday life. BRIEF scores may reflect a different set of executive functions than the test used by Hopkins et al. (2016).

An alternative explanation is that the experimental paradigm employed by Hopkins et al. (2016) was a game with single-utterance turns and a predictable conversation structure. The task used in the present study was more complicated and required active dialogue to complete. It is plausible that the increased cognitive load of our task and the accompanying dialogue required more of the children's cognitive resources than the game used by Hopkins et al. (2016). This could mean that there were fewer cognitive resources available for processes such as remembering that the experimenter used the dispreferred word and inhibiting the use of the preferred word, thus leading to decreased entrainment, especially in children who have more difficulties with tasks that require executive functioning skills. This is in line with existing research that has suggested that increased task demand and cognitive load leads to reduced entrainment (Abel and Babel, 2017).

The latter explanation is further supported by the study conducted by Stabile and Eigsti (2022), who also investigated lexical entrainment (on a global level) during a Maps task, and also measured executive functioning using the BRIEF questionnaire. While results of the study by Stabile and Eigsti (2022) did not reveal any significant associations between BRIEF score and lexical entrainment, results were marginally significant and in the same

Figure 2: Lexical entrainment, measured as the difference between the number of times a child used a preferred versus a dispreferred term, plotted by group and gender.



direction as the findings here: higher BRIEF score and thus more executive functioning difficulties were associated with lower lexical entrainment.

In other words, the results of the lexical entrainment analysis conducted in this experiment thus do not closely follow the predictions of any of the major theories of entrainment. Rather, the findings point towards a more nuanced and complex picture of lexical entrainment, in which various social and cognitive factors may influence the phenomenon.

6 Conclusion

The aim of this study was to investigate lexical entrainment on target words in children with and without ASD during a semi-naturalistic, task-oriented interaction that had less predictable turn-taking than previous studies. Results of our analysis suggest that some social factors such as age, and some (socio-)cognitive factors such as IQ and ToM score, do not significantly predict lexical entrainment. On the contrary, other social and cognitive factors, such as gender and executive functioning, do significantly predict lexical entrainment: girls show lower degrees of lexical entrainment on dispreferred terms than boys, and more executive functioning challenges in every day life are associated with decreased lexical entrainment. Moreover, the number of times an adult used a dispreferred word significantly predicted a child's entrainment on that dispreferred word. Taken together, the results of this study do not follow the predictions of any of the major theories of entrainment, suggesting that the phenomenon is complex and may be mediated by a number of different mechanisms and factors simultaneously.

Limitations

Dialogue was elicited using an experimental paradigm that does not have structured, predictable turn-taking. As with every decision made during a research process, this had advantages and disadvantages. A disadvantage of this decision was that we could not control the dialogue and thus could not control how often the experimenter used a dispreferred term. We tried to account for this by including it as a fixed effect in the LMEM we constructed, and found that this was indeed a significant predictor of entrainment. Importantly, we did not investigate the order in which interlocutors said dispreferred word: due to the unstructured nature of the conversation, it is possible that sometimes, a child referred to an object before the experimenter had a chance to refer to it by its dispreferred term. Future research may take a more qualitative approach, which could shed more light on the development of lexical entrainment during the dialogue.

Additionally, while our sample size was relatively large, we used a large statistical model and there is a chance our analysis was slightly underpowered. However, recruiting larger groups of children with ASD is extremely time- consuming and requires an incredible amount of resources, so this issue applies to many studies that aim to investigate behaviours of this population. Nonetheless, future studies may aim to implement different statistical tests to mitigate this issue.

Ethics Statement

ARCA's ethics board granted ethical approval for both the current experiment and the broader overarching research project in which this study was embedded. Prior to the experiment, informed consent was obtained from the parent(s) or caregiver(s) of the participants. To compensate for their time and participation, participants and their parent(s)/caregiver(s) received gift vouchers.

It is crucial to exercise caution when making assumptions about Theory of Mind (ToM) or social impairments in disorders such as Autism Spectrum Disorder (ASD). Traditionally, ASD has been associated with ToM impairments and inherent social deficits. However, recent empirical evidence challenges this assumption (e.g. Paynter et al., 2016; Gernsbacher and Yergeau, 2019). Instead of perceiving the communication, ToM, and social difficulties of individuals with ASD as their inherent deficits, it is proposed that these challenges arise due to "neurotype mismatches" occurring during interactions between individuals with ASD and neurotypical (NT) individuals. Individuals with ASD may not lack a theory of mind in general, but rather struggle to understand the mind of NT individuals specifically. Importantly, this perspective works in both directions, as NT individuals also seem to lack an understanding of the "autistic" mind (Sheppard et al., 2016; Heasman and Gillespie, 2018). This conceptualization is known as the "double empathy problem" (Milton, 2012), which is often advocated for by individuals with ASD. Given that most existing research on conversation coordination strategies of individuals with ASD has focused on interactions with a neurotype mismatch, it is crucial to consider this perspective.

Acknowledgements

We would like to thank all participants and their parent(s)/caregiver(s) for participating in this study. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588, from the Slovak Granting Agency grant VEGA 2/0165/21, and the Slovak Research and Development Agency grant numbers APVV-21-0373 and APVV-20-0070.

References

Jennifer Abel and Molly Babel. 2017. Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech*, 60(3):479–502.

M.L. Allen, S. Haywood, G. Rajendran, and H. Branigan. 2011. Evidence for syntactic alignment in children with autism. *Developmental science*, 14(3):540–548.

S. Baron-Cohen. 2000. Theory of mind and autism: A review. *International review of research in mental retardation*, 23:169–184.

S. Baron-Cohen, A.M. Leslie, and U. Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

G. Bossaert, H. Colpin, S.J. Pijl, and K. Petry. 2015. Quality of reciprocated friendships of students with special educational needs in mainstream seventh grade. *Exceptionality*, 23(1):54–72.

H.P. Branigan, A. Tosi, and K. Gillespie-Smith. 2016. Spontaneous lexical alignment in children with an autistic spectrum disorder and their typically developing peers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(11).

H.H. Clark and C. Marshall. 1978. Reference diaries. In *Theoretical issues in natural language*, page 2.

H.H. Clark and G.L. Murphy. 1982. Audience design in meaning and reference. In *Advances in psychology*, volume 9, page 287–299. North-Holland.

K. Cornish, N. Rinehart, K. Gray, and P. Howlin. 2010. *Comic strip task*. Monash University Developmental Neuroscience and Genetic Disorders Laboratory and Monash University Centre for Developmental Psychiatry and Psychology, Melbourne.

Michelle Dean, Robin Harwood, and Connie Kasari. 2017. The art of camouflage: Gender differences in the social behaviors of girls and boys with autism spectrum disorder. *Autism*, 21(6):678–689.

E.A. Demetriou, A. Lampit, D.S. Quintana, S.L. Naismith, Y.J.C. Song, J.E. Pye, and A.J. Guastella. 2018. Autism spectrum disorders: a meta-analysis of executive function. *Molecular psychiatry*, 23(5):1198–1204.

J.W. Du Bois, R.P. Hobson, and J.A. Hobson. 2014. Dialogic resonance and intersubjective engagement in autism. *Cognitive Linguistics*, 25(3):411–441.

Riccardo Fusaroli, Ethan Weed, Roberta Rocca, Deborah Fein, and Letitia Naigles. 2023. repeat after me? both children with and without autism commonly align their language with that of their caregivers.

M.A. Gernsbacher and M. Yergeau. 2019. Empirical failures of the claim that autistic people lack a theory of mind. *Archives of scientific psychology*, 7(1):102.

H. Giles, N. Coupland, and J. Coupland. 1991. Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, and N. Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*, page 1–68. Cambridge University Press, Cambridge.

G.A. Gioia, P.K. Isquith, S.C. Guy, and L. Kenworthy. 2000. *Behavior rating inventory of executive function: BRIEF*. Psychological Assessment Resources, Odessa, FL.

B. Heasman and A. Gillespie. 2018. Perspective-taking is two-sided: Misunderstandings between people with asperger's syndrome and their family members. *Autism*, 22(6):740–750.

Z. Hopkins, N. Yuill, and B. Keller. 2016. Children with autism align syntax in natural conversation. *Applied Psycholinguistics*, 37(2):347–70.

A. Kuznetsova, P.B. Brockhoff, and R.H.B. Christensen. 2017. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

C. Lord, M. Rutter, P.C. DiLavore, S. Risi, K. Gotham, and S.L. Bishop. 2008. *Autism diagnostic observation schedule (ADOS): Manual*. Western Psychological Services, Los Angeles.

D.E. Milton. 2012. On the ontological status of autism: the 'double empathy problem'. *Disability Society*, 27(6):883–887.

Julia Parish-Morris, Mark Y Liberman, Christopher Cieri, John D Herrington, Benjamin E Yerys, Leila Bateman, Joseph Donaher, Emily Ferguson, Juhi Pandey, and Robert T Schultz. 2017. Linguistic camouflage in girls with autism spectrum disorder. *Molecular autism*, 8(1):1–12.

Shivani P Patel, Jennifer Cole, Joseph CY Lau, Gabrielle Fragnito, and Molly Losh. 2022. Verbal entrainment in autism spectrum disorder and first-degree relatives. *Scientific reports*, 12(1):11496.

J.M. Paynter, D. Keen, and V.J. Rose. 2016. Systematic review documents limited empirical support for the practical application of the theory of mind model of asd. *Evidence-Based Communication Assessment and Intervention*, 10(3-4):131–139.

M.J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

M.J. Pickering and S. Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(04):329–347.

E. Sheppard, D. Pillai, G.T.L. Wong, D. Ropar, and P. Mitchell. 2016. How easy is it to read the minds of people with autism spectrum disorder? *Journal of autism and developmental disorders*, 46(4):1247–1254.

K.E. Slocombe, I. Alvarez, H.P. Branigan, T. Jellema, H.G. Burnett, A. Fischer, and L. Levita. 2013. Linguistic alignment in adults with and without asperger's syndrome. *Journal of autism and developmental disorders*, 43(6):1423–1436.

Mackenzie Stabile and Inge-Marie Eigsti. 2022. Lexical alignment and communicative success in autism spectrum disorder. *Journal of Speech, Language, and Hearing Research*, 65(11):4300–4305.

H. Tager-Flusberg. 2007. Evaluating the theory-of-mind hypothesis of autism. *Current directions in psychological science*, 16(6):311–315.

A. Taheri, A. Perry, and P. Minnes. 2016. Examining the social participation of children and adolescents with intellectual disabilities and autism spectrum disorder in relation to peers. *Journal of Intellectual Disability Research*, 60(5):435–443.

Classification of Feedback Functions in Spoken Dialog Using Large Language Models and Prosodic Features

Carol Figueroa
Furhat Robotics
Aix-Marseille University
carol@furhatrobotics.com

Magalie Ochs
Aix-Marseille University
magalie.ochs@lis-lab.fr

Gabriel Skantze
KTH
Royal Institute of Technology
skantze@kth.se

Abstract

Feedback utterances such as 'yeah', 'mhm', and 'okay', convey different communicative functions depending on their prosodic realizations, as well as the conversational context in which they are produced. In this paper, we investigate the performance of different models and features for classifying the communicative function of short feedback tokens in American English dialog. We experiment with a combination of lexical and prosodic features extracted from the feedback utterance, as well as context features from the preceding utterance of the interlocutor. Given the limited amount of training data, we explore the use of a pre-trained large language model (GPT-3) to encode contextual information, as well as SimCSE sentence embeddings. The results show that good performance can be achieved with only SimCSE and lexical features, while the best performance is achieved by solely fine-tuning GPT-3, even if it does not have access to any prosodic features.

1 Introduction

In human-human conversations, short feedback tokens such as 'mhm', 'yeah', and 'wow' serve different communicative functions. For example, 'yeah' can indicate a response to a question, express agreement to an opinion, convey surprise, or simply signal that the interlocutor should continue speaking, depending on the prosodic realization, as well as the conversational context. The terms *feedback* and *backchannels* are sometimes used interchangeably. However, in this paper we use the term backchannel to denote a specific type of feedback that signals that the speaking partner should continue speaking.

There has been a lot of work on incorporating user-generated or system-generated feedback in dialog systems and human-robot interactions (Axelsson et al., 2022). Most work on incorporating feedback in dialog systems have focused on the timing of backchannels (Ward and Tsukahara, 2000; Ruede et al., 2017, 2019; Morency et al., 2010; Adiba et al., 2021; Boudin et al., 2021; Ishii et al., 2021). There has also been work on predicting which type of backchannel or feedback to produce (i.e., predicting what function the backchannel or feedback should convey) (Kawahara et al., 2016; Ortega et al., 2020; Adiba et al., 2021; Boudin et al., 2021; Jang et al., 2021; Lala et al., 2022). In this paper, we focus on the classification of the communicative function of short feedback tokens (i.e., assign the function of feedback), given their lexical and prosodic form as well as the preceding conversational context.

A model that automatically classifies the communicative function of feedback can be used for different purposes. When used offline, such a model could be used to automatically annotate the functions of feedback in a speech corpus. The annotated feedback can then be used to, for example, gain insights into human conversational behavior, or to learn how to synthesize feedback with appropriate prosody, given the feedback function. When used online in a spoken dialog system, it could be used to classify feedback coming from the user.

In this paper, we investigate the performance of different models and features for classifying the communicative function of short feedback tokens in the Switchboard corpus (Godfrey et al., 1992). We use our previously proposed annotation scheme (Figueroa et al., 2022), consisting of 10 feedback functions: continue, non-understanding, agree, disagree, yes/no response, sympathy, mild/strong surprise, and disapproval. For the classification task, we use lexical and prosodic features from the short feedback token, as well as contextual features from the preceding utterance of the interlocutor. Since the representation of dialog context is non-trivial, especially considering the limited amount of annotated data at our disposal, we also investigate the use of probability distributions from a pre-trained large language model (GPT-3) as input to a Support Vector Machine (SVM) classifier, along with the

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

previously mentioned features. To the best of our knowledge, probability distributions from GPT-3 have not been used as an input to another machine learning algorithm for this problem before.

2 Related Work

The classification of feedback functions is related to the more general problem of Dialog Act classification, where the goal is to identify the communicative function of an utterance in dialog. However, in most Dialog Act classification schemes, backchannels are typically treated as a single dialog act category and no fine-grained distinctions are made (Stolcke et al., 2000; Dielmann and Renals, 2008; Liu et al., 2017).

When it comes to the more specific problem of classifying the communicative function of feedback, the only related work we are aware of are Prévot et al. (2015), Neiberg et al. (2013), and Gravano et al. (2007). Although Gravano et al. (2007) do not specifically classify feedback, they do classify affirmative words, which function as a backchannel or acknowledgment/agreement. They use JRIP, a machine learning algorithm to classify affirmative words using text-based, timing, and acoustic-prosodic features from both the affirmative words and context preceding and following the affirmative words. While Neiberg et al. (2013) do not propose a classifier for feedback functions, they use semi-supervised annotations and prosodic clustering to investigate how different prosodic realizations of feedback affect the function of feedback tokens. In Prévot et al. (2015), feedback functions are classified into two levels: *base* function, and *evaluation* function, which respectively correspond to *generic* and *specific* listener responses (Bavelas et al., 2000). A Random forest classifier is first used to classify feedback in the *base* level into the following functions: contact, acknowledgement, evaluation-base, answer, elicit or other. If the feedback is classified into the evaluation-base function, another Random forest classifier is used to classify the feedback into the following functions: approval, expectation, amusement, or confirmation/doubt. Lexical, acoustic, and position information is used of the feedback. Bigrams and the function of the context (the previous utterance) are also used for the classification task.

Feedback Function	Count	GPT-3 prompt label
(C) Continue	1024	Continuer
(U) Non-understanding	63	Misunderstand
(A) Agree	435	Agree
(D) Disagree	46	Disagree
(Y) Yes-response	56	Yes-answer
(N) No-response	114	No-answer
(S) Sympathy	82	Sympathy
(MS) Mild Surprise	103	Interest
(SS) Strong Surprise	191	Surprise
(Ds) Disapproval	65	Reproach
(O) Other	77	Other

Table 1: Feedback functions, count of manually annotated data, and corresponding labels in GPT-3 prompt.

3 Communicative Functions of Feedback

A number of annotation schemes have been proposed for annotating the communicative functions of feedback (Allwood et al., 1992, 2007; Bunt, 2009; Buschmeier et al., 2011; Neiberg et al., 2013; Prévot et al., 2015, 2016; Malisz et al., 2016; Figueroa et al., 2022). As mentioned, feedback can be categorized as having two communicative functions: *generic* and *specific* (Bavelas et al., 2000; Prévot et al., 2015, 2016; Ortega et al., 2020; Boudin et al., 2021). Generic feedback can be thought of as *continuers*; they encourage the interlocutor to continue speaking (Schegloff, 1982). Specific feedback can be thought of as *assessments*; they are listener responses that depend on the context of the interlocutor (Goodwin, 1986). The DIT++ taxomony of dialogue acts also categorizes feedback by two functions, *allo-feedback* and *auto-feedback* which carry information about attention, perception, interpretation, evaluation, and execution of the feedback.

Allwood et al. (1992) introduced four communicative functions of feedback:

- *Contact*: whether the interlocutor is willing and able to continue the interaction
- *Perception*: whether the interlocutor is willing and able to perceive the message
- *Understanding*: whether the interlocutor is willing and able to understand the message
- *Attitudinal reactions*: whether the interlocutor is willing and able to react and (adequately)

respond to the message, specifically whether he/she accepts or rejects it.

These four feedback functions are related to the four levels of joint actions of an addressee proposed by Clark (1994) which are important for establishing common ground. The four feedback functions introduced by Allwood et al. (1992) have inspired many annotations schemes for annotating functions of feedback (Allwood et al., 2007; Buschmeier et al., 2011; Malisz et al., 2016; Neiberg et al., 2013).

In this work, we use our previously proposed annotation scheme (Figueroa et al., 2022), consisting of 10 feedback functions: continue, non-understanding, agree, disagree, yes/no response, sympathy, mild/strong surprise, and disapproval. The scheme also includes an Other category that is used to capture lexical tokens that are not feedback but share the same lexical form as feedback, for example, discourse markers ('okay, let's begin') or literal uses ('he was standing on the right'). The feedback functions continue and understanding can be thought of being in the contact, perception, or understanding grounding level, whereas the other feedback functions are on the attitudinal grounding level.

4 Method

4.1 Corpus and feedback functions

We extracted short feedback tokens from the Switchboard corpus (Godfrey et al., 1992), according to the definition and selection criteria given by Figueroa et al. (2022). Switchboard consists of about 2,500 dyadic telephone calls between 500 native speakers of American English, recorded in two separate channels and lasting about 3-10 minutes. The corpus also contains transcriptions and word level time-alignments.

In total, Switchboard contains 85,956 instances of potential feedback tokens, according to the working definition in Figueroa et al. (2022). Note that this definition is based on the lexical form of the token, and thus may include instances which are not in fact feedback, such as discourse markers. Thus, we train our classifier to also classify such instances as Other. From the full set, we compiled a set of 2256 instances, which were manually annotated with one of the 10 communicative functions (plus Other), as identified in Figueroa et al. (2022), by listening to them in context. Table 1 lists these functions and their counts in our data set.

4.2 Feedback features

For the short feedback utterance, we use the lexical token as well as its prosodic realization as features.

Lexical tokens (e.g. 'yeah', 'wow') and non-lexical tokens (e.g. 'mhm', 'hm') were encoded as one-hot encodings using the scikit-learn Python library (Pedregosa et al., 2011).

Prosodic features – duration, mean pitch, pitch slope, pitch range, and mean intensity – were extracted from the feedback instances. We used Parselmouth (Jadoul et al., 2018) to extract pitch (F0 Hz) and intensity (dB) values. The pitch values were first transformed to log scale and then z-score normalized, intensity values were also z-score normalized. The normalization was done per speaker, where the mean and standard deviation for each speaker were computed from their entire conversation. Pitch slope was calculated by subtracting the mean of the z-score normalized pitch values of the second half of the feedback from the mean of the z-score normalized pitch values of the first half of the feedback.

4.3 Context features

We also added contextual features from 4000 ms of the interlocutor's utterance preceding the feedback. Previous work in feedback modeling have extracted features from the context by either setting an arbitrary window length or number of words. We experimented with a window length of between 1500 - 4000 ms and found that 4000 ms often captured full sentences. We decided to only use features from the preceding utterance of the interlocutor (and not any future context) in order to make the model applicable for online classification.

Part-of-speech (POS) tags of the preceding utterance were extracted using the spaCy Python library. From these, POS bigrams were created and sorted by their term frequency-inverse document frequency (TF-IDF), treating the 10 feedback functions as documents and the POS bigrams as terms. From this list, the top 30 bigrams were selected and used as one-hot features.

Dialog Acts were automatically assigned to the interlocutor's utterance using DialogTag (Malik, 2020), a Python libary. We collapsed the following dialog tags into a single 'Question' tag: 'Yes-No-Question', 'Declarative Yes-No-Question', 'Rhetorical-Question', 'Wh-Question', and 'Tag-Question'. The dialog tags were then one-hot encoded.

Sentence Embedding of the previous utterance was obtained using SimCSE (Gao et al., 2021), which is an auto-encoding embedding technique based on contrastive learning. During training, SimCSE uses BERT encodings of the input and then fine-tunes the parameters using the contrastive learning objective which pushes together semantically similar pairs and pushes apart semantically dissimilar pairs. We used the sup-simcse-bert-base-uncased pre-trained model of SimCSE which is readily available on Github (Gao et al., 2021).

4.4 GPT-3

As an alternative to the context features listed above, we also explored the use of GPT-3 from OpenAI (Brown et al., 2020) to encode the previous utterance, as well as the lexical form of the feedback token. We tested three different approaches: zero-short, few-shot, and fine-tuning.

For *zero-shot* classification, we provided GPT-3 with a prompt similar to the one shown in Table 2. The prompt ends with the opening bracket at the end, and GPT-3 is asked to predict the next token (marked in bold). This is done using the davinci-003 model. For *few-shot* classification, we provided an example of each function, with both the dialog and the corresponding label, in addition to the instructions.

The third approach is to *fine-tune* GPT-3. We fine-tuned the davinci base model, since davinci-003 is not available for fine-tuning. For fine-tuning, there are no instructions or examples in the prompt; the model is only given training examples, which consist of input text (the preceding utterance and the feedback token) and its associated output (the function label).

Note that, in Table 2, the feedback function labels in the GPT-3 prompt have been changed from the ones listed in the first column of Table 1. Since GPT-3 generates word pieces, we changed the feedback function labels in the prompt so that they would not start with the same first letters. This way, we can simply inspect the first generated word piece from GPT-3 and map it to one of the functions.

For zero-shot and few-shot classification, we also explored if the prediction could be used as an input feature to the feedback function classifier, rather than using it directly. For this, we use the probability distribution that GPT-3 outputs over potential function labels (or rather their prefix).

GPT-3 Prompt
The following is a list of dialog acts and their description in parentheses: - Continuer (Backchannel) - Misunderstand (Expressing non-understanding) - Agree (Agreeing with a statement) - Disagree (Disagreeing with a statement) - Yes-answer (A positive answer to a yes/no question) - No-answer (A negative answer to a yes/no question) - Sympathy (Expressing empathy) - Reproach (Expressing disapproval or disgust or disappointment) - Interest (Expressing interest) - Surprise (Expressing surprise) - Other (thinking or interrupting conversation) The following is a dialog between two persons. The dialog acts are written in brackets. A: i was mowing the lawn yesterday B: mhm [**continuer]**

Table 2: Prompt given to GPT-3.

From GPT-3, we can get the top five labels that would have been generated by the language model and their corresponding probabilities. For example, given the feedback 'yeah', GPT-3 could predict the following word pieces: 'Ag' 74%, 'Contin' 1.7%, 'Yes' 21%, 'agree' 3%, and 'yes' 0.3%. From these probabilities, we can create a vector where feedback function (A)gree is assigned 77%, (C)ontinue 1.7%, (Y)es-response 21.3%, and all other functions plus the Other category are assigned 0%. These probability distributions can then be used as input features to the main function classifier.

For all GPT-3 models we use the following settings: temperature=0, max_tokens=1, frequency_penalty=0, presence_penalty=0.6, and logprobs=5.

4.5 Function classifier

The task of the main classifier is to classify the feedback function, given the features listed above. As explained, GPT-3 can be used both as a main classifier and as a method for encoding lexical and contextual information, which can then be used as input to another classifier. Since we did not have a large data set to train a deep learning model, we explored three machine learning models which can handle small data sets (Forman and Cohen, 2004): Support Vector Machine (SVM), Logistic Regression, and Random Forest, using the classifiers implemented in scikit-learn a Python library (Pedregosa et al., 2011).

For the three classifiers, we set the parameter class_weight as balanced. For the SVM classifier we used a linear kernel, we also experimented with radial basis function kernel but the linear kernel gave the best results. For the Logistic Regression classifier we set the max_iter to 200. For all other parameters we used the default settings. In general, we found that the SVM classifier performed the best, and thus only report our results from the SVM classifiers.

We experiment with different combinations of input features and evaluate our SVM classifiers using 10-fold cross validation. In order to evaluate the model performance, we use the F1-weighted score. In the cases where we do not fine-tune GPT-3 and use it directly as a classifier, we do not use cross validation, but instead use our entire annotated data.

5 Results

5.1 Classifier performance

Table 3 summarizes the F1-weighted scores of the different models with different combinations of input features. For comparison, we also report the majority-class baseline, as well as the inter-annotator agreement annotations from our previous work (Figueroa et al., 2022). Note that only 1124 feedback utterances were annotated for the inter-annotator agreements. In cases where the annotators could not decide on a single function (e.g. 'A/C'), we chose one of the functions randomly while calculating the F1-weighted score. This procedure was averaged over 10 times.

When only lexical features are used (Model 1), we get a fairly high F1-weighted score (0.63) which outperforms the baseline. We used the majority-class baseline which returns the frequent class label. The prosodic features are not very informative, and adding them to the lexical features do not improve the performance further (Model 16).

Among the contextual features, SimCSE is clearly the most informative (Model 7). Just using GPT-3 as a zero-shot or few-shot classifier or as input features does not appear to be very useful (Models 10,11,13,14), considering that it also encodes lexical information about the feedback utterance; the performance is on par with Model 1 which only uses lexical features. There is also no significant difference between using zero-shot or few-shot (Model 13,14) ($t(18) = 1.585$; $p = 0.13$). Lexical features in combination with SimCSE, on the other hand, do give a better performance: Model 19 performs as well as the inter-annotator agreement score. Figure 1 shows the confusion matrix of Model 19 trained on 1804 examples and evaluated on a 452 test set. We can see that it performs poorly on (D)isagree, (Ds) Disapproval, and (Y)es-response. This poor performance could be due to the few training examples.

To improve this score further, we need to add prosodic features, the dialog act, and the GPT-3 distributions to the lexical and SimCSE features. While Model 20 performs significantly better than Model 19 ($t(18) = 2.509$; $p = 0.02$), the difference is not very big, considering the much larger feature set. Figure 2 shows the confusion matrix of Model 20 trained on the same 1804 examples and

Model #: Features	**F-score**
1: Lexical	0.63
Prosody	
2: Duration	0.10
3: Mean pitch	0.16
4: Pitch slope	0.24
5: Pitch range	0.18
6: Mean intensity	0.15
Context	
7: SimCSE	0.32
8: Dialog act (DA)	0.14
9: Part-of-speech (POS)	0.09
GPT-3	
10: Zero-shot majority*	0.61
11: Few-shot majority*	0.65
12: Fine-tuned*	**0.80**
13: Zero-shot as features (ZS)	0.61
14: Few-shot as features (FS)	0.63
Combinations	
15: Prosody (all)	0.37
16: Lexical + Prosody (LexPro)	0.63
17: Lexical + GPT-3 (ZS)	0.68
18: Lexical + GPT-3 (FS)	0.69
19: Lexical + SimCSE	**0.72**
20: LexPro + SimCSE + DA + GPT-3 (FS)	**0.76**
Majority-class baseline	0.28
Inter-annotator agreement	**0.74**

Table 3: F1 weighted scores for different feature sets. *Uses GPT-3 (and not SVM) as the main classifier.

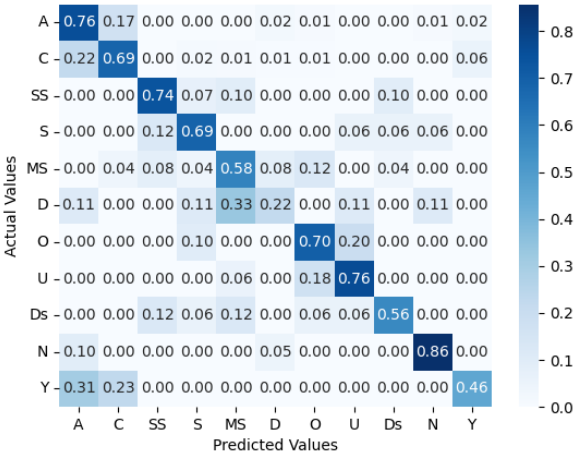


Figure 1: Confusion matrix for Model 19.

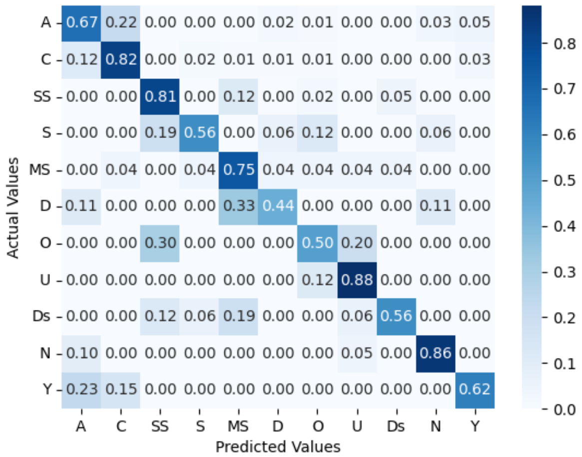


Figure 2: Confusion matrix for Model 20.

evaluated on the same 452 test set. By adding the prosodic features and the dialog act, we can see that it improves the classification for (C)ontinue, (SS) Strong Surprise, (MS) Mild Surprise, (D)isagree, (U) Non-Understanding, and (Y)es-response.

The best performing model is the GPT-3 fine-tuned classifier (Model 12), which performs significantly better than Model 20 (t(18) = -2.803; p = 0.01). Figure 3 shows the confusion matrix of Model 12 trained on the 1804 examples and evaluated on the 452 test set. The GPT-3 fine-tuned classifier improves the classification for (C) Continue, (SS) Strong Surprise, (O) Other, (U) Non-understanding, (N)o-response, and (Y)es-response.

To conclude, models 12, 19, and 20 are all viable classifiers for feedback functions, and they all seem to perform on par with the inter-annotator agreement. The choice of classifier depends on specific requirements, for example whether it should be used offline or online, and whether access to GPT-3 is available.

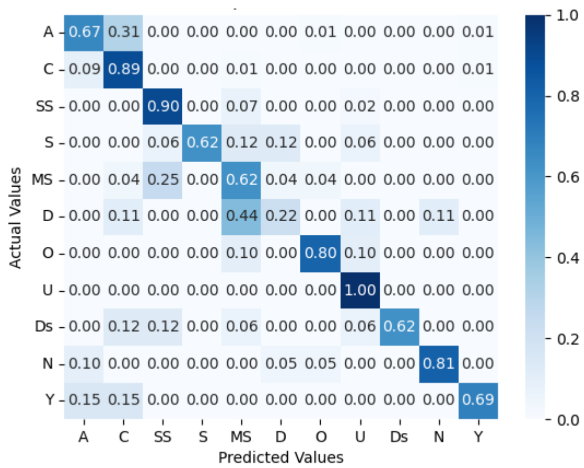


Figure 3: Confusion matrix for Model 12.

5.2 Labeling the remaining Switchboard corpus

Given that we now have working classifiers of feedback functions for Switchboard, we finally experimented with applying one of them to the remaining set of 83,700 potential feedback instances in the Switchboard corpus, in order to study the general distribution of the communicative functions. For this, we used Model 19, as it has a low cost while the performance is relatively good. The distributions are shown in Table 4 which include the distributions of the 2,256 manually annotated lexical tokens and the 83,700 automatically annotated lexical tokens. In total, there were 74,106 instances of actual feedback (not Other), according to the classifier. As can be seen, (C)ontinue and (A)gree are the most frequent feedback functions.

5.3 Investigating sex differences

To illustrate how this classification can be used for further analysis, we also broke down these numbers based on the sex of the listener (i.e., the interlocutor producing the feedback), as provided in the Switchboard corpus. This is shown in Table 4. Note that in the metadata of Switchboard there are only two options for sex, female and male. A chi-square test revealed that sex influences the type of feedback ($\chi^2(9) = 1165.71, p < .001$). Analysis of the standardized residuals ($\alpha = 0.05$) revealed that there were significant differences in most feedback types, as indicated in Table 4. Perhaps most notably, the use of (S)ympathy, and (SS) Strong surprise is much more frequent for females than males. To further investigate whether these effects are also affected by the sex of the interlocutor re-

Function	Total	Tot. %	F %	M %	FF %	FM %	MM %	MF %	Ov %
(C) Continue	39499	51.8	**54.2**	**48.8**	54.7	53.6	48.4	49.3	45.3
(U) Non-understanding	342	0.45	**0.34**	**0.58**	0.31	0.39	0.56	0.60	26.9
(A) Agree	22809	29.9	**26.3**	**34.3**	26.3	26.3	**35.3**	**33.2**	45.3
(D) Disagree	986	1.29	**1.12**	**1.51**	**0.96**	**1.32**	1.56	1.44	36.0
(Y) Yes-response	4101	5.38	**4.93**	**5.93**	5.05	4.77	5.90	5.97	45.1
(N) No-response	787	1.03	1.02	1.05	0.96	1.08	**0.90**	**1.24**	34.6
(S) Sympathy	1775	2.33	**3.19**	**1.25**	3.29	3.06	**1.01**	**1.55**	46.7
(MS) Mild surprise	2325	3.05	3.03	3.07	2.82	**3.31**	3.01	3.13	37.7
(SS) Strong surprise	3023	3.96	**5.03**	**2.64**	4.81	5.31	2.62	2.66	41.8
(Ds) Disapproval	638	0.84	0.81	0.86	0.78	0.85	0.80	0.94	39.8
(O) Other	9671								

Table 4: Distribution of manually annotated and automatically annotated tokens in the Switchboard corpus. Distribution percentages are calculated excluding the Other category. Automatic annotations used Model 19. F=Female, M=Male. FM = Female-to-Male feedback, etc. Bold numbers denote significant deviations from the expected distribution ($\alpha = 0.05$). Ov=Overlap.

ceiving the feedback, we also split these numbers based on the sex of both interlocutors, as can be seen in Table 4. Chi-square tests revealed that there was indeed such an effect, both when the feedback was produced by males ($\chi^2(9) = 46.2, p < .001$) and females ($\chi^2(9) = 34.7, p < .05$). For example, in male-male conversations, there is less use of (S)ympathy, compared to in male-female conversations.

Our analysis also shows that in general, females produce 2.73 feedback tokens per minute, whereas males produce 2.23 feedback tokens per minute. Our findings only reflect observations in the Switchboard corpus and therefore these findings may not be generalizable to other corpora.

5.4 Analysis of overlap

Another example in which this classification can be used is in analyzing whether certain feedback functions overlap more or less with the speech of the interlocutor. In order to determine whether a feedback was overlapping or not, we took the start time of the feedback and searched for that timestamp in the speech of the interlocutor, if that timestamp occurred during or the start of the interlocutor's speech we assigned the feedback as overlapping. If the start time of the feedback occurred during the interlocutor's silence or laughter we assigned the feedback as not overlapping. The percentage of overlap for each feedback type is shown in Table 4.

Using this method, we find that (U) Non-understanding, (D)isagree, (N)o-response, (MS) Mild Surprise, and (Ds) Disapproval tend to not overlap as much with the interlocutor's speech. Listeners may wait to produce a feedback function (U) Non-understanding until the end of the interlocutor's turn in order to first see if they can repair their comprehension of what was said or being said. Listener's may also wait to produce feedback functions with negative connotations such as (D)isagree, (N)o-response, and (Ds) Disapproval, in order to decide whether they should take the turn, or to further respond to what the interlocutor has said.

We had expected feedback functions (C)ontinue, (A)gree, (S)ympathy, (MS) Mild Surprise, and (SS) Strong Surprise to overlap with the speech of the interlocutor. However, we find that for (C)ontinue, (A)gree, (S)ympathy, (SS) Stong Surprise, these feedback functions almost equally overlap and not overlap. Further analysis should be done to see if the silences of the interlocutor's are short breaths or longer pauses. It would be interesting to do an analysis similar to the one done by Goodwin (1986) where they compared assessments and continuers. They found that although assessments and continuers share similar contexts (they are said during the speech of the interlocutor), continuers bridge turn-constructional units of the interlocutor, whereas assessments do not interrupt the subsequent unit of the interlocutor. This type of analysis which takes into consideration conversational units of the interlocutor may give more insight into where exactly these feedback functions occur within the interlocutor's turn, as well as give information whether the feedback functions which occur during the interlocutor's silences are between or within turn-constructional units.

6 Discussion

Although fine-tuning GPT-3 (Model 12) performs the best, it may not be suitable for an online setting or for annotating large corpora. This model is dependent on OpenAI's API which can have downtime and using it can be costly. Model 19 (SimCSE + Lexical), which can fairly accurately predict feedback functions on par with human annotators, can be an option for online settings or for annotating large corpora.

The prosodic features performed poorly, and in the end they did not contribute much to the best-performing models. The best model, using a fine-tuned GPT-3, did not use any prosodic features at all. This is perhaps a bit surprising, since prosody should help to disambiguate feedback tokens which are not easily classifiable given only textual information, such as 'no' when it is used as negative agreement. On the other hand, it might be the case that the preceding context contains redundant information, and could for example help to disambiguate a question (preceding a No-response) vs. a statement (preceding a Disapproval). In any case, future work should explore better prosodic features, using distributed, self-supervised speech representations (Lin et al., 2023). It is also interesting to note that discrete representations of dialog context, such as Dialog Acts and Part-of-speech, performed much more poorly than the distributed representations (SimCSE).

One reason that the Dialog Acts may not have performed well (Model 8) could be due to the preceding context being misclassified with the incorrect dialog act. Therefore, this is an error that was propagated into the model. This propagation of errors can also be said for the probability distributions by GPT-3 (Models 13, 14). For future work, further analysis should be done on how these errors affect the model.

There is also more experimentation that could be done with GPT-3. Without fine-tuning, the probability distributions from zero-shot and few-shot classifications did not perform better than the lexical one-hot encodings. Experimentation with different prompts could improve the GPT-3 features. In future work, we would like to use a separate training set to fine-tune the GPT-3 model so that we can evaluate the probability distributions of the fine-tuned model, and potentially combine them with other features. One potential route could also be to add prosodic information to the prompts by discretizing them. For example, pitch slope could be disctretized by describing it as flat, rising, or falling.

Our classification models have only been trained and evaluated with the Switchboard corpus, it would be interesting to see how our best models perform with other corpora, such as corpora where the interlocutors are speaking face-to-face.

7 Conclusion

In this paper, we proposed different models which can automatically classify 10 communicative feedback functions: continue, non-understanding, agree, disagree, yes/no response, sympathy, mild/strong surprise, and disapproval. We experimented with different combinations of lexical and prosodic features from the feedback utterances, as well as context features from the preceding utterance of the interlocutor as input to a SVM classifier. For contextual features, we investigated the use of probability distributions from the predicted function labels from a zero-shot or few-shot GPT-3 classifier, as well as SimCSE sentence embeddings. Finally, we also compared with a fine-tuned GPT-3 classifier.

Our experiments show that just using lexical features and SimCSE gives a fairly good performance, on par with inter-annotator agreement. While using GPT-3 in a zero-shot or few-shot fashion does not contribute much, a fine-tuned GPT-3 model outperforms all other models, even though no prosodic information is used.

The automatic annotations of the communicative functions of feedback in the Switchboard corpus by Model 19 can be found in this repository: https://github.com/carolfigPhD/FeedbackAnnotationScheme.

Limitations

We are aware that one limitation in terms of reproducibility is that GPT-3 may not return the same labels if the experiments were to be run again. Moreover, GPT-3 is like a black-box, when we fine-tune the model we do not know what exactly is being fine-tuned. Another limitation is accessibility, not everyone will have access to GPT-3 which can be costly and is dependent on the services of OpenAI.

We have also not examined if there are differences in feedback in face-to-face conversations compared to telephone conversations. As mentioned, we have only trained and evaluated our classifiers with the Switchboard corpus but have not evaluated with a face-to-face corpus.

Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 859588.

References

Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7408–7412.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling feedback in interaction with conversational agents—a review. *Frontiers in Computer Science*, 4.

Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A multimodal model for predicting conversational feedbacks. In *International conference on text, speech, and dialogue*, pages 537–549. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

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 *Proc. Interspeech 2011*, pages 2057–2060.

Herbert H Clark. 1994. Managing problems in speaking. *Speech communication*, 15(3-4):243–250.

Alfred Dielmann and Steve Renals. 2008. Recognition of dialogue acts in multiparty meetings using a switching dbn. *IEEE transactions on audio, speech, and language processing*, 16(7):1303–1314.

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, and Gabriel Skantze. 2022. Annotation of communicative functions of short feedback tokens in switchboard. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1849–1859.

George Forman and Ira Cohen. 2004. Learning from little: Comparison of classifiers given little training. In *Knowledge Discovery in Databases: PKDD 2004: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, 2004. Proceedings 8*, pages 161–172. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, 9(2):205–217.

Agustín Gravano, Stefan Benus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Proc. Interspeech 2007*, pages 1613–1616.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 131–138.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. Bpm_mt: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3452.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *Proc. Interspeech 2016*, pages 2890–2894.

Divesh Lala, Koji Inoue, Tatsuya Kawahara, and Kei Sawada. 2022. Backchannel generation model for a third party listener agent. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, pages 114–122.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G Ward. 2023. On the utility of self-supervised models for prosody-related tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111. IEEE.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Bhavitvya Malik. 2020. Dialogtag.

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(2):411–442.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems*, 20:70–84.

Daniel Neiberg, Giampiero Salvi, and Joakim Gustafson. 2013. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55(3):451–469.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Laurent Prévot, Jan Gorisch, and Roxane Bertrand. 2016. A cup of cofee: A large collection of feedback utterances provided with communicative function annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3180–3185.

Laurent Prévot, Jan Gorisch, and Sankar Mukherjee. 2015. Annotation and classification of french feedback communicative functions. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 298–306.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*, pages 247–258. Springer.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings. In *Proc. Interspeech 2017*, pages 879–883.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:93.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Grounding with particles

Ahmad Jabbar*
Stanford University
jabbar@stanford.edu

Veda Kanamarlapudi*
Stanford University
vedakan@stanford.edu

Abstract

We focus on a *sui generis* grounding move in Hindi-Urdu dialogue, namely *voh hi na*. A dataset consisting of minimal pairs of dialogues is presented to get a better sense of the move. Using dynamic models of discourse structure, we propose a semantics for *voh hi na* in terms of its update effects.

1 Introduction

Grounding moves are an important part of any dialogue (Clark and Schaefer, 1989; Ginzburg, 1996). *Inter alia*, they are important for purposes of coherence and cooperativity in dialogue. Recent work has also shed light on their importance for understanding clause-types *per se* (Farkas and Bruce, 2010). Moreover, the recent past has seen a boom in the literature on discourse particles (Rojas-Esponda, 2014; Theiler, 2021; Yuan, 2020). Within this boom, there has also been a focus on exploring the rich ground for Hindi-Urdu discourse particles (Brown, 2022; Deo, 2022, 2023b; Jabbar, f.c.).

We bring the two lines of work, grounding moves and discourse particles, together to study a *sui generis* grounding move in Hindi-Urdu, *voh hi na*. What is noteworthy about this string is that it has two discourse particles, *hi* and *na* appended to a propositional anaphor *voh*.[1] *Voh* is a third-person pronoun that can also function as a propositional anaphor in dialogues.[2]

To give a sense for its use, which we make more precise below, *voh hi na* is licensed only in contexts where the interlocutor comes to see the speaker's point of view.

(1) A: The pizza's stale.
B (stubbornly takes a bite): Yeah, it tastes pretty bad.
A: Voh hi na.

In its above sense, *voh hi na* is similar to *told you so*. That's a good place to start, which we modify step-by-step in light of the data we present here. We proceed as follows: in §2, we make a few ground-clearing remarks. In §3, we construct minimal pairs of contexts and dialogues to see when it is felicitous to use *voh hi na*. We propose a semantic account in §4. Using the account, in §5, we explain the data we present. In §6, we conclude.

Via this paper, we seek to contribute in the following ways to the literature. Although there's a lot of insightful work on discourse and question particles in Hindi-Urdu, to our knowledge, there's not much work on the semantics and pragmatics of Hindi-Urdu dialogue.[3] We motivate inquiry into Hindi-Urdu dialogue by presenting a unique grounding move. We also observe that the string *voh hi na* is interesting in that two discourse particles *hi* and *na* contribute a compositional meaning. This is noteworthy in light of the dearth of work on discourse particles composition.[4] For this paper, we only focus on the compositional meaning, without breaking down the individual contributions of the two particles.[5]

2 Preliminaries

Before we move on to our explicandum, we remark briefly on grounding moves. The simplest way to understand the nature of grounding, for the purposes of this paper, is to note that assertions are

*Equal contribution.

[1] *toh* is another particle in Hindi-Urdu, and interestingly, one can infix *toh* in *voh hi na* as in *voh hi toh na* to form another felicitous string. For now, we focus on *voh hi na* and add remarks about infixing *toh* in our conclusion below.

[2] See Bhatia and Bhatt (2023) for some data on this pronoun.

[3] For the polar question particle *kya*, see Bhatt and Dayal (2020); Biezma et al. (2022).

[4] Zimmermann (2011)'s overview on discourse particles includes discussion of some thorny issues surrounding scope and composition.

[5] The reader is directed to Bhatt (1994); Deo (2023a); Jabbar (f.c.) for more on these two particles.

tentative; assertions are proposals to make some content common ground. In this way, assertions are subject to acceptance for their content to be made common ground. With this notion of acceptance in mind, one can understand a subclass of grounding moves as moves in the discourse that accept and thereby acknowledge assertions.[6] In this paper, we focus on a member of this subclass of grounding moves in Hindi-Urdu. Given this characterization, we can say that A's final move in (1) grounds B's assertion. Sure the content of the grounded assertion becomes common ground, but we take it upon ourselves to show that *voh hi na* does more than just accepting the content of the grounded assertion.

We take *voh hi na* to be a grounding move because a speaker cannot use it, on its own, without a prior utterance by an interlocutor. Simply, it cannot be used as the first move in a discourse by a speaker. Something ought to have come before it which it grounds. Its uniqueness arises out of the conditions of its felicitous use. Given that *voh hi na* can only be used to ground, and never out of the blue, and given that it is licensed under very specific conditions, we take that the unique grounding update is conventionally encoded in the string as its compositional meaning. Let's take a closer look at this string.

Voh is a third person pronoun, and translates to *that* in English. Just like English *that*, *voh* can also be used to form referring expressions that can be used deictically to refer to salient individuals in discourse.[7] However, in its referential uses in Hindi-Urdu, *voh* cannot be appended with a sequence of particles as in *voh hi na*. In other words, a speaker cannot point to a person X and say *voh hi na* to refer to X. *Voh hi na* can be used felicitously only when *voh* is anaphoric on an antecedent proposition, as can be noted in the dialogues we present. That is why we call the *voh* in felicitous *voh hi na* strings, *propositional anaphor*.

The above way of characterizing *voh hi na* is helpful. We can break down the string into the the antecedent proposition for *voh*, call it p for now, and isolate the contribution of *hi na* as is standard in the literature on discourse particles and clause-types. To wit, *hi na* somehow relates p to the structure of the discourse or some (epistemic) states or preferences of the participants (cf. Kaufmann (2011); Condoravdi and Lauer (2012); Rett (2011)). Therefore, we break down our inquiry into *voh hi na* as consisting of the antecedent proposition for *voh*, and the specific way the proposition is coherent in the discourse. These coherence conditions specify the felicitous distribution of *voh hi na*. In future work, we intend to explore how *hi* and *na* interact compositionally to yield the felicity conditions we specify for *voh hi na*. In this paper, we simply offer the felicity conditions.

Given that we have already broken down *voh hi na* in noting that it consists of a propositional anaphor and two discourse particles, we don't present glosses for our dialogues below. Instead, for brevity's sake, each of the dialogues consists of English sentences. In all dialogues, the final move is *voh hi na*, which we hold constant, varying only in its felicity, across dialogues.

3 Dialogues and analysis

Our strategy in this section is to situate each dialogue in a context. Each context specifies the information states of the discourse participants and other related facts. While dialogues are numbered as usual, contexts are given names for ease of recall later. First, consider (2) in STROLL 1 and (3) in STROLL 2.

[STROLL 1]: A and B are in Manchester, and it has been rainy for the past few days. It's a new day now. B puts on her jacket to prepare for taking a stroll, when A expresses his suspicion that it may be raining.

(2) A: You're being optimistic. It's probably raining outside.
B (checks the weather app): Yeah, there's a 100% chance of rain.
A: Voh hi na.

[STROLL 2]: A and B are in Manchester, and it has been rainy for the past few days. It's a new day now. A and B are excited to take a pre-planned stroll outside. As a last minute consideration before leaving, B checks the weather app.

(3) B (checks the weather app): Oh, there's a 100% chance of rain.
A: # Voh hi na.

The minimal pair of dialogues above helps bring out the following point: to be able to felicitously use *voh hi na* to ground an utterance u, the speaker had to have made a prior commitment q in the dis-

[6] See Clark and Schaefer (1989) for a hierarchy of grounding moves.

[7] Complex demonstratives as in *that man with the mustache* serve just this purpose. See King (2001) for an overview.

course. Moreover, q must occur prior to u and the content of u must verify or validate q somehow. Here, we use the notions *verification* and *validation* pre-theoretically; we make them precise below. Then, with a notion of validation, yet to be made precise, we can state an observation below.

(4) **Observation 1**: The speaker who grounds u with *voh hi na* ought to have made a prior commitment that the interlocutor validates with u.

In the same vein, consider another context.[8]
[VEGAN]: A and B are discussing whether there are good restaurants on campus. A is vegan, while B is not. They have the following exchange:

(5) A: I haven't had any good food from a restaurant here on campus yet.
B: The restaurant in the south end has really nice burgers ... oh, wait, but you're vegan. You can't go there.
A: Voh hi na.

A's use of *voh hi na* above is felicitous. However, the use would have been infelicitous had A not been vegan, and had B's recommendation for the south end restaurant been helpful for A. Note that this recommendation by B is presented primarily to guide A's future actions. This recommendation would have been effective, according to B, had A not been vegan.
[NOT VEGAN]: A and B are discussing whether there are good restaurants on campus. Neither A nor B is vegan. They have the following exchange:

(6) A: I haven't had any good food from a restaurant here on campus yet.
B: The restaurant in the south end has really nice burgers.
A: # Voh hi na.

Now, knowing about the south end restaurant can influence A's actions in the following way. After coming to know that there's a nice burger place in the south end, A might not consider eating on campus to be as sub-optimal as A was considering it prior to knowing about the burger place. It's also quite possible that A may still consider eating on campus to be as sub-optimal after coming to know about the restaurant. However, if B knows that the recommendation for the south-end restaurant is futile, B wouldn't offer it—or so A thinks. According to A, what drives B to offer the recommendation is the following open possibility: that in light of B's contribution, A might come to consider eating on campus to be not as sub-optimal. From A's perspective, that's exactly what motivates B to make that specific contribution in the first place. In VEGAN however, B comes to recall that A is vegan. This knowledge makes B no longer believe that *the south end restaurant has nice burgers* would make A change A's preference about eating on campus. As the proposition that A is vegan is specifically stated by B, A is privy to B's mental state that B's prior contribution *the south-end restaurant has really nice burgers* preserve A's preferences over the set of actions, as A is vegan. The minimal difference in the two contexts again brings out the difference in the felicity of *voh hi na*. What observation can we distill here?

From STROLL 1 and 2, we were able to understand that to felicitously ground with *voh hi na*, the speaker ought to have made a prior commitment that the utterance preceding the grounding somehow verifies.[9] Let the content of the interlocutor's utterance be p. In light of (5) and (6), we observe that the speaker should have the following belief about p: that the interlocutor thinks that updating the speaker information state with p does not change the speaker ranking for a salient set of alternative actions. Using this condition, we can understand the notion of verification, introduced above, precisely. What is verified or validated is then the speaker ranking for a set of alternative actions. We note this as observation 2 below and refer to it as *rank preservation* alternatively.

(7) **Observation 2/Rank-preservation**: The speaker thinks that according to the interlocutor, the speaker ranking for a salient set of alternative actions does not change once the antecedent proposition for *voh* as in *voh hi na* is made common ground.

Note that (7) involves reasoning about the interlocutor's mental state. This mental state is about a set of alternative actions $\mathcal{A}$, a proposition p, and the speaker preferences over $\mathcal{A}$ in light of p. Such

[8] For the following pair, for ease, we assume that it's common ground between A and B that none of the places on campus serve vegan burgers.

[9] All throughout the paper, by *speaker*, we mean the speaker of *voh hi na*, and by *interlocutor*, the participant whose move gets grounded by *voh hi na*.

reasoning about other's mental states is crucial in dialogue; not only for the purposes of fully understanding each other (van Rooij, 2003; Gunlogson, 2008; Goodman and Frank, 2016), but also for grounding each other's assertions in dialogue (Benz, 2006; Stone and Lascarides, 2010). For building intuition for such reasoning, note that a speaker may assert that there's a strike only if the speaker thinks that the interlocutor doesn't know that there's a strike. Reasoning about each other's information states guides the sort of contributions speakers make. We can build intuition even for (7).

Let's say there's a speaker preference between a and $\neg a$. There's an interlocutor belief about that speaker preference. In light of the contribution that the interlocutor makes, the speaker can reason about the interlocutor's mental state. The interlocutor may come to believe that the speaker preference does not change in light of their contribution. When the speaker thinks that the interlocutor comes to believe that the speaker preference does not change, *voh hi na* is felicitous to ground the interlocutor contribution. That's the idea we work with for now, until we make it precise in §4.

Further, we can separate two things in (7). Although it is the ranking as done by the speaker of *voh hi na*, i.e. A in our examples, that remains unchanged according to B, the agent whose actions get ranked by A's preferences need not be the speaker. This is especially vivid in the STROLL minimal pair, where the relevant agent is not (just) the speaker. It is both A and B who deliberate over taking a stroll in STROLL. In VEGAN, it is just A who deliberates over a set of actions. Let's capture this in the observation below.

(8) **Observation 3**: The agent x to whom the action set is relativized, as in $\mathcal{A}_x$, is contextually determined.

The contextual determination of the agent is not surprising, as the relevant set of actions is contextually determined too.

Although we present our account more fully in §4, another clarificatory remark is in order. Following Kolodny and MacFarlane (2010)'s remark "For in addition to talking of what agents ought to do, we talk of what thinkers ought to believe" (Kolodny and MacFarlane, 2010, page 132), we take believing to be an action too.[10] In other words, we construe *action* broadly so as to include doxastic actions. How does this help? Both of the above contexts were set such that there was a salient action available to at least one discourse agent. Below, we construct a context where an agent is divided on what to believe.

[RAIN]: A and B are talking about how it rains so much in San Francisco. B is under the impression that it's not raining today. But, it has rained all days of the week, including today. The following exchange occurs.

(9) A: It has rained all week.
B: Oh, but it's not raining today. (Takes a peek out of the window.) Oh wait, it is raining.
A: Voh hi na.

Now, RAIN is set such that there's no salient action apart from believing or not that it is raining, and A and B are deciding between that. Therefore, what unites all of the contexts so far is rank-preservation over a set of actions, where the conception of *action* includes doxastic actions too.

The contexts so far might give the following impression: that the interlocutor has to update their belief state to align with the speaker's; and that this alignment makes it felicitous for the speaker to ground the interlocutor's discourse move with *voh hi na*. This generalization doesn't hold, and baking this into the semantics for *voh hi na* would underspecify its felicitous uses. Consider the context below.

[HIKE]: It's raining very heavily. A had planned to go on the hike, but now A is put off by the rain. The following exchange occurs between A and B.

(10) A: I don't think I'll go to the hike.
B: The trail must be very slippery too today.
A (glumly): Voh hi na.

B doesn't come to update their belief state or preference for an action. All that occurs is that B says something that validates A's preference for not going to the hike.

Moreover, in (10), A has no credence or degree of positive belief in the proposition that the trail is very slippery. A has a preference for an action, and the proposition that the trail must be very slippery preserves A's preference for not going to the hike. The important thing to note is that we cannot make a generalization about speaker's prior belief about

[10] Also see McCready (2008)'s interplay between actions and beliefs within an information state.

p, where p is the proposition which the speaker grounds with *voh hi na*. Prior to interlocutor's assertion of p, the speaker may have a belief w.r.t. p or may be agnostic. *Voh hi na*'s felicity doesn't co-vary with a prior belief w.r.t p.

The above contexts may give the impression that the speaker who grounds a discourse move with *voh hi na* ought to have ranked the set of available actions such that a unique action among the set comes out to be preferred. However, it is quite possible that the current information state of the speaker doesn't break the tie between two alternative actions in the action set. Consider the following context and dialogue.

[PROPOSAL]: A and B are friends, and A is dating Mohan since a year now. A thinks that Mohan will propose to her, but A is divided between whether she should say *yes* or not. Mohan is loving towards A, which A loves, but Mohan is rude towards work staff, which A hates. A expresses this problem to B, and continues by saying,

(11) A: I'm really not sure about Mohan.
B: I get you! He is such a loving guy, but he comes off as super arrogant occasionally. Now, how does one decide?
A: Voh hi na.

In the above context, A hasn't made up her mind about Mohan. More specifically, A hasn't made up her mind as to whether she should accept Mohan's proposal or not. It is this indecision that she expresses to B. B's utterance only confirms A's state of indecision. While in contexts like VEGAN, and RAIN, B's assertion confirmed A's *preference* for an action, in PROPOSAL, B's assertion confirms the lack of preference. This illustrates the importance of the way we defined rank-preservation above.

In defining rank-preservation, we said that the speaker ranking for a salient set of alternative actions does not change once the content of the utterance preceding *voh hi na* is made common ground. Now the speaker ranking may be such that two actions acquire the same order in the ranking. This insight informs our formalization, as we would not always want the action set to be strictly ordered.

(12) **Observation 4**: The contextually determined set of actions need not be strictly ordered for *voh hi na* to be used felicitously.

With all of the above contexts in mind, it may start seeming as if all that *voh hi na* grounds is agreement by the interlocutor. In other words, one can propose that *voh hi na* is felicitous to use only if the prior move expresses agreement with what the speaker of *voh hi na* had said earlier in the discourse. First, such an account would need to employ a rather broad notion of agreement. In (10), B simply adds new information to the common ground, i.e., the trail must be very slippery. This is not how we canonically understand agreement. This can also be noted with other contexts like VEGAN. If such moves by the interlocutor are construed as agreement, they certainly don't target the content of the preceding utterance. For instance, in (10), B, in noting the slipperiness of the trail, doesn't explicitly target the content of *I don't think I'll go to the hike*. In addition, we can construct a dialogue where the interlocutor agrees with the content of what the speaker utters earlier, but to ground with *voh hi na* turns out to be infelicitous.

[WEATHER]: A and B are talking about the weather in San Francisco.

(13) A: The weather here is terrible.
B: I agree!
A: # Voh hi na.

The infelicity of *voh hi na* in (13) illustrates that its felicity conditions cannot be defined by the following requirement only: that the move it grounds expresses agreement with what the speaker of *voh hi na* had said earlier. It may be that the content of the move by B that *voh hi na* grounds ought to be consistent with what A had said earlier in the discourse. However, that is exactly what we have been trying to work towards: a clear and precise understanding of *the way* in which the prior move is consistent with speaker commitments or preferences that the speaker makes public. To that end, we provide our semantic account below.

4 The semantic account

4.1 Nuts and bolts

In the previous section, we noted multiple things along the way in light of the contexts. However, we did not list all of them as observations as in observations 1-4. For instance, in light of HIKE and (10), we observed that the interlocutor whose move gets grounded by *voh hi na* need not have updated their belief state. Moreover, due to (10), we also noted that the speaker who grounds with *voh hi na* need not have a prior degree of positive belief in the proposition that is grounded. The reason for making these notes along the way was to merely point out that these features of the dialogues are orthogonal to the felicitous use of *voh hi na* as a grounding move. Therefore, we didn't highlight these as observations, unlike observations 1-4.

We use observations 1-4 to build our semantic account for *voh hi na*. First, we must get a better understanding of how the four observations are dialectically related to each other. An informal account will fall out of this understanding.

Observation 1 tells us that the speaker of *voh hi na* ought to have made a prior commitment that gets validated by the interlocutor. As is obvious, the notion of prior commitment is underspecified, and the notion of validation is not defined. To specify these two notions in a precise manner, we introduce the notion of an action set. Here, we specify the properties of the action set explicitly.

(14) An action set $\mathcal{A}$ has the following properties:
  a. $\mathcal{A}$ is the set of alternative actions.
  b. $\mathcal{A}$ is relativized to an agent x, as in $\mathcal{A}_x$.
  c. The order on $\mathcal{A}$ may be weak.
  d. $\mathcal{A}$ can include doxastic actions.

The condition of alternative-hood makes $\mathcal{A}$ such that, for ease, one can divide up the action space to include two mutually exclusive actions, along with (an optional) third catch-all OTHER category.[11] This can especially help in formalization. Secondly, the agent-relativization is contextually determined. While in VEGAN, $\mathcal{A}$ is relativized only to A, in STROLL, $\mathcal{A}$ is relativized to both A and B.[12] Here, we don't probe the mechanisms of context-sensitivity involved in determining $\mathcal{A}$. We make the simplifying assumption that a context and a dialogue will provide such a salient $\mathcal{A}$, which will be relativized to an agent. This assumption relies on the dialogue agents' ability to infer such a set. If such an inference is not made and an action set isn't available, our theory predicts that *voh hi na* cannot be used as a grounding move. Further, we note that the order on $\mathcal{A}$ may be weak. This amounts to the feature that an agent can have an absence of preference for two actions in the set. This is illustrated nicely in PROPOSAL/(11). And lastly, we construed *action* to include doxastic actions too.

Note that Observation 3 and 4 fall out of making the notion of action set precise, as in (14-b) and (14-c). Moreover, in §3, Observation 2 was presented as a precisification of Observation 1. In turn, Observation 2 makes reference to $\mathcal{A}$ and Observation 3 and 4 define features of $\mathcal{A}$. That's how all of the observations are dialectically related. Here, we transmute Observation 2 to a semantic account of *voh hi na*.

(15) Statement: A speaker s can felicitously ground u, as said by an interlocutor, with *voh hi na* only if, given a contextually determined $\mathcal{A}$, (i) the speaker ranking $\prec$ over $\mathcal{A}$ is public; (ii) s thinks that, according to the interlocutor, $\prec$ remains unchanged when $[\![u]\!]$ is made common ground.[13]

There are many things to be made precise here. First, we need a notion of publicity of preferences. Second, although we talked about how the discourse participants can reason about each other's information states, we face the challenge of implementing this formally. And thirdly, we haven't said anything about what it means for the ranking to be unchanged. We take on these tasks in the next section and present a formal model.

4.2 The formal model

We use Cariani (2013)'s influential work on deontic modals to couch our account above in an intensional semantics framework.[14] We construe action-types as sets of worlds. More specifically, we can think of actions as functions from agents to sets of worlds. For instance, A's going to the hike can be modeled as the action type *going to the hike* taking A as an argument and yielding the set of worlds where A goes to the hike. As at a given time, an agent can either go to the hike or not, the set of alternative actions specifies a partition over the logical space. This partition divides the logical space such that worlds w and v occupy the same cell in the partition if and only if the agent performs that same action in w and v. Before things start to

[11] This is indeed what Cariani (2013) does. For the most part, we ignore the OTHER.

[12] There's a wide variety of context-sensitive expressions in language (Kaplan, 1979; Lewis, 1981; Kratzer, 1981; Lasersohn, 2005; Stephenson, 2007; MacFarlane, 2014; Jabbar, 2021). Moreover, there's recent work that aims to specify more quantitatively effects of context-dependence for semantic interpretation and pragmatic inference (Beddor and Egan, 2018; Kursat and Degen, 2020)

[13] For any utterance u, we take $[\![u]\!]$ to be its content.

[14] Cariani in turn cites Belnap et al. (2001) as inspiration.

get any more wordy, let's start formalizing.

Where W is the set of all worlds and ΓW is an equivalence class over W, we can define the set $\mathcal{O}_x$ for an agent x as the following equivalence class using an action set $\mathcal{A}$:

(16) $\mathcal{O}_x = \Gamma W$ s.t. for $\gamma \in \Gamma W$, $w \in \gamma$ and $v \in \gamma$ iff $w \in a(x)$ and $v \in a(x)$, where $a \in \mathcal{A}$

In words and more simply, $\mathcal{O}_x$ is the appropriate equivalence class over worlds as specified by action-types taking agents to propositions, where action-types are provided by a contextually determined action-set.

We can certainly rank worlds in the Kratzerian (Kratzer, 1981, 2012) fashion by ordering sources. However, given that *voh hi na*'s felicity tracks preference over a set of actions, following Cariani, we rank the cells of $\mathcal{O}_x$. More importantly, we let this ranking be provided by the preferences of the speaker of *voh hi na* (A in our dialogues). This is what we have been calling *speaker-ranking* all along. We can state this more explicitly below.

(17) $\prec$ is the ordering over $\mathcal{O}$

We can relativize $\prec$ to an agent x as in $\prec_x$ to represent x's preferences over $\mathcal{O}$. In addition to $\prec$, $\mathcal{O}$ can be relativized to an agent too. However, our model only bakes in relativization of $\prec$ to specific agents. Although $\mathcal{O}$ will be relativized to (a group of) agents too, who these agents are will be contextually determined. In VEGAN, it is the speaker. In STROLL, it is both the speaker and the interlocutor. This relativization will be contextually determined.

Below, we can show how preferences over an action-set can be lifted to preferences over a partition $\mathcal{O}$. In (19), we also define the identity relation for preferences, which we use later.

(18) For $\sigma, \tau \in \mathcal{O}$, $\sigma \prec_x \tau$ iff x prefers a_1 over a_2 and $a_1(x) = \sigma$ and $a_2(x) = \tau$

(19) $\prec_1 = \prec_2$ iff for $\sigma, \tau \in \mathcal{O}, \sigma \prec_1 \tau$ and $\sigma \prec_2 \tau$

Moreover, following work in discourse structure and dynamic semantics, we take each discourse participant to be associated with an information state.[15] Discourse moves can be analyzed by their effects on information states. Where p is a proposition and s an information state, we denote update by the following notation:

(20) $s[p]$

The dynamic effect of such an update can be modeled as:

(21) $s[p] = \{w \in s \mid p(w) = 1\}$

We can take $\prec$ to be sensitive to s_x, x's information state. This sensitivity can be denoted by $\prec_{s_x}$. The thought behind this sensitivity is simple; your preferences are determined by what you think the world is like. If you believe that it is raining, you may bring your umbrella with you. Moreover, in §3, we noted that participants reason about each other's mental states in dialogue. Specifically, (ii) in (15) states that the speaker thinks that, according to the interlocutor, the speaker preference remains unchanged. To be able to implement such reasoning about mental states, we introduce the interlocutor's construction of the speaker information state. While for the speaker x, we denote x's state by s_x, the interlocutor's construction of it is denoted by s^i_x. s^i_x will not always be an accurate construction of s_x.

Preferences aren't just sensitive to information states, but crucially to subjects too, as we noted in our discussion of (17). For modeling purposes, one might suggest that preferences come out to be sensitive to subjects by way of being sensitive to their information states. However, such a trickle-down is not ideal. Why not? Given the interlocutor's construction of a speaker's information state s^i_x, we are left with the choice of whether the relativization to the subject should trickle down from s^i_x to i or x. We see no *prima facie* reason for trickle-down to i over x, or *vice versa*. Therefore, for modeling purposes, we let $\prec$ be sensitive to two parameters: an information state and a subject. For instance, $\prec_{x,s^i_x}$ denotes x's preferences given i's construction of x's information state. This is to model x's preferences from the interlocutor's perspective. To illustrate with an example, it's quite possible that B doesn't know that A is vegan. In such a scenario, B may think that *there's a really nice burger-place in the south end of campus* is such that once A's information state updated with its content, A would want to eat on campus. This is a case where the agent for the $\mathcal{O}$ is A. $\prec_A$ over $\mathcal{O}_A$ is also sensitive

[15] Stalnaker (1978) first models the effect on an assertion as an intersective update on the context set. We use just that notion of update and remain agnostic about the sense in which it is dynamic (Rothschild and Yalcin, 2016).

to A in modeling A's ranking over $\mathcal{O}_A$. Moreover $\prec$ here is sensitive to B's construction of A's information state, which is inaccurate in not taking into account that A is vegan, which is a crucial piece of information for A that determines $\prec_A$.

In addition to individual information states, there's a scoreboard that all discourse participants contribute to. More theoretically, this is termed *the common ground*, call it cg for short. cg simply contains all of the propositions that are public knowledge.[16] Stalnaker's formalization of cg, the context set, is achieved by set intersection of all of the propositions in cg. We introduce cg to model the publicity of speaker preference in discourse. We noted this as a requirement for the felicity of *voh hi na* in (15). As we know from Stalnaker (1978), when a proposition p gets added to cg, more than p is added to cg, including the proposition that p has been added to cg. Similarly, we can let the following be a proposition: that x has the preference $\prec_{x,s_x}$ over a set of actions $\mathcal{A}$. This proposition is separate from the preference itself. If this proposition is cg, then we can say that x's preferences w.r.t. $\mathcal{A}$ are public. More explicitly:

(22) x's preferences w.r.t. $\mathcal{A}$ are public iff the proposition that x's preferences over $\mathcal{A}$ are provided by the order $\prec_{x,s_x}$ is common ground.

The above way of modeling the preference being public helps to keep the model simple. Note that while it may be public that x's preferences are provided by $\prec_{x,s_x}$, the information state s may still be private to x. You can refrain from eating meat due to your belief that the meat is contaminated; let it be public that you don't want to eat meat, while keeping private your belief about its contamination.

If we take $[\![u]\!]$, the content of the utterance u that *voh hi na* grounds, to be the antecedent for *voh*, we can take *hi na* to be operating on $[\![u]\!]$. Where $p = [\![u]\!]$, x and i are the speaker and interlocutor respectively, and $\mathcal{O}$ is formed via a contextually supplied $\mathcal{A}$,

(23) *voh hi na* can felicitously ground u only if for $\prec_{x,s_x}$ and $\prec_{x,s^i_x}$ over $\mathcal{O}$, x believes that[17]

a. $\prec_{x,s_x}$ is public
b. $\prec_{x,s_x} = \prec_{x,s^i_x[p]}$

First, note crucially that in (23), both (23-a) and (23-b) are stated in the scope of what x believes. Therefore (23) boils down to the following: (i) the speaker believing that for a salient action, the speaker's preferences w.r.t. it are public; (ii) the speaker believing that the speaker ranking over the action set, given what the speaker knows, is the same as the speaker ranking over the salient action set given what the interlocutor thinks the speaker knows once updated with the interlocutor's contribution. Let's understand (23) even more vividly. Using the two conditions in (23), we walk the reader through two full calculations below.

5 Explaining *voh hi na*

Given that in all our dialogues, A grounds with *voh hi na* and B is the interlocutor, we use s_A to denote the speaker information state and s^B_A for the interlocutor's construction of s_A. Now, we can use our model above to explain the dialogues we presented in §3. First, we take STROLL 1 and 2. In STROLL 1, use of *voh hi na* is felicitous, while in STROLL 2, it isn't. In STROLL 1, A's preference for believing that it's raining is made public by A's utterance *You're being optimistic. It's probably raining outside.* In STROLL 2, there's nothing as such made public. Our account, more specifically (23-a), explains this distribution. Now, the difference in *voh hi na*'s felicitous use in VEGAN and NOT VEGAN can be explained using (23-b). Let's walk through this calculation carefully.

(24) a. The contextually salient action-set is $\{$*eat on campus*, *not eat on campus*$\}$.[18]
b. Both VEGAN and NOT VEGAN are set such that *not eating on campus* $\prec_{A,s_A}$ *eating on campus*. This means that A ranks not eating on campus higher than eating on campus.
c. In NOT VEGAN, B asserts that there's a really nice burger place in the south end. Call this proposition *nice burger*.
d. Once s^B_A—what the interlocutor thinks the speaker information state

[16] To define and even understand common ground is not an easy task. See Lederman (2018) for why the classical ways of understanding common ground may be inaccurate.

[17] We add x *believes that* because it is x who reasons about the interlocutor's contribution and grounds it.

[18] Alternatively, we can say that the action set in VEGAN and NOT VEGAN contains *believing that there's no good food on campus* and *believing that there is good food on campus*. Every set containing actions can be reduced to a a set containing doxastic actions.

is—is updated with *nice burger*, $\prec_{A,s_A^B}$—the speaker ranking, given s_A^B, for what the speaker does—is such that *eating on campus* $\prec_{A,s_A^B}$ *not eating on campus*.
We note in §3 that this action-guiding potential of *nice burger* is what serves as motivation for B to assert *nice burger*.

e. Given (24-b), (24-d), and the identity conditions for any two $\prec$ (cf. (19)), $\prec_{A,s_A} \neq \prec_{A,s_A^B[\textit{nice burger}]}$.

Via (24-e), we witness a direct violation of one of the conditions for felicitous use of *voh hi na*, as outlined in (23-b). Thus, our account correctly predicts that *voh hi na*'s use to ground *nice burger* would be infelicitous in NOT VEGAN. We get the same calculation for RAIN, HIKE, and PROPOSAL. In RAIN, the preference for believing that it is raining is reified. In HIKE, the preference for not going to the hike is preserved. In PROPOSAL, the state of indecision survives a tie. Each of these contexts differs from each other in some way. Then, (23)'s ability to explain these contexts at least suggests that our model specifies felicitous uses of *voh hi na* at the right amount of fit.

6 Conclusion

In this paper, we have presented one way of thinking about a grounding move. Our inquiry was guided by the following observations. There's an utterance that ought to have been made prior to *voh hi na*. *voh hi na* itself involves a pronoun *voh*. We made the safest assumption that the proposition expressed by the prior utterance serves as the antecedent proposition for *voh*. This led us to analyze the two appended particles in a non-compositional way where we understood the contribution of *hi na* as establishing a relation between the antecedent proposition and a feature of the discourse structure. Given that we didn't decompose *hi na*, our account comes out to be partly compositional. Although the current account explains the data presented, a more complete account will seek to derive the meaning contribution of (23) entirely compositionally. This will give us a nice insight into how meanings of multiple discourse particles can compose with each other. Such a line of inquiry is exciting also because we can infix another particle *toh* to form *voh hi toh na*. It would be interesting to see if the felicitous distributions of *voh hi na* and *voh hi toh na* vary, which can help us understand discourse particles better, especially the infixed *toh*. Ideally, a compositional account for *voh hi na* would propose individual meanings for particles that don't diverge too much from the ones proposed in the literature already. In the other direction, semantic accounts of particles in the literature can be tested against their ability to derive the felicitous distribution of *voh hi na*. On the theoretical side, our paper adds to the body of work on linguistic expressions that are sensitive to decision-problems in context. In this vein, our work is most comparable to Davis (2009)'s work on the Japanese discourse particle *yo*.

Acknowledgments

Special thanks to Judith Degen for her feedback on this paper! Thanks also to Ashwini Deo, Xuetong Yuan, and two anonymous reviewers for their insightful comments and time!

References

Bob Beddor and Andy Egan. 2018. Might do better: Flexible relativism and the qud. *Semantics and Pragmatics*, 11(7):1–44.

Nuel Belnap, Michael Perloff, and Ming Xu. 2001. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press.

Anton Benz. 2006. Utility and relevance of answers. In *Game theory and pragmatics*. Springer.

Sakshi Bhatia and Rajesh Bhatt. 2023. Copular agreement in Hindi-Urdu. *Glossa: a journal of general linguistics*, 8(1).

Rajesh Bhatt. 1994. The semantics of *hi*. *Unpublished ms, University of Massachusetts, Amherst*.

Rajesh Bhatt and Veneeta Dayal. 2020. Polar question particles: Hindi-urdu kya. *Natural Language & Linguistic Theory*, 38:1115–1144.

María Biezma, Miriam Butt, Farhat Jabeen, and Benazir Mumtaz. 2022. Urdu/hindi polar kya as an expression of uncertainty. *Manuscript submitted for review, https://lingbuzz. net/lingbuzz/006997*.

Madelaine Brown. 2022. *Naan* as a tag question and a discourse marker in hindi-urdu. *Proceedings of Formal Approaches to South Asian Languages (FASAL)*, 10:1–12.

Fabrizio Cariani. 2013. *Ought* and resolution semantics. *Noûs*, 47:534–558.

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

Cleo Condoravdi and Sven Lauer. 2012. Imperatives: Meaning and illocutionary force. *Empirical issues in syntax and semantics*, 9:37–58.

Christopher Davis. 2009. Decisions, dynamics and the japanese particle yo. *Journal of semantics*, 26(4):329–366.

Ashwini Deo. 2022. Could be stronger: Raising and resolving questions with hindi= *to*. *Language*, 98(4):716–748.

Ashwini Deo. 2023a. Coordinated on the context: the many uses of marathi= ts. *Formal Approaches to South Asian Languages*.

Ashwini Deo. 2023b. Preference for independent shared commitments. *Presentation at 41st West Coast Conference of Formal Linguistics (WCCFL 41)*.

Donka F Farkas and Kim B Bruce. 2010. On reacting to assertions and polar questions. *Journal of semantics*, 27(1):81–118.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation*, 1:221–237.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Christine Gunlogson. 2008. A question of commitment. *Belgian Journal of Linguistics*, 22(1):101–136.

Ahmad Jabbar. 2021. Pluralism for relativists: a new framework for context-dependence. *Logic and Engineering of Natural Language Semantics (LENLS)*, 18:3–16.

Ahmad Jabbar. f.c. The Hindi-Urdu *NA* and reasonable inference. *Proceedings of the 59th meeting of the Chicago Linguistic Society (CLS 59)*.

David Kaplan. 1979. On the logic of demonstratives. *Journal of philosophical logic*, 8:81–98.

Magdalena Kaufmann. 2011. *Interpreting imperatives*, volume 88. Springer Science & Business Media.

Jeffrey C King. 2001. *Complex demonstratives: A quantificational account*. MIT Press.

Niko Kolodny and John MacFarlane. 2010. Ifs and oughts. *The Journal of philosophy*, 107(3):115–143.

Angelika Kratzer. 1981. The notional category of modality. *Words, worlds, and contexts*, pages 38–74.

Angelika Kratzer. 2012. *Modals and conditionals: New and revised perspectives*, volume 36. Oxford University Press.

L Kursat and J. Degen. 2020. Probability and processing speed of scalar inferences is context-dependent. In *CogSci*.

Peter Lasersohn. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and philosophy*, 28:643–686.

Harvey Lederman. 2018. Uncommon knowledge. *Mind*, 127(508):1069–1105.

David Lewis. 1981. Index, context, and content. *Philosophy and grammar: Papers on the occasion of the quincentennial of Uppsala University*, pages 79–100.

John MacFarlane. 2014. *Assessment sensitivity: Relative truth and its applications*. OUP Oxford.

Elin McCready. 2008. What man does. *Linguistics and Philosophy*, 31:671–724.

Jessica Rett. 2011. Exclamatives, degrees and speech acts. *Linguistics and philosophy*, 34:411–442.

Tania Rojas-Esponda. 2014. A discourse model for überhaupt. *Semantics and Pragmatics*, 7:1–45.

Daniel Rothschild and Seth Yalcin. 2016. Three notions of dynamicness in language. *Linguistics and Philosophy*, 39:333–355.

Robert C Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics*, pages 315–332. Academic Press.

Tamina Stephenson. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics & philosophy*, 30(4):487.

Matthew Stone and Alex Lascarides. 2010. Coherence and rationality in grounding. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 51–58. Poznán Poland.

Nadine Theiler. 2021. *Denn* as a highlighting-sensitive particle. *Linguistics and Philosophy*, 44(2):323–362.

Robert van Rooij. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy*, 26:727–763.

Xuetong Yuan. 2020. Challenging the presuppositions of questions: the case of ba-interrogatives. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 469–484.

Malte Zimmermann. 2011. Discourse particles. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An international handbook of natural language meaning*, volume 2, page 2012=2038. Walter de Gruyter.

Exploring the Semantic Dialogue Patterns of Explanations – a Case Study of Game Explanations

Josephine B. Fisher
Psycholinguistic
Paderborn University
TRR 318 Constructing Explainability
j.b.fisher@upb.de

Amelie S. Robrecht
Social Cognitive Systems
Bielefeld University
TRR 318 Constructing Explainability
arobrecht@techfak.de

Stefan Kopp
Social Cognitive Systems
Bielefeld University
TRR 318 Constructing Explainability
skopp@techfak.de

Katharina J. Rohlfing
Psycholinguistic
Paderborn University
TRR 318 Constructing Explainability
katharina.rohlfing@upb.de

Abstract

Contributing to the research on social design of explainable AI, we studied 51 German dyadic explanations to reveal how an explanation process is unfolding and to what extend both, the explainer (EX) and the explainee (EE) are contributing to the content. In this paper, we exploratively examine semantic dialogue patterns of semi-naturally and spontaneously occurring explanations of the game Quarto, which are – compared to an expert explanation – less restrictive. We apply the notion of explanation nodes to identify explanation blocks as well as their order that constitute the internal structure of these explanations. In particular, we analyse which information is covered by an explanation dialogue in terms of both, coverage and frequency. Our results reveal the engagement of both interlocutors and provide a basis for the study of adaptivity in explanations and its realisation in dialogue systems.

1 Introduction

Explanations provide an interesting case for the study of how semantic structure is built up during a dialogue: As explanations have the goal to result in understanding, it is reasonable to assume that both partners need to contribute to the structure (Rohlfing et al., 2021). However, little is known about how this joint co-construction unfolds. At the same time, there is a growing need for understanding how explanations succeed. In the last years, Explainable AI (XAI) is driven by the General Data Protection Regulation (GDPR) (Carey, 2018) and the right to have an algorithm explained. This is particularly the case for "blackbox" machine learning algorithms. While many approaches to how to make the blackboxes inspectable exist, the process of explaining, i.e., the way of how to present the relevant content (the *explanandum*) and how to ensure sufficient understanding of it, receives little attention (Anjomshoae et al., 2019). The research area of XAI seems to be unbalanced, prioritising what aspects and features to explain instead of how to explain (Baniecki et al., 2023). Thus, empirically-driven studies are demanded to address the research gap from the perspective of a more user-centred and social interaction (Madumal et al., 2019). To create systems that are adaptive and provide an explanation that addresses the users' knowledge gap, it is crucial to explore how humans achieve an adaptive process when interacting with each other in explanatory dialogues.

In our investigation we address this gap by focusing on the co-constructive character of explanations, subscribing to the view that explanation is a social and co-constructive process (Rohlfing et al., 2021; Miller, 2018). How this co-constructive process is reflected in the dialogues can be addressed by contrasting the distributions of semantic contributions of the interlocutors. Thereby, we take into account the influence the EE and EX can take within a dyad. This allows both interaction partners to shape the content of the discourse. Who is planning and structuring and who is confirming the explanation?

While there is a well established research focus on modelling the structures of direction-giving (guiding a person to a specific place via verbal instruction) by extracting different phases out of human-human interactions (Psathas and Martin, 1976; Ewald, 2010), there is little done on spontaneous explanation dialogues. Due to this research gap, this paper will describe the semantic dialogue patterns of human-human everyday explanations to point out reoccurring structures. By introduc-

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

ing an explanation node scheme, we also allow a more fine-grained semantic analysis. To reach this goal, we combine linguistic analyses with methods from computer science to work towards an implementation of humans' adaptive capabilities. The linguistic analyses focus on the explanation structure by introducing an explanation node scheme where each explanation node – which is the smallest unit in the system (see Sec. 3.3) – captures a semantic dialogue pattern which can be observed in the interaction. We employ this explanation node scheme to study the semantic dialogue structure of explanations between two interlocutors engaged in explaining a board game. The structures and relations that are represented by the explanation node scheme can be transferred to an ontology and, e.g., serve as a knowledge base in an adaptive explanation dialogue system.

Based on current research, we expect a game explanation to be sequential and co-constructive. (1) Concerning sequentiality, we expect sequential patterns comparable to the phases in direction-giving introduced by Psathas and Martin (1976). In addition, because the setting is eliciting everyday explanations, (2) we expect the EE to be an active participant (Rohlfing et al., 2021; Fisher et al., 2022) having the opportunity to introduce explanation nodes on their own. (3) Based on Rohlfing et al. (2021) and Miller (2018), we further expect the explanations to be co-constructive. For that, we will investigate the EE's contributions and how they are addressed jointly. If the EE is the first to mention an explanation node, we expect the EX to take it up.

2 Background and Related Work

Much work on how information is established during an interaction was characterised by Clark (1996) as introduced by his theory of *common ground*. It displays how conversational partners agree on their shared information, during the course of an interaction. Any type of discourse is a joined activity in which the common ground between interlocutors increases, and in which "sections and subsections [are]n't fixed beforehand, but [are] negotiated as [they go] along" (Clark, 1996, p.36). This includes "the knowledge, beliefs, and suppositions they believe they share about the activity" (Clark, 1996, p.38).

2.1 Structures of Explanations and Tutoring

Taking a broader perspective towards human explanatory dialogue, each explanation involves two conversational partners with an asymmetric knowledge distribution: an EX, who is more knowledgeable, and an EE, who is less knowledgeable. The subject of the explanation is the so-called *explanandum* which is constituted by different types of *explanans* (Rohlfing et al., 2021). Looking at an explanation as a process, it unfolds because the EX and the EE work together on specifying what information is needed for the EE to understand (Klein, 2009) as well as what is or should be the object of explaining (*explanandum*). Klein (2009) claims that there are several types of explanations; they relate to the How, the Why and the What. Scientific explanations rather focus on the Why, whereas everyday explanations reveal a variety in their types. The subtype of everyday explanations we are focusing on, are game explanations which cover different aspects, such as rules, figures and the game board. Kotthoff (2009) classifies game explanations in more detail as procedural explanations. This goes in line with the categorisation of Klein's (2009) definition of How explanations.

One can define an explanation process as a sequence of phases that contain explanation and verification blocks (El-Assady et al., 2019). How to find the optimal order of these blocks and which explanation strategy to choose depends on the level of detail, the EE, and the desired amount of interactivity. In this paper, we investigate such explanation blocks in human-human explanations and study how to extract their internal structure from explanation dialogues. An explanation involves two processes, the cognitive process, which can be described as the planning and construction of the explanans, and the social process, which focuses on the interaction between the EX and the EE (Miller, 2018). This paper will put the spotlight on the explanation as a conversation, by focusing on the content structure.

Similar to explanation, in the context of tutoring, a knowledge asymmetry exists. However, the addressee is supposed to learn, which is not necessarily the case in explaining, where the focus is on understanding or enabling (Rohlfing et al., 2021). Research on tutoring (Chi et al., 2008; Miyake, 1986) has established mind maps, in which the different elements that are part of a topic are listed and numbered in individual nodes, to account for the contents which were already discussed and understood in a conversation. These mind maps differ from what is known in linguistics as semantic maps that sound similar. Haspelmath (2003) has proposed the semantic map method that displays the lexical relatedness of words. It uses graphs to present relatedness of co-expressed meanings, connecting nodes by edges to describe which concepts can be expressed by the same words. However, it does not focus on the semantic relatedness of the explanation elements and is thus little relevant for the idea of the mind map. Explicitly in the work by Chi et al. (2008) on scientific explanations, the problem solving nodes were based on the verbal explanations of the tutors when they explained the steps alone. There, the individual nodes relate to a problem solving step. Miyake (1986) similarly listed the different elements in a hierarchical fashion which belong to a problem regarding the stitches of a sewing machine. For this purpose, the framework was called "the function-mechanism hierarchy". In this, the contents are differentiated in two ways. They address the function – what is taking place – and the mechanism – how it is performed. They are in such a way connected that the mechanism at a lower level is needed to describe the function of the next higher level. Here, the categorisation of the elements and its level of detail is justified as being appropriate to examine the ongoing process of understanding (Miyake, 1986).

2.2 Models in Computer Science and XAI

In contrast to the previously introduced node system, using an ontology or a knowledge graph (KG) to store information is a common method in dialogue systems (Axelsson and Skantze, 2023; Robrecht and Kopp, 2023; Axelsson and Skantze, 2020; Ma et al., 2015; Lin et al., 2015). The KG concept was first introduced by Minsky (1968), who called them *semantic networks*. Today it is used in approaches such as *the semantic web* (W3C, 2012) or *Wikidata* (Vrandečić and Krötzsch, 2014). The domain that is stored using a KG varies from scientific publications or E-commerce to social networks and geopolitical events (Kejriwal et al., 2021). While most ontologies are defined by the Resource Description Framework (RDF), other approaches or variations – such as Resource Description Framework Schema (RDFS) – are used. We will focus on RDF, as popular languages, such as the *Ontology Writing Language* (OWL) (W3C, 2012) derives from it. An RDF graph consists of a set of triples, each consisting of a *subject*, an *object* and the connection *predicate*. In other words: Two entities (subject and object) are connected via a relationship (predicate). Further information and restrictions on designing an RDF ontology can be found in Kejriwal et al. (2021). The subject, object or predicate – the smallest unit in an ontology – captures only one single entity or relationship. Therefore, a node might be, but not necessarily has to be, broken down into multiple triples, when transforming an explanation node scheme into an ontology.

In human robot interaction (HRI), the majority of research aims to create *Explainable Agency* or *Goal-Driven XAI*. As the agent explains behaviour and decisions, the interaction becomes predictable to the user (Anjomshoae et al., 2019). Next to predictability, understandability is a key goal when thinking about explainable agency. Both can be increased by improving the agents human-likeness. By looking at the processuality of human-human explanations, we aim to find patterns that can be transferred to HRI settings at a later state. Currently, effects on communication and explanation structure are usually measured using interaction studies (Stange and Kopp, 2020, 2021). Subsequently, the explanation is adapted to the best condition. There is research that uses a bottom-up approach by analysing multiple explanation interactions for their model (Madumal et al., 2019), but none of the considered dialogue types, the model is based on, are verbal everyday human-human explanations. Nevertheless, in the final study the agent performs an explanation on the board game "ticket to ride", which can be considered an everyday explanation in an agent-human setting. Most of the current papers define the communication of the explanation as their most important future work project (Anjomshoae et al., 2019).

3 Method

3.1 Participants

A subset of 51 game explanation dyads with a total of 102 participants from the *ADEX* (Adaptive Dialogical Explanations) corpus, which we collected in the project A01 *Adaptive Explanation Generation* in the *TRR 318 Constructing Explainability*[1], was considered. In the recorded (video and audio)

[1] `https://trr318.uni-paderborn.de/en/projects/a01`

Figure 1: Study design of ADEX corpus[2]

dyadic explanation dialogues 60 female, 38 male and three diverse subjects took part (age range: 18-55 years). 96 participants were native German speakers and five were second language speakers. Lastly, 94 of them were students and seven had other occupations [3]. The study was conducted in six phases (Fig. 1). Phase 1, 3, 5 and 6 were different questionnaires, which included psychological and understanding instruments. In Phase 2, the participants took part in the explanation without the game being present. Before the study, the EX was asked to learn the game Quarto. Quarto is a strategic board game that includes game figures with four different characteristics. The goal for each player is to place four figures in a row that have one of those characteristics in common. They were free to use any resources they liked for their preparation. We provided them with some exemplary sources and the possibility to take a look at the physical game before the study. After the first phase, the EX was instructed to spontaneously explain the game in such a manner that the EE would have the chance to win the game. The EE was told to actively take part in the explanation. The participants had no time restrictions for the explanation phase. Consequently, the explanations can be considered diverse because the subjects were free in their preparation of the game and their speech. In Phase 4, the dyads were instructed to play a couple of games of Quarto and to continue explaining. This phase was excluded in the current analysis.

3.2 Linguistic Coding

To explore the semantic dialogue structure of the explanations, we coded the speech according to their content with the program *ELAN* (Wittenburg et al., 2006)[4]. Therefore, we adapted the node scheme from scientific explanations to game explanations of Quarto [5]. The speech is divided into moves by the conversational partners. Following the work of Chi et al. (2008) moves are defined as statements including a single idea presented by a single speaker within one turn. Thus, the explanation nodes serve as a foundation for the speaker move analysis. Backchannels (such as *mh, yeah and okay*) are not considered in the analysis because they do not function as separate turns that attempt to take the conversational floor (Dideriksen et al., 2019). For the reliability check, six explanations (about 12% of the data) were coded concerning the blocknodes by two researchers. Thereby, an unweighted Cohen's kappa yielded an inter-coder reliability that can be considered almost-perfect (Landis and Koch, 1977) (k=0.90). The majority of mismatches related to the count of the parent - when one of their childnodes was discussed. Henceforth, deviations between the two coders were smoothed via post-hoc agreement. Based on this, the analysis of the whole data set was adjusted.

3.3 Explanation Structures

In contrast to the hierarchical and sequential order of scientific explanations, game explanations occur in a more flexible manner. In our approach, the semantic dialogue structure is captured in an explanation node scheme. Each *explanation node* captures specific semantic information. The explanation nodes are connected via arrows, whose direction represents an increase of detail. A parentnode is an explanation node at an upper level, while the next more detailed explanation node connected by an arrow is referred to as a childnode. A group of explanation nodes referring to the same semantic category form an *explanation block*, the highest node in a block is called *blocknode*. Together, the explanation nodes form a map that can be revisited

[2]©Paderborn University, Patrick Pollmeier

[3]One data point each is missing due to technical problems.

[4]Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands https://archive.mpi.nl/tla/elan

[5]The preliminary ADEX Codingscheme for Explanation Nodes can be found at https://go.upb.de/ADEX_Explanation_Nodes.

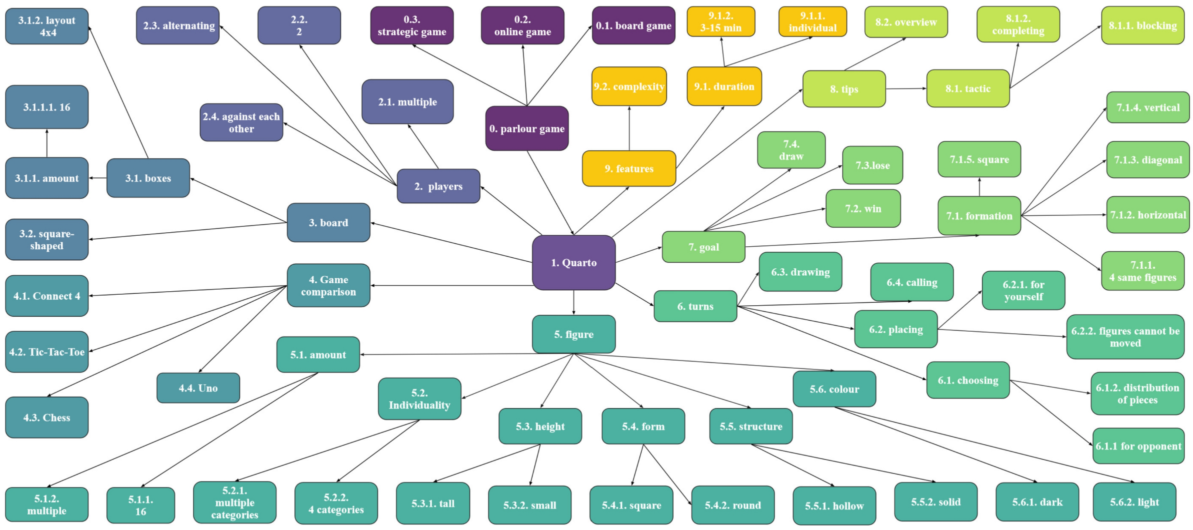


Figure 2: Node scheme: Each block is represented by a specific colour. The colour coding is consistent within all presented figures.

by the interlocutors (Fig. 2).[6] In contrast to (Chi et al., 2008), in which the tutor's explanation served as the only source for the node system construction, we instead used the explanation nodes in interaction. First, the blocknodes were established and in an iterative process the subnodes were added. We adjusted the level of detail to the topical occurrences in the data. As one can see, the game explanations cover ten blocknodes divided further in several subnodes. Taken together, 69 explanation nodes were identified.

The *Quarto* block only contains one node, its name, all the other nodes are placed around this central node. In the *Parlour Game* block, the game is put into the broader game context. All information related to the players, how many there are and in which mode they play, are grouped in *Players*. The third block captures the different characteristics of the *Board*. A special block is the game *Comparison* which contains the games that are frequently compared to Quarto. *Figures* is the largest block describing the characteristics of the game pieces. In *Turns*, the required game turns are listed and *Formations* names the possible formations of the figures and their impact on the goal of the game. Tactical tips are depicted in block *Tips* and the final block, *Features*, includes general features of the game, such as duration and difficulty. The block dependency is expressed through the colours, while each node has its own reference number.

[6]In the empirically developed explanation node scheme nodes were divided in subnodes if mentioned separately.

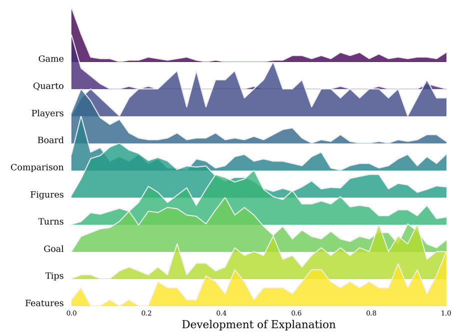


Figure 3: Reference to blocks by EX in relation to the time in all dialogues

4 Results

4.1 Order and Sequentiality

As previously introduced, we hypothesise the explanation blocks occur in certain patterns. These patterns will be described by focusing on the order the explanation blocks and nodes are either introduced or mentioned in. The order in which the blocks are mentioned by the EX can be seen in Figure 3[7].

It becomes apparent that the blocks *Game* and *Quarto* – if mentioned at all – are discussed in the very beginning of the explanation. The blocks *Board* and *Figures* are discussed subsequently, followed by the blocks *Goal* and *Turns*. The expla-

[7]The length of the interaction is normalised and the frequency of appearance is normalised for each block independently

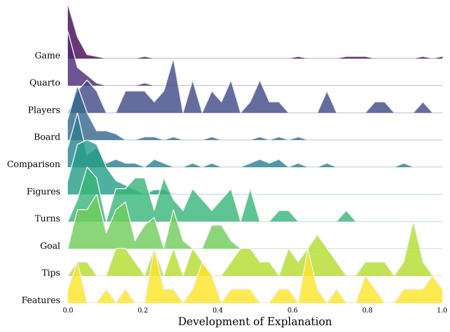


Figure 4: Introduction of the blocks by EX in relation to the time in all dialogues

nation is typically closed by referring to *Tips* and *Features* of the game. The block *Players* is not as explicitly connected to a specific part of the explanation; it can be addressed in the very beginning or at the end of the explanation or at both times. Apart from the discussed blocks, there is one block, the *Game Comparison*, that can be relevant at each state of the interaction. This shows how comparisons differ from the other blocks, as a *Game Comparison* unites nodes by their function and not primarily by their semantic meaning. Figure 4 displays the occurrence of a block being mentioned by the EX for the first time. Especially the blocks that are discussed in the beginning, such as *Parlour game*, *Quarto*, *Board* and *Figures*, are typically introduced in the beginning as well. The moment the block *Players* is mentioned first, shows a higher variance. Some explanations refer to the block *Players* at an early stage, while others first mention the block only in the second half of the explanation. Blocks that are discussed in the second half of the explanation, such as *Turns*, *Goal*, *Tips* and *Features*, are nevertheless often already introduced in the first half.

When distributing the explanation nodes separately (App. A Fig. 6), it becomes clear, that explanation nodes connected to certain blocks, such as *Figure* or *Goal*, tend to be explained close together at more or less the same place in the discourse, while blocks such as *Players* or *Turns* are spread over the whole interaction. This can be explained by either the fact that they are mentioned several times, as their semantic connection to other blocks is very strong, or that the order of the blocks differs in each dialogue. The explanation blocknode *2.0. Players* is not mentioned by any EX. A reason for this might be that the EX prefers the other – more detailed – explanation nodes of the block. Considering the individual explanation nodes, helps to understand, why the block *Game Comparison* is spread over the whole explanation. There are explanation nodes in the comparison block, that appear close to others due to their semantic relation (e.g. *4.4. Uno* and *6.4. Calling*): Similar to Uno, one also has to verbally indicate in the game that one has won. In Example 1 the EE notices the upcoming game comparison and brings in the name. Thereby, they co-construct the explanation and the EE displays their active participation.

Example 1 from D02

EX: Äh und dann ja hat man das Spiel gewonnen also es ist nen bissel dieser Ausruf kennt man ja so [von] genau von Uno letzte Karte.
Uh, and then yes, you won the game, so it's a bit like this exclamation that you know [from] exactly from Uno last card.[a]

EE: [Uno]
[Uno][a]

[a]English translation of the German transcripts.

Other comparisons, such as *4.2. Tic-Tac-Toe* or *4.1. Connect Four* can be used to compare multiple aspects of the game, as they have several semantic relations to Quarto.

4.2 Coverage and Frequency

In the following, the coverage and frequency analysis of the explanation nodes will be presented[8]. This includes answering the questions: (a) How many explanation nodes are addressed in the explanations and (b) How often is an explanation node addressed in an explanation (and by whom)? Turning to the coverage of the explanation nodes by the conversational partners. On average, the EX mentions 49% (min. 33% and max. 67%, SD = 8.0) of the explanation nodes in their explanations. In other words, about half of all explanation nodes are covered by the EX in the explanations. In contrast, the EE addresses on average 20% of the explanation nodes (min. 4% and max. 48% , SD = 11.0). Therefore, the EE relates to the explanation nodes less frequently and contributes less to the overall map coverage. We will now take a look at how the individual explanation nodes are covered in coverage and frequency. There are ex-

[8]For the analysis ELAN Annotation Frequency and Coverage (Biermeier, 2023) was used.

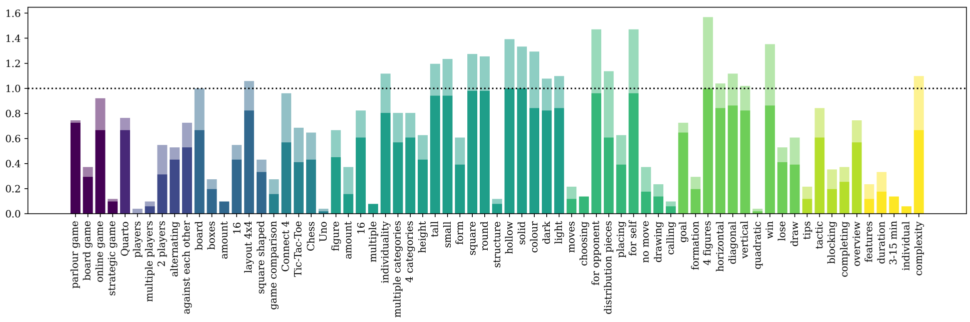


Figure 5: Explanation node coverage by EX (bottom bar) and EE (top bar) - each bar displays in how many of the dialogues the explanation node is mentioned. It shows the proportionate occurrence of the explanation nodes in the entire data set.

planation nodes in each block that are covered in almost every explanation, while others are discussed rather sparsely. When looking at the frequency of the explanation nodes (Fig. 5), it becomes clear that neither the more general blocknode [9] nor the more specific childnodes have a higher frequency of being mentioned. No explanation node specific patterns can be found, but block specific tendencies are observable. When describing the categories of figures, in three of four cases the contrastive characteristics are used more often (*5.3.1. tall - 5.3.2. small* (94.12%) – *5.3. height* (43.14%), *5.4.1. square -5.4.2. round* (98.04%) – *5.4. form* (41.18%), *5.5.2. solid - 5.5.1. hollow* (100%) – *5.5. structure* (9.8%), while in one case the category is used slightly more often (*5.6. colour* (88.24%) *5.6.1. dark - 5.6.2. light* (82.35%-84.31%). In general, the more detailed contrasting information is preferred. The comparison that is used in most explanations is *4.1. Connect 4*, which is mentioned in 60.78% of the explanations, while *4.4. Uno* is only used as a comparison in 1.96%. The explanation nodes that are mentioned in every explanation by the EX (Fig. 5 EX-darker bar) are *5.5.1. hollow, 5.5.2. solid* and *4 Figures*. There is no explanation node that is addressed in every explanation by the EE. The explanation node with the highest frequency is *7.1.1. 4 Figures* (Fig. 5 EE-lighter bar). The explanation nodes that are mentioned in less than 10% of the explanations are: *2. Players* (3.9%), *2.1. multiple (players)* (7.84%), *4.4. Uno* (1.96%), *5.1.2. multiple (figures)* (7.84%), *5.5. structure* (9.8%), *6.4. calling* (7.84%), *7.1.5. square* (1.96%) The option to arrange the figures in a quadratic shape is an optional rule and is not

[9]There is only one block where the blocknode has the highest frequency (*0. parlour game*).

Example 2 from D36

EE: Wie lange dauert das?
How long does it take?[a]
EX: Ne Runde höchstens zehn Minuten.
One round, ten minutes at the most.[a]

[a]English translation of the German transcripts.

captured in every external explanation of the game. As each participant was supposed to learn Quarto in advance with a source of their choice, this might be the reason for the low coverage. and *9.1.1. individual* (5.88%).

When looking at the frequency of an explanation node in a dialogue, it is considered to be discussed in depth, if it is mentioned more than five times by either the EX or the EE. There are only three explanations where no explanation nodes are discussed in depth and each block has at least one explanation node that is deeply discussed in either of the dialogues. Especially the fact, that a line needs four figures and that the figures are picked for the opponent are deeply discussed in more than half of the explanations (Tab. 1). The other explanation nodes that are deeply discussed occur in fewer explanations. They occur in a range of six till eighteen explanations. Overall, 221 times an explanation node is discussed in depth. In 95.48% of these, the EX is referring to the explanation node more often, than the EE. In these cases, the EE is rather passive. Nevertheless, there are explanations with a highly active EE. On the one hand, the EE can contribute nearly as many moves as the EX. On the other hand, the EE can introduce new explanation nodes (see example 2). Table 2 displays all explanation nodes that were referred to in more depth

#Dialogues	Label	Node
28	7.1.1.	4 Figures
28	6.1.1.	For Opponent
18	6.2.1.	For Self
16	5.5.1.	Hollow
12	7.2.	Win
11	5.4.2.	Round
11	5.6.	Colour
10	5.4.1.	Square
9	5.3.2.	Small
8	5.3.1	Tall
8	5.5.2	Solid
7	5.6.2	Light
6	8.1.	Tactic

Table 1: Number of explanations an explanation node is discussed in depth (>5) by either of the interlocutors

D-Number	Label	Node	EX	EE
D16	5.2.	Individuality	5	7
	5.5.1.	Hollow	5	5
	5.5.2.	Solid	5	11
D17	7.2.	Win	3	5
	8.1.	Tactic	3	6
D23	4.1.	Connect 4	2	6
	8.1.1.	Blocking	2	5
D42	5.2.	Individuality	3	7
D49	4.3.	Chess	2	5
	5.1.1.	16	5	5

Table 2: Dialogues in which explanation nodes are mentioned more frequently by the EE

by the EE than by the EX. Both, visualisations and examples show, how much the explanations differ from each other concerning their coverage and frequency. Finally, in our analysis, we addressed the question how explanations are co-constructed. For this purpose, the explanation nodes by the EE were analysed in detail to investigate which explanation nodes were introduced by the EE and whether the EX addressed these and when. In almost all of the dyads (50/51), the EE introduced a new explanation node. On average, the EEs initiated 4.2 new explanation nodes in a conversation (min. 0 and max. 10, SD = 2.6). Out of 212 explanation nodes that were introduced by the EE in the whole data set, the EX took up the explanation node directly 152 times (72%); 19 times (9%) they did not directly address the explanation node, but later on in the conversation. In 41 cases (19%), the EX did not take up the explanation node at all. Out of the 69 explanation nodes in total, the EEs introduced 48 (69%) throughout the different dyads. The results taken together show that an explanation is a unique interaction and highly depends on both conversational partners.

5 Discussion and Conclusion

In this paper, we introduced an explanation node scheme as a tool to model and explore the semantic dialogue structures of explanations. This tool allowed us to investigate the contributions of both dialogue partners to the domain knowledge. Concerning our research question (1), we were able to show that a game explanation is a sequential interaction. Nevertheless the patterns are not as restrictive as in a scientific explanation. A reason for this is likely to be the active participation of the EE, which we addressed in research question (2). In contrast to this, in an everyday explanation, the EE can be more active by demanding a more detailed explanation or pointing out knowledge gaps. In more naturally occurring explanations Fisher et al. (2022) also found a lot of variance in interaction patterns.

Further and with respect to research question (2), we expected the EE to be actively involved. We found support for this in our data showing that in each of the dyads in the corpus, up to 48% of the explanation nodes were covered by the EE showing also a high variance in the EEs' verbal contributions (Fisher et al., 2022). For future work, we hypothesise that the more active the EE is in the explanation, the less predictable it is to the EX, who has to adapt their explanation accordingly. This might account for why the sequentiality of the explanation nodes varies, even in our semi-natural game explanations. For naturally occurring explanations, we expect a higher variance in the contribution of the EE. This highlights the need for adaptive dialogue systems.

In research question (3), we set out to examine the relationship between the explanation nodes introduction by the EE and their uptake by the EX. The findings regarding the explanation node introduction by the EE being uptaken by the EX indicate that an explanation is a joined activity (Clark and Schaefer, 1989) in which the conversational partners co-construct their content (Rohlfing et al., 2021). To what extent the mentioned explanation nodes correlate with the EX's speaker moves, should be investigated in the future to provide more foundations for adaptive dialogue systems.

To conclude, based on first exploratory empirical results, we were able to display the content of explanations via explanation nodes. Thereby, we highlighted the active involvement of the EE by their explanation node introduction. The co-

construction of explanations is demonstrated by the take up of the explanation nodes by the EX.

The next steps in the linguistic analysis are: (1) to connect the explanation nodes with the verbal behaviour (speaker moves) of the conversational partners, as it was done in the works of Miyake (1986) and Chi et al. (2008). By making use of the nodes, one can keep track of the interaction history, i.e., the progress of the dialogue. Hereby, the explanation nodes can serve as a tool to support the future speaker move analysis because one is capable of telling whether information was already discussed and compare whether it has been modified. Chi et al. (2008) adds the concept of substantiveness to the contributions of the conversational partners. We hypothesise that the explanation nodes will correspond to this concept. This can be considered in future analyses. We only considered how the EX takes up the the explanation nodes the EE brings into the explanation and not all of their contributions. This could be an additional step for further analyses. When taking the modelling of human-agent explanation into account, the results will also be beneficial to the enhancement. The observed semantic dialogue patterns will be implemented into the dialogue system *SNAPE* (Robrecht and Kopp, 2023) . The order of the blocks will be used to define transition probabilities for a high level semantic decision process.

6 Limitations

We have to stress that because little is known about semantic structure being built by both partners during explanations, we followed an explorative approach. In our current analysis we excluded backchannels because they do not attempt to take the conversational floor. Nevertheless, backchannels might contribute to the dialogue. We attempted

Current Node *t*	**Next Node** *t+1*	**Frequency**
For Opponent	For Yourself	55
Square	Round	40
Vertical	Horizontal	35
For Opponent	For Opponent	32
Light	Dark	29
Hollow	Round	25
Tall	Small	23
For Yourself	For Yourself	23
Quarto	Parlour Game	21
Individuality	Individuality	21

Table 3: Explanation nodes with a cooccurence > 20

to find clusters in the explanation nodes by seeking high frequent bigrams, to not only see patterns in the block, but also in the explanation node sequences. The combinations in Table 3 were the ones that appeared more than 20 times. Some are repetitions of the same explanation node which can be interpreted as a deeper discussion of a particular explanation node. The others with a strong semantic connection are the (contrastive) characteristics for the figures and the categorisation that Quarto is a parlour game. With the exception of these bigrams, we were unable to find sequential patterns on the explanation node level. This can be either due to the interlocutors' co-construction or due to the size of the dataset. Following the first assumption, it could be that in expert explanations that occur without the involvement of the EE and in a more monological form (Klein, 2009), more patterns on the explanation node level can be found. With the current data size and method, we cannot provide clear indications. It might be possible to find patterns on the explanation node level within the explanation of the EX, if one controls the behaviour of the EE. Thereby, the influence of the EE on the explanation dialogue can be minimised. As we analyse only a subset in this paper, a next step is to expand the analysis to the whole study and with similar data from other projects.

Ethics Statement

The study with adult participants was approved by the Paderborn University Ethics Committee.

References

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*, volume 19, Montréal, Canada.

Agnes Axelsson and Gabriel Skantze. 2023. Do You Follow?: A Fully Automated System for Adaptive Robot Presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 102–111, Stockholm Sweden. ACM.

Nils Axelsson and Gabriel Skantze. 2020. Using knowledge graphs and behaviour trees for feedback-aware presentation agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, Virtual Event Scotland UK. ACM.

Hubert Baniecki, Dariusz Parzych, and Przemyslaw Biecek. 2023. The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery*.

Kai Biermeier. 2023. ELAN Annotation Frequency and Coverage: Commented Release.

Peter Carey. 2018. *Data protection: a practical guide to UK and EU law*, fifth edition edition. Oxford University Press, Oxford, United Kingdom.

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. 1996. Common Ground. In *Using Language*, pages 92–121. Camebridge Usiversity Press, Cambridge UK.

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

Christina Dideriksen, Riccardo Fusaroli, Kristian Tylén, Mark Dingemanse, and Morten H. Christiansen. 2019. Contextualizing Conversational Strategies: Backchannel, Repair and Linguistic Alignment in Spontaneous and Task-Oriented Conversations.

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.

Jennifer D. Ewald. 2010. "Do you know where X is?": Direction-giving and male/female direction-givers. *Journal of Pragmatics*, 42(9):2549–2561.

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*, 36(3-4):317–326.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *The new psychology of language*, pages 217–248.

Mayank Kejriwal, Craig A. Knoblock, and Pedro Szekely. 2021. *Knowledge Graphs: Fundamentals, Techniques, and Applications*. MIT Press.

Josef Klein. 2009. Erlären-Was, Erklären-Wie, Erklären-Warum: Typologie und Komplexität zentralier Akte der Welterschließung. In R. Vogt, editor, *Erklären: Gesprächsanalytische und fachdidaktische Perspektiven*, pages 25–36. Stauffenburg Verlag, Tübingen.

H. Kotthoff. 2009. Erklärende Aktivitätstypen in Alltags- und Unterrichtskontexten. In J. Spreckels, editor, *Erklären im Kontext. Neue Perspektiven aus der Gesprächsund Unterrichtsforschung*, pages 120–146. Schneider, Germany.

Richard J. Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–74.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Yi Ma, Paul A. Crook, Ruhi Sarikaya, and Eric Fosler-Lussier. 2015. Knowledge Graph Inference for spoken dialog systems. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5346–5350, South Brisbane, Queensland, Australia. IEEE.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence.

Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences.

Marvin Minsky. 1968. *Semantic Information Processing*. MIT Press.

Naomi Miyake. 1986. Constructive Interaction and the Iterative Process of Understanding. *Cognitive Science*, 10(2):151–177.

George Psathas and Kozloff Martin. 1976. The Structure of Directions. *Semiotica*, 17(2).

Amelie Sophie 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*, ICAART '23. SCITEPRESS.

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 Hullermeier, 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.

Sonja Stange and Stefan Kopp. 2020. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, pages 619–627, Cambridge United Kingdom. ACM.

Sonja Stange and Stefan Kopp. 2021. Effects of Referring to Robot vs. User Needs in Self-Explanations of Undesirable Robot Behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, pages 271–275, Boulder CO USA. ACM.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

W3C. 2012. OWL 2 Web Ontology Language Document Overview (Second Edition). *OWL 2 Working Group*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

A Appendix

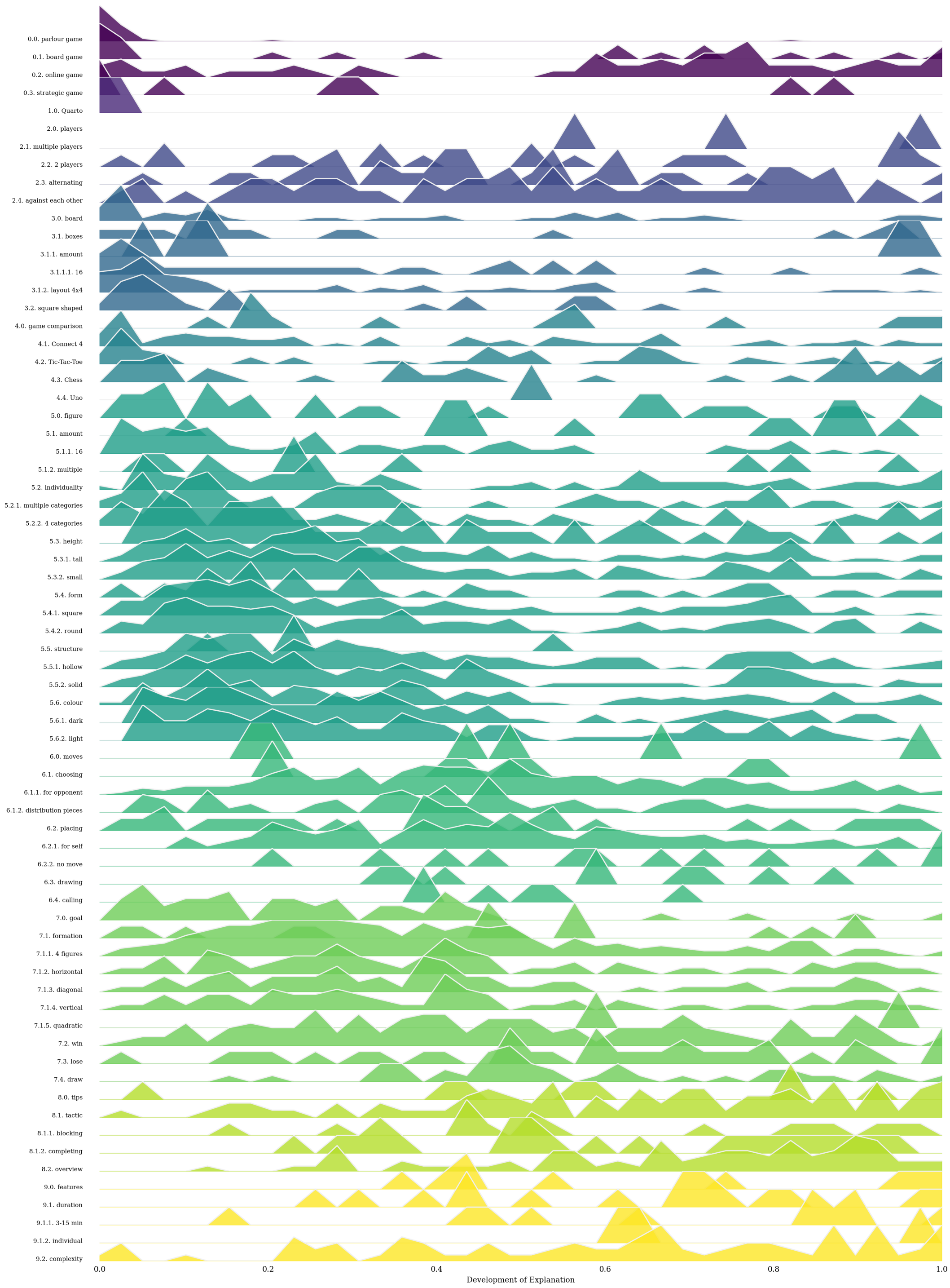


Figure 6: Reference to explanation nodes by EX in relation to the time over all dialogues

Predicting grounding state for adaptive explanation generation in analogical problem-solving

Lina Mavrina and **Stefan Kopp**
Social Cognitive Systems Group,
Bielefeld University, Bielefeld, Germany
{lvaronina,skopp}@techfak.uni-bielefeld.de

Abstract

This paper's main contribution is a Bayesian *hierarchical grounding state prediction model* implemented in an adaptive explainer agent assisting users with analogical problem-solving. This model lets the agent adapt dialogue moves regarding previously unmentioned domain entities that are similar to the ones already explained when they are instances of the same generalised schema in different domains. Learning such schemata facilitates knowledge transfer between domains and plays an important role in analogical reasoning. An explainer agent should be able to predict to what extent the explainee has learned to induce a schema in order to build up on this in the explanation process and make it more cooperative. This paper describes the approach of hierarchical grounding state prediction, introduces the analogy-based explanation generation process and the agent architecture implemented for this approach, as well as provides some example interactions as the first developers' evaluation of the system in preparation for upcoming empirical studies.

1 Introduction

Explanations are complex social processes that are actively shaped by both explainer and explainee throughout the course of their interaction (Miller, 2019, Rohlfing et al., 2021). Dynamic changes in the mental states of the explainee pertaining to their understanding of the *explanandum* (i.e., the object of the explanation) should be monitored and predicted by the explainer based on observable evidence, such as conversational feedback or clarification requests posed. These predictions should be then used to continuously re-conceptualise the *explanans* (i.e., the way in which the explanandum is presented by the explainer during explanation) (Rohlfing et al., 2021). Similar principles can be applied to human communication in general: active cooperation of the interlocutors and their stepwise co-construction of the interaction and the *common ground*, i.e., "their mutual, common, or joint knowledge, beliefs, and suppositions" (Clark, 1996, p. 93), as well as mentalising over relevant mental states of each other based on observable evidence (Kopp and Krämer, 2021). However, these principles of cooperative communicative behaviour are rarely applied in modern dialogue systems.

While explanations in a narrower sense serve as answers to *why?*-questions and explain causes of events, they can also serve other functions such as providing process narratives or instructions (Miller, 2019). In assistive scenarios, instruction and guidance during problem-solving are important functions of explanations. Analogy-based explanations specifically can help people transfer knowledge from one domain to another, for instance, via the process of schema induction. During this process, a generalised schema, i.e., "an abstract category that the individual analogs instantiate in different ways" (Gick and Holyoak, 1983, p. 8), can be induced from a range of specific examples and then applied to a new target domain.

This paper introduces an architecture for an assistive agent that guides the user through the process of problem-solving via adaptive explanations. The agent presents analogous stories from other domains hinting at the desired solution of the target problem, and helps the user understand similarities and differences between these stories, as well as induce and apply generalised schemata instantiated in the stories. In order to find good analogies, the agent uses graph-based knowledge representation to compare the examples and the target problem according to *structure-mapping theory* (Gentner, 1983). In order to be adaptive, the agent bases its explanation generation on predictions of grounding state of domain entities (DEs). These predictions are continuously updated via Bayesian inference.

The main contribution of the current research is the *hierarchical grounding state prediction model*.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

This approach allows the agent to adapt dialogue moves regarding previously unmentioned domain entities if they are related to the ones already explained via a common schema. The model and the architecture facilitating this kind of inference will be described in more detail in section 3, and their limitations will be discussed in a special section after the conclusion.

So far, the system has only been tested by the authors using different types of feedback and observing the behaviour of the agent. Some example dialogues showcasing the adaptivity of the system will be presented in section 4. The empirical evaluation of the system requires a series of laboratory studies in order to gain a comprehensive understanding of the impact of various factors present during adaptive spoken interaction. These studies are currently being planned and prepared for by the authors.

2 Background and related work

2.1 Adaptive explanation generation

With the rise of machine learning and specifically deep learning, the focus of research on explanations in human-machine interaction has been primarily on explanations of artificial systems and their decisions (Mueller et al., 2019). However, often these explanations are conceptualised and presented in a one-off and static way that may not be sufficient for diverse stakeholders interested in them (Suresh et al., 2021; Lakkaraju et al., 2022). An increasing amount of research is currently calling for incorporation of findings from social and cognitive sciences into explanation generation to make it interactive and adaptable towards specific goals, needs, expertise and changing levels of understanding of the explainee (Miller, 2019; El-Assady et al., 2019; Shvo et al., 2020; Sokol and Flach, 2020; Dazeley et al., 2021; Rohlfing et al., 2021; Lakkaraju et al., 2022).

The process of explanation generation can be divided into two parts: the *cognitive process* responsible for the generation of causes, and the *social process* responsible for construction and presentation of the explanans, as well as interpretation of the signals of explainee's understanding (Dazeley et al., 2021). The social process can also be seen as an interaction pattern consisting of joint actions that are facilitated by the processes of co-construction and scaffolding, during which the explainer should strive to build explanations from the knowledge the explainee already possesses, yet enrich it with additional relevant information (Rohlfing et al., 2021). This work focuses on the social process of explanation generation that can be studied and applied across a multitude of domains, not just in the field of explainable artificial intelligence.

An explainer agent incorporating the complexity of the explanation generation process requires (1) a rich and dynamic explainee model, describing relevant mental states and the level of understanding of the explanandum with appropriate granularity, (2) representations of domain knowledge, dialogue state and history, as well as (3) capabilities to continuously reason over these representations to select explanation strategies, dialogue moves and content under uncertainty inherent to communication. Concepts such as *Theory of Mind*, i.e., the ability to attribute mental states such as beliefs, goals and intentions to self and others (Premack and Woodruff, 1978), *mentalising*, i.e., the ability to predict the actions of others based on their desires, knowledge and beliefs (Frith and Frith, 2006), and *common ground* play an important role here (Miller, 2019; Shvo et al., 2020; Kopp and Krämer, 2021; Rohlfing et al., 2021). A major challenge for this research is the lack of high-quality training data for explanation dialogues, which means that the parameters of the models are hard to pre-train in advance and the system has to be able to adapt online relying only on the data observed during interaction.

Hereby, approaches used in older expert and tutoring systems can be revisited and adapted. One example is the EDGE explanation system described in Cawsey (1993). Here, inference rules are used to update the level of knowledge of the explainee stored in the user model. There are direct inference rules that concern entities under discussion and indirect inference rules that concern unmentioned entities. The former are based on the user input and update the user model, while the latter are conditions that are checked against the user model if the system requires the corresponding information to construct an explanation. The system presented in this paper similarly aims to infer the grounding state of unmentioned entities, but realises it with a hierarchical probabilistic model.

Speaking of implemented systems adapting the social process of explanation generation, here are some more recent examples. Robrecht and Kopp's (2023) SNAPE model uses online planning in form of Monte Carlo Tree Search to solve a non-stationary Markov Decision Process for explanation generation, where transition probabilities depend on the level of understanding for concepts under discussion as observed by the system from user feedback. Axelsson and Skantze (2023) work on adaptive presentation. Their agent adapts its generation behaviour based on the grounding levels of various concepts as inferred from observed multimodal user feedback and stored in a knowledge graph.

2.2 Models of common ground in dialogue systems

As previously mentioned, the concept of common ground is important for adaptive explanation generation, as well as adaptive dialogue in general. Empirical evidence suggests that representations of common ground in humans are richer than a mere binary of grounded vs. ungrounded, however, these representations are still required to be efficient to support real-time language use (Brown-Schmidt, 2012). Stone and Lascarides (2010) distinguish between two types of grounding models: symbolic approaches based on discourse coherence and probabilistic approaches based on inference from observed evidence. Both of these approaches have been used in earlier-generations dialogue systems, a prominent example of the former is Traum and Larsson (2003), while the latter was pioneered by Paek and Horvitz (2000). Stone and Lascarides (2010), however, point out that both of these approaches have limitations. For instance, the probabilistic approaches were primarily used to predict whether the system had understood the user during slot-filling, i.e., collecting of the parameters of the user's query. Yet for cooperative dialogue, predicting whether the user had understood the system is equally important. Thus Stone and Lascarides (2010) integrate both types of approaches in a theoretical framework consisting of a dynamic Bayesian network (DBN) model of dialogue that represents the relationships between interlocutors' mental states, evolving dialogue context, discourse moves and observable evidence produced by interlocutors over time.

Buschmeier and Kopp (2018), too, use a DBN to represent the dependency of the probabilistic grounding state on the so-called *attributed listener state* (ALS) over time. The ALS consists of several variables based on communicative functions of linguistic feedback (Allwood et al., 1992; Kopp et al., 2008), namely contact, perception, understanding, acceptance and agreement which are inferred within the DBN based on incoming multimodal data and interaction context.

Axelsson and Skantze's (2023) adaptive presenter agent stores grounding as labels of properties in the domain knowledge graph, and these labels are updated based on the user feedback category obtained from a random forest classifier (positive, negative or neutral feedback). SNAPE (Robrecht and Kopp, 2023) similarly represents the grounding state via level of understanding (a concept can be either grounded or not) regarding relationships in a knowledge-graph-based domain model. Di Maro et al. (2021) focus on detecting conflicts during interaction leading to inconsistent state of common ground. They conceptualise their common ground representation in terms of personal common ground consisting of dialogue history, and communal common ground consisting of domain knowledge that is shared between the agent and the user. On the technical level, their common ground representation is implemented as a graph database. A similar approach is also pursued in this work.

2.3 Analogical problem-solving

The general principle of analogical reasoning lies in the concept of *mapping*, wherein correspondences are found between the *source* (also called *base*, i.e., known body of information) and the *target* (problem to be solved) of the analogy (Gick and Holyoak, 1983). Gentner (1983) defines the so-called *structure-mapping theory* describing interpretation rules for analogies. This theory postulates that an analogy is characterised by the mapping of structural relations between entities within base and target, rather than the surface-level similarity of their features, and that this mapping is governed by the principle of *systematicity*, i.e., the existence of related higher-order relations. The key concepts of the structure-mapping theory are supported by empirical evidence (Gentner and Maravilla, 2017).

As mentioned before, analogical reasoning is closely related to the process of schema induction, during which a generalised schema is extracted from specific examples. Gick and Holyoak (1983) found that, when given two analogy sources, the participants were able to derive the generalised problem schema as a byproduct of comparison of the sources, and that the quality of the generated schema was a positive predictor for the transfer of the analogy to the target. Similar results were obtained by Gentner et al. (2003) who additionally showed that increasing the degree of guidance during analogy training increased the rate of transfer during the exercise. These findings suggest that an adaptive explainer/tutoring agent may have a positive effect on the success of analogical transfer in problem-solving.

To be able to interpret this work in the bigger context of research on analogical reasoning, a set of frequently used problems from the experiments by Gick and Holyoak (1983) was chosen as the use case for the agent. The explainee is required to solve the *Radiation* problem first posed by Duncker (1945) with the help of various analogs from different domains. In the *Radiation* problem the user is asked to imagine they are a doctor and have to find appropriate treatment for a patient with an inoperable tumor. The tumor can be destroyed with high-intensity radiation, but such procedure would also destroy the healthy tissue the radiation would pass through on the way to the tumor. While there are several possible solutions to the *Radiation* problem, the desired one is the so-called *convergence solution* where multiple weaker forces converge on the target, such as several low-intensity radiation rays from different directions that will not damage the healthy tissue, but combined will destroy the tumor. Further information on the use case will be provided in section 4.

3 Agent architecture

The core components of the architecture facilitating predictive grounding state inference are depicted in figure 1:

1. the *dialogue manager* based on the `flexdiam` architecture described in Yaghoubzadeh and Kopp (2017), extended for grounding state prediction and explanation generation, and

2. the *memory* component in form of a graph database that stores multiple types of information, such as *domain model*, *conversational record* (i.e., interaction-related information that was made public to interlocutors) (Thomason, 2003) and *dialogue information state* (DIS) incorporating the agent's prediction about current grounding state of domain entities (Buschmeier and Kopp, 2012).

In this section, these will be described in more detail. Additionally, a subsection will be devoted to the natural language understanding (NLU) component of the architecture to discuss an example use of state-of-the-art large language models (LLMs) in adaptive dialogue interaction.

3.1 Memory component

The memory component stores all information that is available to the agent at runtime in the form of a graph defining relationships between various types of entities (figure 2). Currently, these include the following.

- DE nodes: structured representation of domain knowledge is important for the application of the *structure-mapping theory* in order to determine the best analogy for the target among the sources. This representation includes abstractions of relations and actions in the form of generalised schemata, as well as instances of these schemata in source and target examples. The model can support a higher granularity of domain knowledge representation if necessary. DE nodes are initialised at the start of the interaction and do not change throughout.

- DIS nodes for domain entities: DE nodes for schemata and schema instances have corresponding DIS nodes that store the parameters of the probability distribution describing the current belief of the agent about the grounding state G of an entity. These parameters are initialised when the entity first becomes significant, for instance, by being introduced by the system, and updated whenever relevant evidence of understanding is provided by the user.

- Conversational record nodes: these store information about employed dialogue moves and user feedback concerning a specific DIS node. New nodes in this category are continuously created throughout the interaction, but once added to the graph, they remain unchanged.

The memory component is implemented using the graph database framework Neo4j[1].

[1] `https://neo4j.com/`

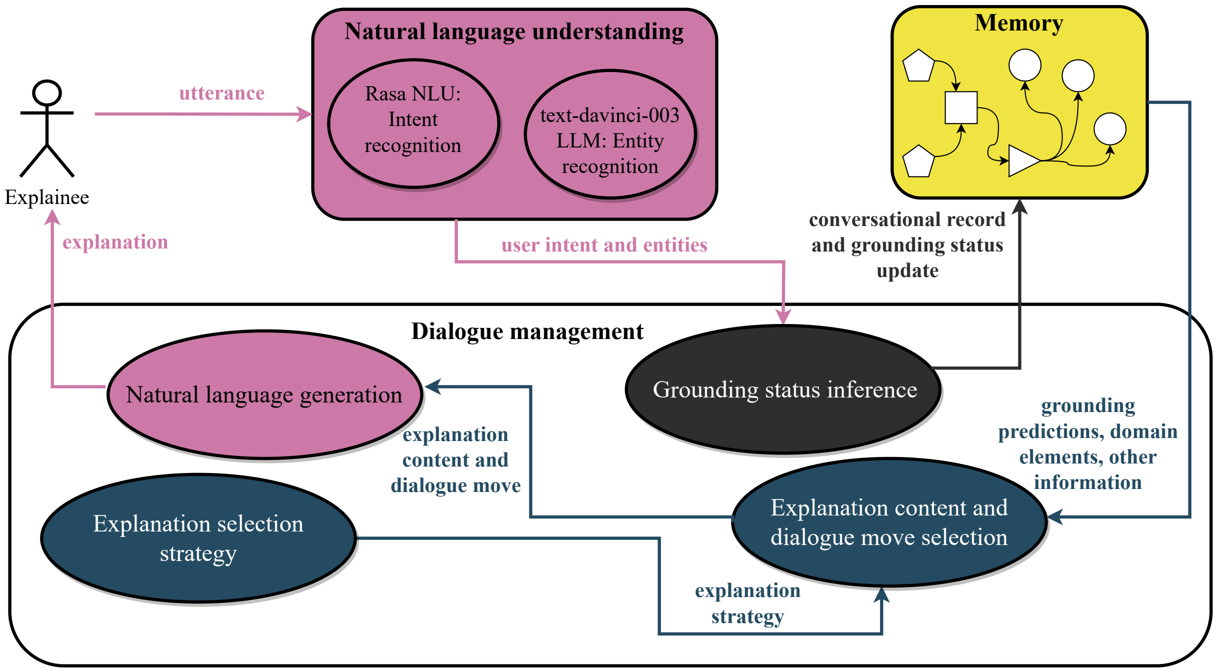


Figure 1: The architecture of the explainer agent.

3.2 Dialogue manager

As can be seen in figure 1, dialogue management essentially consists of two subsystems: *grounding state inference* and *explanation planning*. The dialogue manager is implemented in Python using the architecture called `flexdiam` (Yaghoubzadeh and Kopp, 2017) that was developed for spoken interaction in assistive settings utilising approaches tailored to dialogues with high degrees of uncertainty, which is also beneficial for a tutoring scenario.

Predictive grounding state inference

As previously mentioned, the belief of the agent about the grounding state G of a DE is described by a probability distribution. The parameters of this distribution are initialised when the entity becomes relevant for the first time during the explanation process. This initial distribution constitutes a uniform prior over the grounding state belief $P(G)$. When evidence of understanding U relevant to the entity is observed by the agent, it is used to calculate the posterior distribution $P(G|U)$ based on the Bayes' theorem:

$$P(G|U) \propto P(G) \times P(U|G) \tag{1}$$

Once the posterior is computed, it becomes the new prior distribution for the grounding state belief. In order to make the calculation of the posterior tractable at interaction time, the system uses conjugate priors for corresponding evidence likelihoods (Lambert, 2018). As the model for grounding state inference is hierarchical, two pairs of likelihoods and conjugate priors are used in the system, depending on the type of DE they are assigned to.

The lower level of the inference model deals with beliefs about the grounding state of schema instances. A belief about the grounding state of a schema instance is thus described by the beta distribution with probability density function (PDF) defined as

$$f(g;\alpha,\beta) = \frac{g^{\alpha-1}(1-g)^{\beta-1}}{B(\alpha,\beta)} \tag{2}$$

where $g \in [0;1]$ is the realisation of the random variable G representing the grounding state of a DE, $\alpha, \beta > 0$ are the shape parameters of the distribution, and $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function acting as the normalisation constant (where Γ is the gamma function defined for positive integers as $\Gamma(y) = (y-1)!$).

When the explainee reacts with positive or negative feedback to the agent's utterance, this feedback is interpreted by the system as evidence of understanding or non-understanding, respectively. This binary outcome is modelled using Bernoulli likelihood to which the beta distribution is the conjugate prior. Thus, the posterior is also a beta distribution with updated parameters

$$\alpha' = \alpha + \sum_{i=1}^{n} u_i \quad \text{and} \quad \beta' = \beta + n - \sum_{i=1}^{n} u_i \tag{3}$$

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

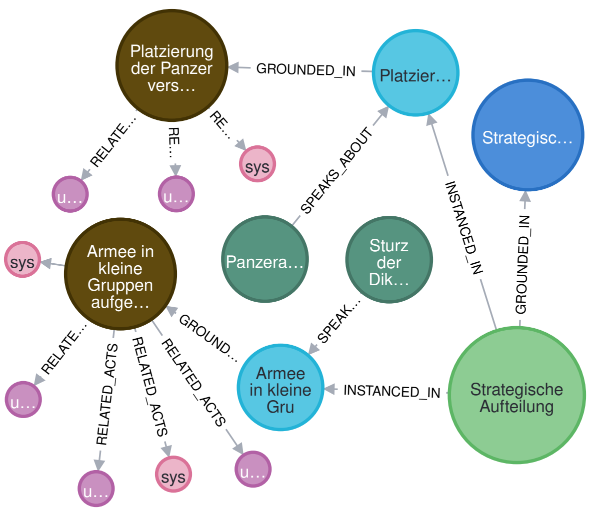


Figure 2: Part of the graph associated with the schema "strategic division" ("Strategische Aufteilung", big light green node). Following relationships and node types are represented here: `:GROUNDED_IN` as relationship between a DE and a DIS node, `:INSTANCED_IN` as relationship between schema and its instances, `:SPEAKS_ABOUT` as relationship between example and its schema instances, `:RELATED_ACTS` as relationship between a DIS and a conversational record node.

where $u \in \{0;1\}$ is the evidence of non-understanding ($u = 0$) or understanding ($u = 1$). As maximum of one instance of evidence per DE can be observed each turn, $n = 1$.

The higher level of the inference model deals with beliefs about the grounding state of generalised schemata. A new posterior for grounding state belief distribution of a schema is calculated if the distribution parameters of at least one of its instances were updated. The general update rule defined by the Bayes' theorem (equation 1) is applied here as follows. The mean value μ of the newly calculated posterior distribution $P(G|U)$ for the related schema instance is assigned to categories "low", "medium" and "high". These categories are defined in an overlapping fashion to express uncertainty within the model, for instance, μ that equals 0.45 is categorised as both "low" and "medium". The evidence of understanding is then defined by a categorical variable $\boldsymbol{u} = (u_{low}, u_{medium}, u_{high})$ where u is the number of occurrences of each category. So, for μ equals 0.45, the evidence of understanding used on the higher level of inference is $\boldsymbol{u} = (1, 1, 0)$.

The conjugate prior to the categorical likelihood is the Dirichlet distribution with PDF defined as

$$f(g_1, ..., g_K; \alpha_1, ..., \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} g_i^{\alpha_i - 1} \quad (4)$$

where $g_i \in [0;1]$ for all $i \in \{1;K\}$ and $\sum_{i=1}^{K} g_i = 1$ is the realisation of the random variable G representing the grounding state of a DE, $\boldsymbol{\alpha} > 0$ is the vector of concentration parameters of the distribution and $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$ is the multinomial beta function where the gamma function is expressed for positive integers in the same way as above. In the inference model, $K = 3$ for the categories "low", "medium" and "high".

Considering the definitions above, the parameter update rule for the Dirichlet distribution is

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{u} \quad (5)$$

A special case of feedback regarding a schema instance can occur if the agent poses an open question to the user in order to encourage them to apply a schema with a high grounding state belief to a new example by themselves, similarly to Cawsey (1993). This is a way to obtain high-quality evidence of understanding. If the user manages to successfully generate the schema instance, the update rules for the lower level of the inference model defined in equations 3 are superseded in order to distinguish such maximising feedback from regular positive feedback such as responding with "yes" to an agent's utterance. In this case, the parameters of the distribution are directly adjusted so that the mean of the distribution lies exclusively within the "high" category. A special label is added to the corresponding DIS node in the memory graph as well, denoting that its DE was generated by the user. The update of the higher level of the inference model then proceeds normally with $\boldsymbol{u} = (-1, -1, 1)$ to increase the impact of the evidence of understanding resulting from a user-generated utterance.

Explanation planning

The planning of explanations in the architecture is also hierarchical. On the higher level of abstraction, the agent can implement different general strategies that define the principles for explanation content and dialogue move selection, while on the lower level of abstraction, it selects new content and dialogue moves for every explanation turn based on predictions of the grounding state of DEs and rules defined by the high-level strategies. High-level planning can thus be seen as an instance of the *cognitive process* of explanation generation as defined by Dazeley et al. (2021), while low-level planning belongs more to the *social process* of explanation generation, and was therefore the primary focus of research so far.

Currently, high-level explanation planning is kept constant by predefined rules. For instance, the agent always starts with examples that are most similar to the target in terms of the *structure-mapping theory*. In future research, however, it can be attempted to formalise high-level strategies as adaptable *pathways*, building up on the definition of El-Assady et al. (2019), and explore the impact of this level of adaptation in empirical studies.

Concerning low-level planning, first, the main content of the next explanation turn is determined according to principles predefined by the high-level strategy. When the agent needs a new example, it is selected based on its structural similarity to the target. It is calculated using the Jaccard similarity coefficient:

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|} \tag{6}$$

where S and T are sets of analogy-relevant relationships within the source and target example, respectively. For instance, all relationships of the type `:SPEAKS_ABOUT` (figure 2).

Schema instances within an example are selected based on the high-level strategy. The memory component is hereby queried for corresponding grounding state predictions to inform the system's dialogue move selection via predictive inference.

Consider the general update principle of the grounding state belief of a DE in equation 1. This equation can be used to estimate the posterior distribution $P(G_m|U_m)$ given the most likely evidence of understanding $u{*}_m$ the agent would receive after a dialogue move m. The dialogue move resulting in the highest posterior distribution is selected by the explanation planner. The system currently supports two dialogue moves relating to introduction of new schema instances: "elicit generation" and "present alignment". Section 4 shows how the system chooses between these alternatives using predictive grounding state.

Determining $u{*}_m$ is not trivial and ideally requires a model of explainee's feedback generation. Right now, this value is defined by a set of rules for each available dialogue move. It is decided based on the category with the highest expected value in the grounding state belief distribution of the schema corresponding to the instance selected for the explanation turn. However, data of interactions with real users that will be collected in future empirical studies could be used to construct a generative model of evidence of understanding that can be used to estimate $u{*}_m$.

3.3 Natural language understanding

Previously, the `flexdiam` dialogue management architecture used the Rasa NLU[2] framework for intent and entity recognition. The language model employed there is based on word vectors that worked well for use cases with more structured user input where entity recognition was used primarily for slot-filling. However, in order to allow the users to answer open questions freely and use diverse expressions to refer to complex concepts and schemata, a different type of NLU component was required. This component should be capable of reformulating and summarising user utterances to obtain DEs that can be easily matched to the definitions in the agent's domain model. This kind of task is highly suitable for a pre-trained large language model (Yang et al., 2023), especially in the absence of high-quality training data.

These requirements led to a hybrid approach for NLU where intent recognition is still done with the Rasa NLU framework for a higher degree of control, while entity recognition is done with a pre-trained large language model based on the transformer architecture (Vaswani et al., 2017), namely, `text-davinci-003` from the GPT 3.5 family. Once the intent has been recognised by Rasa NLU, a prompt corresponding to the required entity recognition task is constructed. Currently, the pre-trained model is used "as-is", taking advantage of the LLMs' capabilities for few-shot learning from a small amount of handcrafted examples (Brown et al., 2020). However, the authors are preparing to evaluate the use of a smaller open-source model instead of `text-davinci-003` and are currently creating a data set for model fine-tuning.

While using an LLM can lead to unpredictable output such as hallucinations (i.e., undesirable text generation) (Ji et al., 2023), these risks were deemed acceptable, as the adaptive nature of the agent is expected to mitigate potential downstream errors caused by undesired language model out-

[2] https://rasa.com/docs/rasa/nlu-only

put through interaction, serving a function similar to repair of miscommunication in human-human interaction (Albert and de Ruiter, 2018).

4 Worked examples

This section offers more details about the use case for the agent, as well as some dialogue excerpts showcasing its behaviour in response to different types of user feedback. These are real conversations a user can have with the agent as it is implemented at the moment. Natural language generation is currently done with templates that were pre-generated using the `text-davinci-003` language model and manually edited. The possibility of using an LLM for online natural language generation is currently being evaluated. The agent converses with the user in German, however, for illustration purposes, the dialogues were translated into English by the author of the paper.

As mentioned in section 2.3, the use case chosen for the agent is based on experiments by Gick and Holyoak (1983). The user is required to find the *convergence solution* to the *Radiation* problem. The desired solution can be learned from analogous examples from other domains adapted from Gick and Holyoak (1983, Appendix II). These examples include various stories such as *Fall of the Dictatorship* (originally *The General*), in which the attacking army needed to be divided into smaller groups that converged on the fortress for a coordinated attack in order to avoid triggering the mines on the roads to the fortress.

Example interactions in the following dialogue boxes illustrate the adaptivity of the agent regarding presentation of instances of a known schema in previously unseen stories. The evidence of understanding for the schema instance in the story under discussion is used to update the probability distributions for the grounding state of the instance, as well as that of the underlying schema. These new probabilities are later used by the system to adapt dialogue moves happening potentially several turns after the evidence was provided by the user. In the example dialogues, the adaptation happens in the last turn of the agent when a new story containing a known schema is introduced to the user. For reasons of simplicity, the exchanges in the dialogue boxes happen in the system where the prior probability distributions for grounding state have just been initialised and were not changed by previous turns. For update rules and descriptions of nodes from the memory graph, see sections 3.2 and 3.1, respectively.

Without user-generated schema instance

AGENT: Can you name an important point from the story "Fall of the Dictatorship" that describes the problem in the story or its solution?

USER: No.

[*(1) Update $P(G|U)$ for all DIS nodes for schemata instances in "Fall of the Dictatorship" with $u = 0$, leading to μ' in category* ***"low"***
(2) Update $P(G|U)$ for all DIS nodes for schemata in "Fall of the Dictatorship" with $\boldsymbol{u} = (1, 0, 0)$]

AGENT: No problem, I will explain the important points of the story to you in a moment!

AGENT: One important aspect of the story "Fall of the Dictatorship" was that the big army of the general had to be divided in smaller groups, so that the mines on the road don't get triggered. Thus we see that the principle of "strategic division" was important here for the successful resolution of the problem.

USER: Okay.

[*(1) Update $P(G|U)$ for the DIS node for schema instance with schema name "strategic division" and story name "Fall of the Dictatorship" with $u = 1$ leading to μ' in categories* ***"low"*** *and* ***"medium"***
(2) Update $P(G|U)$ for the DIS node for schema called "strategic division" with $\boldsymbol{u} = (1, 1, 0)$]

[*(1) After several unrelated turns, when the next example containing "strategic division" gets presented, the DIS node for this schema has the highest expected value for $P(G|U)$ in category* ***"low"***
(2) Apply rules to determine $u{*}_m$ for category ***"low"***: *when m is "elicit generation", $u{*}_m$ is expected to be negative, predicting* ***"low"*** *posterior; when m is "present alignment" $u{*}_m$ is expected to be positive, predicting* ***<u>"medium"</u>*** *posterior*]

AGENT: A relevant feature of the story "Tank Assault" was that the available tanks were divided and positioned on different bridges, so that they could cross these narrow bridges. It means that the principle of "strategic division" is important here. Similarly, in the story "Fall of the Dictatorship" the troops were divided into smaller groups, so that mines on the road don't get triggered.

When a story is introduced to the explainee, they are asked whether they can identify any "important points", i.e., schemata present in it. In response, the explainee can describe any concepts that in their opinion contributed to the solution of the problem in the story. The system then evaluates whether valid schemata instances were named and/or described and updates the probability distributions of related DEs. This results in different states of the system and different behaviour later on.

In the first example, the user does not identify any schemata. Using this negative feedback and the grounding state update rules presented previously, new posterior distributions $P(G|U)$ are calculated for the grounding state of all schemata present in the story under discussion. The system is then required to introduce a schema (e.g., *strategic division*) to the user by describing its instance in the current story. The user can give feedback to signal their level of understanding. "Okay" is interpreted as positive feedback by the system. This leads to updates of grounding state distributions on relevant memory graph nodes. Later in the dialogue, when a new story containing the schema *strategic division* is presented to the user, the system chooses the appropriate dialogue move by predicting the most likely evidence of understanding $u*_m$ for each of the available dialogue moves m considering the expected grounding state category of the schema and the uniform prior over the grounding state corresponding to the unseen schema instance. Here, the expected grounding state category is "low" based on previous interactions, so the system chooses the dialogue move *present alignment*, i.e., to explicitly tell the user about the instance of *strategic division* in the new story and compare it with the instance of this schema from the previous story.

In the second example, the user correctly identifies the schema *strategic division* in the current story, which is recognised by the system using the approach described in section 3.3. In this case, the grounding state distributions related to the generated schema and its instance in the story under discussion are updated using previously defined special rules for high-quality evidence of understanding. Later on, when a new story containing the schema *strategic division* is introduced, the system, as described above, chooses the appropriate dialogue move by predicting $u*_m$. Now, however, the expected category of the schema grounding state is "high", and so the system predicts that the dialogue move *elicit generation* will result in successful generation of the schema instance in the new story by the user, giving the system the best possible evidence of understanding. Given the uniform prior over the grounding state of the new schema instance, this would result in the best expected value of the posterior distribution, therefore, the move *elicit generation* is chosen.

With user-generated schema instance

AGENT: Can you name an important point from the story "Fall of the Dictatorship" that describes the problem in the story or its solution?

USER: Yes, it was an important point in the story that the army was divided into smaller groups so that the mines on the road don't get blown up.

[*(1) Recognise that user is speaking about the schema called "strategic division"*
*(2) Update $P(G|U)$ for the DIS node for schema instance with schema name "strategic division" and story name "Fall of the Dictatorship", so that μ' is in category **"high"***
(3) Update $P(G|U)$ for the DIS node for schema called "strategic division" with $\boldsymbol{u} = (-1, -1, 1)$]

AGENT: You have correctly identified one of the important principles of this story. It is called "strategic division". Now I will help you identify other concepts that played an important role in the solution.

[*(1) After several unrelated turns, when the next example containing "strategic division" gets presented, the DIS node for this schema has the highest expected value for $P(G|U)$ category **"high"***
*(2) Apply rules to determine $u*_m$ for category **"high"**: when m is "elicit generation", $u*_m$ is expected to be maximising, predicting **"high"** posterior; when m is "present alignment" $u*_m$ is expected to be positive, predicting **"medium"** posterior*]

AGENT: Can you recognise the concept of "strategic division" in the story "Tank Assault"? If so, then please tell me how it is expressed here.

5 Conclusion and future work

This paper presents a *hierarchical grounding state prediction model* that allows an explainer agent to adapt dialogue moves regarding previously unmentioned domain elements.

This is necessary in scenarios such as instructions in analogical problem-solving, as through the principle of schema induction users can learn generalised schemata and apply them to new domains autonomously. Following the principle of *scaffolding* (Rohlfing et al., 2021), the agent has to be able to predict the grounding state of relevant domain entities in order to build up on the available knowledge in the explanation process and make it more engaging and cooperative. While these are the expectations placed on the agent, the system can only be comprehensively evaluated in a series of empirical studies. Preparing for these is the next step in the project. Interaction data with real users needs to be collected in order to construct a generative model for evidence of understanding to move away from the rule-based approach currently employed for posterior prediction of grounding state after a specific dialogue move. Additionally, it would be interesting to expand the research on high-level strategies for explanation planning and investigate whether and how those could/should be adapted.

Limitations

Inference of common ground in humans incorporates complex cognitive processes the exact combination of which is not fully understood. A computational model dealing with these processes naturally features a lot of limitations. A system that strives to be co-constructive in conversations with humans also needs to be efficient and interpretable. For reasons of efficiency, the proposed grounding state prediction model uses conjugate priors for Bayesian inference. However, conjugate priors often do not capture the full complexity of real-life data and events. Only results of an empirical study can show whether they are sufficiently good for the intended application.

Another limitation relates to the interpretation of feedback. While feedback fulfils a variety of communicative functions in interaction, such as signalling contact or perception (Allwood et al., 1992), the proposed agent interprets it as evidence of understanding. However, the system should react differently to a user signalling negative perception than to a user signalling negative understanding.

Additionally, the interpretability requirement makes the use of "black box" machine learning models inside the system problematic. Tools such as LLMs are powerful and can allow the system to engage in more complex dialogues, where tasks such as text summarisation and paraphrasing are required on behalf of the system. However, the risk posed by hallucinations of the language model cannot be eliminated completely. Even though the adaptive nature of the developed agent should mitigate it, potential downstream errors might still have a negative effect on understanding, overall success in the problem-solving task or acceptance of the agent. Despite this, multiple applications of LLMs should be considered in more detail, for example, automatic generation of domain knowledge representations, including abstractions of generalised schemata and analogous concepts from text descriptions of problems, as well as generation of training data for classic NLU approaches.

Ethics Statement

This research does not have any particular ethical implications, apart from general considerations on the use of artificial agents in dialogue interactions with humans. The authors are aware of risks associated with the use of large language models, however, some of them such as generation of potentially harmful or misleading content are contained through downstream processing in the architecture, as the output of the language model is not directly present in the utterances of the agent.

This article does not contain any studies involving animal or human participants performed by any of the authors. The authors also declare no conflict of interest.

Acknowledgements

This work was mostly done during the project IMPACT[3] (*The Implications of Conversing with Intelligent Machines in Everyday Life for People's Beliefs about Algorithms, Their Communication Behaviour and Their Relationship-Building*) which was funded by the Volkswagen Foundation and ended in March 2023. It is now continued in the project SAIL[4] (*Sustainable Life-Cycle of Intelligent Socio-Technical Systems*) that is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant number NW21-059A.

References

Saul Albert and J. P. de Ruiter. 2018. Repair: The Interface Between Interaction and Cognition. *Topics in Cognitive Science*, 10(2):279–313.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26.

Agnes Axelsson and Gabriel Skantze. 2023. Do You Follow?: A Fully Automated System for Adaptive Robot Presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 102–111, Stockholm, Sweden.

[3] https://www.impact-projekt.de/
[4] https://www.sail.nrw/

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.

Sarah Brown-Schmidt. 2012. Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, 27(1):62–89.

Hendrik Buschmeier and Stefan Kopp. 2012. Using a Bayesian Model of the Listener to Unveil the Dialogue Information State. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*, pages 12–20, Paris, France.

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)*, pages 1213–1221, Stockholm, Sweden.

Alison Cawsey. 1993. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3:221–247.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, MA, USA.

Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. 2021. Cutting melted butter? Common Ground inconsistencies management in dialogue systems using graph databases. *Italian Journal of Computational Linguistics*, 7(1 | 2):157–190.

Karl Duncker. 1945. On problem-solving. *Psychological Monographs*, 58(5):1–113.

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. In *Proceedings of Human-Centered Machine Learning Perspectives Workshop at CHI'19*, Glasgow, Scotland, UK.

Chris D. Frith and Uta Frith. 2006. How we predict what other people are going to do. *Brain Research*, 1079(1):36–46.

Dedre Gentner. 1983. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7(2):155–170.

Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2):393–408.

Dedre Gentner and Francisco Maravilla. 2017. Analogical Reasoning. In L. J. Ball and V. A. Thompson, editors, *International Handbook of Thinking & Reasoning*, pages 186–203. Psychology Press, New York, NY.

Mary L. Gick and Keith J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology*, 15(1):1–38.

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).

Stefan Kopp, Jens Allwood, Karl Grammer, Elisabeth Ahlsen, and Thorsten Stocksmeier. 2008. Modeling Embodied Feedback with Virtual Humans. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, volume 4930 of *Lecture Notes in Computer Science (LNAI)*, pages 18–37. Springer Berlin Heidelberg.

Stefan Kopp and Nicole Krämer. 2021. Revisiting Human-Agent Communication: The Importance of Joint Co-construction and Understanding Mental States. *Frontiers in Psychology*, 12.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective.

Ben Lambert. 2018. *A Student's Guide to Bayesian Statistics*. SAGE Publications, Inc., London, United Kingdom.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review. Synopsis of Key Ideas and Publications and Bibliography for Explainable AI.

Tim Paek and Eric Horvitz. 2000. Conversation as Action Under Uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, pages 455–464, Stanford, CA, USA.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

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.

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.

Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith. 2020. Towards the Role of Theory of Mind in Explanation. In *Proceedings of EXTRAAMAS 2020: Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 12175 of *Lecture Notes in Computer Science (LNAI)*, pages 75–93, Online. Springer International Publishing.

Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Künstliche Intelligenz*, 34:235–250.

Matthew Stone and Alex Lascarides. 2010. Coherence and Rationality in Grounding. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2010)*, Poznań, Poland.

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan.

Richmond H Thomason. 2003. Accommodation, Meaning, and Implicature: Interdisciplinary Foundations for Pragmatics. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 325–363. MIT Press, Cambridge, MA, USA.

David Traum and Staffan Larsson. 2003. The Information State Approach to Dialogue Management. In J. van Kuppevelt and R. W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology (TLTB)*, pages 325–353. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Ramin Yaghoubzadeh and Stefan Kopp. 2017. Enabling robust and fluid spoken dialogue with cognitively impaired users. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Saarbrücken, Germany.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.

“Why Do You Say So?” Dialogical Classification Explanations in the Wild and Elicited Through Classification Games

Jana Götze and **David Schlangen**
Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
jana.goetze,david.schlangen@uni-potsdam.de

Abstract

Enabling classification models to deliver successful explanations requires such models to not just deliver an explanation on top of their classification, but to adapt to the explainee in presenting arguments and details that the explainee may ask about. We present data collection settings that aim at eliciting such dialogical classification explanations in the context of visual dialog where dialog players need to draw conclusions based on this image. We then describe data from a naturally occurring setting as well as two game settings and how the preliminary data we have collected can inform model building.

1 Introduction

Neural network methods have pushed the boundaries of automating classification tasks in many areas of research. For tasks that involve language and image data, such deep learning models are able to reach ever increasing accuracy. A common concern with any kind of automated classification method is its opacity in terms of how it derives its decision based on the training data it was fed. This concern does not just target the human desire to understand what is going on internally but also the necessity to detect and correct possible mistakes or undesired biases in the underlying data. Furthermore, a model's reasoning can inform us about patterns in the data or training setup that can either inform human decision-making or correct misconceptions about cause-effect relations, as case studies from deployed models show (Caruana et al., 2015).

The area of Explainable AI has become more active again in recent years, focusing on different types of methods that can make a model understandable in some way.[1] These methods commonly present a user with an explanation alongside their classification, without the user being able to further question this explanation. In Miller (2019)'s terms, these methods focus on the *cognitive process* of finding a decision rather than the *social process* of delivering the decision to a specific interaction partner.

In many everyday scenarios, human decisions are open to debate: a decision can be questioned or challenged, or a listener may want to ask questions about how the decision came about. For example, students want to understand a grading decision or patients a medical diagnosis. These explanations are user-dependent, i.e. the explainer takes the listener's previous knowledge or intent into account (Miller, 2019).

The structure of such explanatory dialog has been studied by Walton (2009) and has been empirically affirmed by Madumal et al. (2019), who have annotated naturally occurring data of different multi-party explanation settings.

In his survey of sociological research on explanation, Miller (2019) points out that current research on Explainable Artificial Intelligence systems misses some features that human explanation-giving is known to posses from various studies in sociology and neighboring disciplines. One aspect that we take up here is the aspect of interactivity. In conversation, explanations are not given in isolation, but are embedded into the context of a decision that is reached by two or more speakers in a collaborative process and that takes into account the explainee's knowledge about the issue. Research has shown that conclusions that are reached in such a collaborative fashion are more often correct than an average of individual conclusions (Karadzhov et al., 2022b). Agents that can explain their claims on human terms are therefore

[1] We acknowledge that there is no consensus about terms such as transparency, explainability, interpretability, justifiability and others. We use the term *explainable* in a loose way that incorporates aspects of any of the other commonly used terminology. We are interested in models that can express their reasoning about a decision in natural language, thus giving a human-understandable explanation.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

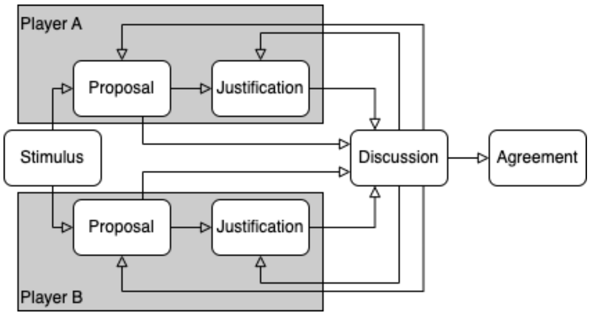


Figure 1: Schema of dialogical classification dialog.

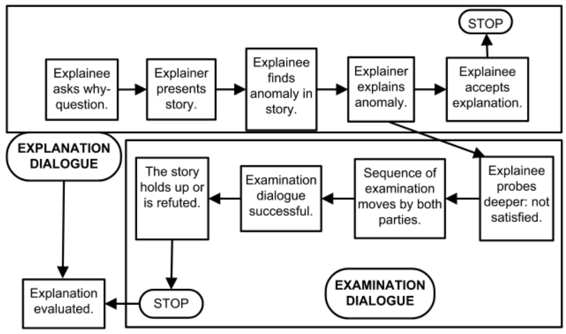


Figure 2: A model of argumentation and explanation in dialog proposed by Bex and Walton (2016, Fig. 4)

likely to help humans make better decisions than they would individually.

Even though some data exists from which such agents can be modeled, it consists largely of only textual data that deal with different topics of discussion, i.e. there is no visual dialog context. In our effort to build explanatory agents that can deal with multimodal input we introduce two new tasks: WORDLE with images, based on the popular Wordle word game, and a collaborative image classification task that comes in two variants: human-human (which we call JUJU for "judgment justification") and human-agent (which we call SIJU for "single-player justification"). We also present data from a naturally occurring dialogical classification setting in discussion forums (below called FORUM). The games we propose differ in their settings and can be used to test the ability to make associations between words and images, and the ability to give and evaluate reasons for proposals.

In the next section, we go into detail about how we intend to extend Miller (2019)'s and Walton (2009)'s models in the context of dialogical classification. We then describe the data collection settings (Section 3) that we use to elicit reasoning chains in interactions. We present example dialogs along with qualitative analyses in Sections 3.3 and 3.4. We end by reviewing related work (Section 5) and summarizing our conclusions with an outlook on future work (Section 6).

2 Collaborative Explanation

Rather than looking at explanations in a general sense, we want to constrain ourselves to explanations of a specific type: Explanations for a classification decision, i.e. given a label $l \in L$ and an input instance $i \in I$, where L is a set of labels and I is a set of images, we want to see dialog around the question "Why was l assigned to i?".

More specifically, we wish to see dialog that fulfills the following constraints:

- *Grounded justifications*: Justifications for a classification label must be grounded in the input data, i.e. when presented with a label, the explanation must refer to (features of) the particular instance and the domain of the classification.

- *Reasoning towards an agreement*: Rather than explaining a past decision, e.g., "Why did you do this?", we want to see how a decision agreement about a future event is reached, e.g., "What decision should we make and why?". These reasoning chains typically start with a proposal for a solution that is then explained for an explainee but the explainee can make a counter-proposal that can become the solution.

- *Symmetric roles*: We are looking for scenarios in which neither dialog participant knows the ground truth. We are particularly interested in seeing how humans reach a conclusion based on the input that they see and what it takes for one player to agree to the other and when counter-proposals are made.

Looking at Bex and Walton (2016)'s model of explanation dialog that we show in Figure 2, we can see how explanations are initiated by a why-question and how an examination subdialog happens when the explainee requires more insight into the explainer's reasoning. In this model, the dialog ends when either the explainee accepts the explainer's justifications or the explanations (the *story*) are refuted. Table 2 shows some examples of data that Madumal et al. (2019) have gathered to empirically verify the model. From these examples we can see that the topics of discussion in this data is rather abstract and cannot be grounded in the

Task/Data	FORUM	SIJU	JUJU	WORDLE
Sources/Collection method	Reddit, Whatbird	slurk, AMT, Prolific	slurk, AMT	slurk, Prolific
Images	Reddit, Whatbird	CUB	CUB	ImageNet, WikiCommons
Class descriptions	—	Whatbird	Whatbird	—
# dialogs	400	38	11	8

Table 1: Summary of collected data and sources. Appendix A lists details about the data sources.

Question	Source
Over time, did you go, "I need to think this through"?	Journalist Interview
To what extent are you concerned that we might see this problem emerging as a significant challenge [...]	
What does that suggest to you?	Chatbot transcript
What are your feelings now?	
How did you guys decide who would walk on the moon first.	Reddit
Why were you wearing this BYU shirt?	
What difference would that have made under the terms of the Hobbs Act?	Court transcript
Why isn't that enough?	

Table 2: Examples of questions that initiate an explanation in the data analyzed by Madumal et al. (2019). Answers to these questions cannot be grounded in the external dialog context.

#	A	B	C
1	 [A⟩B,C] New birder - need help to see if I'm able to get a positive ID on this Sparrow! Options I'm seeing are Swamp Sparrow, Lincoln Sparrow or Song Sparrow. Understand positive ID might not be possible since I don't have the belly in the photo.		
2		[B⟩A] That is a Song Sparrow.	
3			[C⟩A] I agree with Song Sparrow
4	[A⟩B,C] What's the giveaway for you - the pattern on the back?		
5		[B⟩A] Pattern and color.	
6		[B⟩A] Lincoln's Sparrows have buff mustachial stripes and cleaner back patterns.	
7		[B⟩A] Swamp Sparrows have more orangey brown wings and lack the thick dark malar stripes that Song Sparrows have.	
8	[A⟩B,C] Thanks so much for your help!		

Figure 3: A shortened transcript of a forum interaction.

dialog context and that the roles of explainer and explainee are fixed throughout the dialog, e.g., as interviewer and interviewee. The model and data thus gives insight into the structure of explanation in general and especially shows how the explainer must coordinate with the explainee to reach a common understanding. However the model is not fine-grained enough to capture the phenomenon that we wish to model in which the dialog participants both need to reach an agreement.

Figure 1 shows a schematic diagram of the dialog structure we aim to collect data for. The roles between explainer and explainee can change at any time as players propose a class label for the stimulus. Discussion can lead to players to retract their proposal, e.g., when the other player detects an inconsistency, or to decide on which proposal is better. The dialog ends when players reach an agreement.

3 Data collection settings

In this section, we describe four different settings of dialogical classification, and discuss examples. We start with showing examples of dialogical classification "in the wild": forum discussions in which users are looking to classify images of birds (Section 3.1). We discuss how this setting, while being close to our target, is unsuitable to model dialogical classification as an agreement game. The following three sections describe tasks, data, and methods from our own pilot data collections, for which participants were recruited via crowdsourcing platforms. In two of these settings (JUJU, SIJU, Section 3.3), players interact in the same domain of bird classification. In the fourth setting (WORDLE, Section 3.4), players need to take into account several pieces of information (including visual) in order to successfully win a game that they play together. Before describing the settings, we briefly describe the data collection method in Section 3.2.

3.1 Forum data

Dialogical classification happens as part of everyday life. For example, in the domain of bird classification, it is common for people wanting to classify a particular bird that they have spotted in their backyard or during a walk. In specialized forums like WhatBird and general forums like Reddit, anyone can post an image of a bird and ask others to help them classify the bird. One or more users can then propose a label, explain their choice, and discuss with other users to reach agreement.

Tasks and Data We have collected 130 forum threads from the platform Reddit[2] and 270 threads from the platform WhatBird[3] in which a user posted an image of a bird and asks for help with classifying the image. This data has the potential of showing dialogical classification explanations according to the schema in Figure 1 because a dialog can have a variable number of players in the explainer role. The user posting the request for help is a fixed explainee in this setting but can also ask questions and make proposals. Explanations are likely to be grounded in the specific image uploaded with the request.

For the purpose of this work, we only include forum threads that start with a user posting an image and making a clear classification request that is followed by at least one proposal for a label. Before doing any analysis, the data is cleaned of sensitive information such as usernames and urls. Usernames are translated into neutral identifiers.

Example forum data Figure 3 shows an example forum thread. The dialog starts with A requesting a label and explanation for an image and also giving 3 initial proposals. Two users B and C each provide the same label (turns 2 and 3, with C explicitly agreeing with B) but no explanation. A requests an explanation for the label and B explains the decision in turns 5–7. In this particular case, the explanations also use the principle of exclusion in that they refer to features not present in the image. The dialog ends with A accepting the proposal in turn 8.

Dialog structure Even though both forums are specifically targetted at birders, users in these forums may have different motivations for contributing their knowledge. Providing a label for an image is a courtesy and there is no consequence in being incorrect. In fact, for this data, like for most of the data analyzed by Madumal et al. (2019), no ground truth is available to us as for the correctness of the final answer.

In order to get insight into whether this data exhibits the structure of dialogical classification, we look for contributions in the threads in which a user rejects a proposed label or explanation and thus starts a detailed discussion about whether a label matches an instance. We find more than 500 such instances in the 400 dialogs and show examples in Table 3. In the examples, explainers go into detail about why they reject a label and seem to be using a similar strategy of describing concept boundaries that Myrendal (2019) has described for discussions around word meanings: The explanations contrast features of different classes with each other in reference to the stimulus image. These contrasts naturally contain many negations, e.g., by mentioning what an instance of another label would look like (cf. Table 3). While this strategy is valid and interesting to analyze in the future, we are looking for explanations that can be grounded more directly in the image or dialog context, i.e. for positive evidence rather than negative evidence. In the examples, explainers very often reference prototype birds that are not immediately available in the data. As such, these forum interactions are an interesting next step in which a classification model must also learn to reason about the class representation it has built. As a first step however, we work towards the settings where the explanations can be grounded in the immediate context. In addition, this forum data cannot be extended with additional threads on demand as can be done with the tasks we present in the following sections.

#	Example
1	Great Potoos are distinctly more white/pale-colored than this bird and they have black eyes.
2	Too large to be Cattle Egrets, and I would have discerned the yellow bills.
3	The white throat really is a great field mark for White-throated Sparrows; it's distinctive even if they have the drab/tan head stripes
4	Goldfinches have shorter more conical beaks compared to the Scarlet Tanager's long slightly curved beak.
5	Even at this angle, it has a slight recurve

Table 3: Examples of forum contributions that reject a proposal or explanation.

3.2 Collecting chat data

For the following three tasks (SIJU, JUJU, and WORDLE), we have used the slurk chat framework (Schlangen et al., 2018; Götze et al., 2022) to set up a chat environment in which players play the game via their internet browser. In each task, players see a chat interface, the game instructions and any game-specific visual material, such as images.

[2] https://www.reddit.com, subreddits whatbirdisthis, whatisthisbird, and whatsthisbird

[3] https://forums.whatbird.com/

Examples of the interfaces are shown in Figure 4.

An automated game bot that we call *GameMaster* helps the players navigate and informs them about invalid actions and the game score. In the case of the SIJU task, the *GameMaster* is also the player's dialog partner, asking questions to elicit more explanations.

We collect interactions via the platforms Amazon Mechanical Turk and Prolific. Table 1 shows our sources for materials and participants. All data was collected in English. All chat logs are stored on a local server at the authors' institution. Worker IDs are connected to chat ids via a token that players obtain after they finish playing. This allows us to track players who play the game repeatedly and potential repetitions of the same images or words.

We pay workers an average of about $13.00 per hour over all collected data.

3.3 SIJU and JUJU: Judgment Justification

As we have described in Section 3.1, the naturally existing forum data contains aspects of the dialogical classification structure that we aim at but still contains many explanations that cannot be grounded in the immediate dialog context. In this section, we describe two variants of a game in which we constrain the dialog context in a way that allows players to use contrastive explanations that can still be grounded in the context.

We create two variants of a game in which players are tasked to match an image with a description: In the setting that we call JUJU, two players are tasked to create a mapping between 3 images and 3 descriptions. In the setting that we call SIJU, we replace one of the dialog participants with a bot agent that takes the role of a critic. Collecting interaction data for synchronous settings poses an additional challenge in timing participants. With the SIJU setting, we want to investigate whether the illusion of a dialog can be created to the extent that a player will elaborate on their decision further.

JUJU Task Two players are shown 3 images of birds and 3 class descriptions and are tasked to create a mapping between the images and the descriptions. Figure 4 (top right) shows the visual dialog context that participants see. Both images and descriptions are labeled for easy reference (`A/B/C` and (`1/2/3`, respectively). In order to avoid the players making decisions by exclusion, the mapping need not include all the images or all the descriptions, there can be images without descriptions and descriptions without images and the players are informed about this. One round ends when the players both enter their joint decision in free text. The game ends after three rounds.

The *GameMaster*'s role is to keep track of time and contributions and keep the players informed about the state of the game. For example, the *GameMaster* will not accept decisions without prior discussion between the players.

SIJU Task In this variant, the player's task is to decide whether a description fits an image. The player only sees 1 image and 1 description, Figure 4 (top) shows the interface. The game first asks "Does the description fit the image?" and the player answers by clicking one of the buttons `yes`, `no` or `maybe`.

The *GameMaster* that has both the roles of game manager and dialog partner then asks for a justification of the decision. Players click `next` when they think they have explained enough. The game ends after three pairs.

The *GameMaster*'s role is to challenge the player in their justification by using simple checks of the player's input. The *GameMaster* asks for additional explanation when the explanation falls below a minimum length of 20 characters, fewer than 10% of the features mentioned in the description were taken up in the explanations, or when the explanation is a substring of the description, i.e. the player used copy/paste to answer.

For the images and descriptions, we use the USC-Birds Dataset (Welinder et al., 2010). Besides the bird images, the dataset contains symbolic attribute representations for each bird species. This information is used by the bot to determine which attributes have been mentioned and also allows us to create image-description pairs that have a substantial overlap in order to create pairs where the decision is sufficiently difficult to make.

Collaborative classification in SIJU and JUJU Figures 5 and 6 show examples of collaborative classification dialog in the SIJU dialogs. In Figure 5, the *GameMaster* is not satisfied with the initial explanation and prompts player A to explain further some of the attributes that have not been mentioned before. Player A goes into detail about how the image and description match – making turn #4 a good example of a grounded explanation. In Figure 6, the player goes into detail about the attributes that are mentioned, using positive and

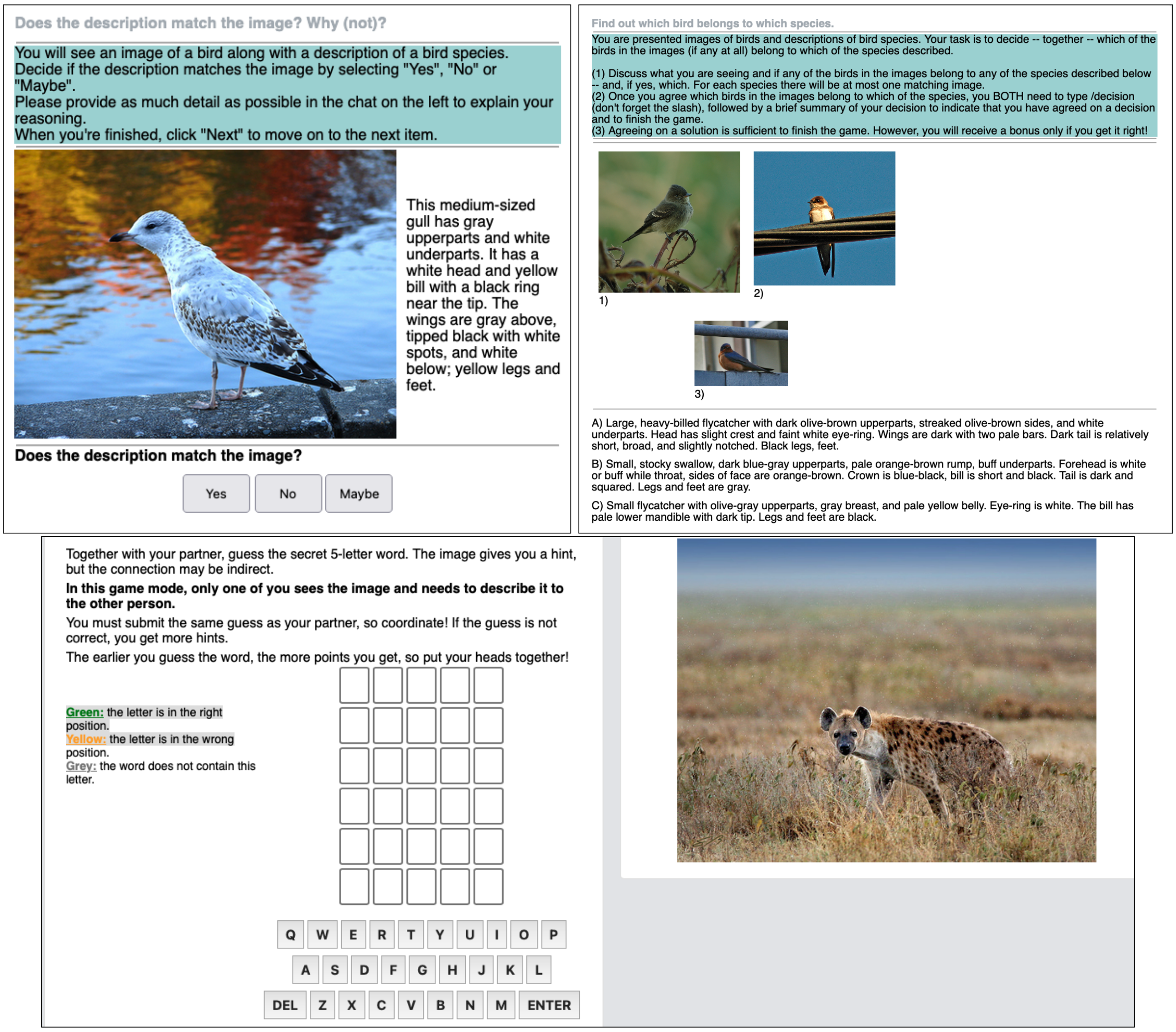


Figure 4: The visual interfaces for the SIJU (top left), JUJU (top right), and WORDLE (bottom) games. The interface also includes a chat area that is not shown here. The complete interfaces are shown in Appendix C.

negative references (*"The upperparts do not look reddish."*).

Figure 7 shows an extract of a JUJU dialog in which the players first propose a solution (turns #1 and #2: B: *"3c for sure"*—A: *"yeah I was thinking C could fit 2 or 3"*) and then take turns to explain it in more detail (turns #3 to #6). Here, too, do the players use specific details from the descriptions and ground them in the images, using both the existence and absence of features they see in their explanations (*"B says pale yellow belly and 2 and 3 have white bellies"*, *"1's belly is not pale yellow and it doesn't have an eye ring"*, *"The wings are dark with two white bars"*).

3.4 Wordle with Images

This 2-player game is an extension of the popular Wordle word game[4] in which participants need to find a 5-letter word. In addition to cues that specify whether single letters are in the correct position or not, our version also includes an image as a cue for the target word. In this setting, the decision that players make is not a label for the image, but a sequence of letters that needs to adhere to a number of constraints: The target word needs to be related to the image and it needs to fit the letter cues that the *GameMaster* provides as the game progresses. Participants need to agree on their next guess before entering it. The maximum number of guesses is six. The *GameMaster*'s role is to provide letter-based cues and guide players through the mechanics of the game, e.g., informing them when their partner has entered a guess. Figure 4 (bottom) shows a screenshot of the visual game interface.

For our pilot data collection, we have manually mapped images with target words. In order to elicit meaningful dialog, the mapping must be

[4]https://www.nytimes.com/games/wordle

Medium, active warbler with black upperparts and hood, distinctive orange-red patches on wings, sides, and long, fanned tail, and white underparts. Bill, legs, and feet are black.

#	A	B
1	[A⟩B] The description matches the image - all components have a corresponding part to the image. You could disagree on the disambiguation of Warbler but the colour, pattern, plummage and characteristics of the bird match the description.	
2	[A⟩B] ACTION: BUTTON-NEXT	
3		[B⟩A] I'm not convinced yet. What do you think about the other attributes, such as the side and the feet?
4	[A⟩B] The description states that the sides have distinctive orange-red patches. The image shows that the sides have distinctive orange-red patches. The description states that the feet are black. The image shows that the feet are black.	

Figure 5: An extract of a SiJu interaction. This example shows how the bot prompts the player to be more precise in their explanation. The player then adds more details that can be grounded in the image details.

sufficiently indirect (showing an image of a piano for the target word *piano* does not require the players to discuss more than their agreement). We also add difficulty by showing the image to only one of the players so that this player needs to describe details of the image to their partner.

Collaborative classification in wordle In the wordle task, the players now need to observe additional constraints: The target word must be related to the image in some way, it must have exactly 5 letters, and it must adhere to the letter-based feedback. Figure 8 shows an extract in which the players discuss one aspect of the constraints – the letter-based feedback. This subdialog is a good example of one player (A) verbalizing the visual feedback by making a proposal for what the next guess must look like (*"None of the other letters are in the word. Just C."*). The other player B then takes up the proposal by rejecting it because it has misinterpreted the feedback (*"there is an a just in a different position"*). A accepts the counter-proposal in turn #5 and makes a new, more specific proposal in turn #6.

Medium-sized wren with rufous upperparts and buff underparts. Eyebrows are white, wings and tail are dark barred with white flecks. Throat and chin are white. Bill is decurved. Legs and feet are pink-gray.

#	A	B
1	[A⟩B] Medium sized is plausible as this looks larger than the average wren. White-ish eyebrows are visible and throat and chin are arguably white. Bill is decurbed.	
2	[A⟩B] I don't know what rufous means, but I'm guessing reddish. The upperparts do not look reddish. Buff underparts are evidence. Dark wings are evident, but white flecks are not. Tail, legs and feet are not visible.	

Figure 6: An extract of a SiJu interaction.

4 Collaborative explanation models

In the previous section, we have shown data examples of different classification settings that all show aspects of the dialogical classification explanations we aim to capture. Specifically, we see how the dialog partners propose solutions and reason about details in images in descriptions in the SiJu and JuJu settings, using both the presence and absence of feature values in an image (*grounded justifications*). In the JuJu and wordle settings, we can see how players make proposals, explain them, and agree on them towards the game goal (*reasoning towards agreement*). Especially in the wordle setting, the larger number of constraints on the solutions seems sometimes to elicit incorrect proposals that the other player can counter (cf. the example in Figure 8). Also in the JuJu and wordle settings, players' roles are equal, meaning that they alternate between being explainer and explainee since neither has the game solution (*symmetric roles*).

For classification models to give dialogical explanations, what the data in the SiJu setting shows can be considered a minimum capability: On request, a model must be able to go into more detail with an initial explanation, mentioning additional features and possibly admitting that certain features cannot be determined from a given image. The modeling efforts of Li et al. (2018) and Park et al. (2018) are close to this capability, however they lack either the continued dialogical explanation or

1  2  3 

A	B	C
Large, crested flycatcher with olive-green upperparts. Head, throat, and upper breast are gray, belly is yellow, and undertail coverts are lemon-yellow. Bill is heavy and black. Wings are dark with rufous patches. Tail is rufous.	Small flycatcher with olive-gray upperparts, gray breast, and pale yellow belly. Eye-ring is white. The bill has pale lower mandible with dark tip. Legs and feet are black.	Medium-sized flycatcher with dull olive-gray upperparts and pale olive-gray underparts. Head has darker cap and slight crest. The wings are dark with two white bars.

#	A	Game Master	B
1			[B⟩A] 3c for sure
2	[A⟩B] yeah I was thinking C could fit 2 or 3		
3	[A⟩B] A and B don't seem to fit any of them		
4	[A⟩B] B says pale yellow belly and 2 and 3 have white bellies. 1's belly is not pale yellow and it doesn't have an eye ring so I don't think B fits any of them		
5			[B⟩A] slight crest. The wings are dark with two white bars. was what has me thinking 3c
6			[B⟩A] 1 and 2 have no crests on their heads that I see
7	[A⟩B] I agree		
8	[A⟩B] so for sure 3c		
9			[B⟩A] yeah, 1 and 2 are confusing

Figure 7: An extract of a JUJU interaction.

#	A	Game Master	B
		[GM⟩A,B] SUBMIT GUESS: cakes – FEEDBACK: C A K E S	
1	[A⟩B] OK. It's not cakes. But it starts with C		
2	[A⟩B] None of the other letters are in the word. Just C.		
3			[B⟩A] there is an a
4			[B⟩A] just in a different position
5	[A⟩B] Oh yeah, sorry.		
6	[A⟩B] Try Cotta		

Figure 8: An extract of a WORDLE interaction.

the questions that the model has to answer are more specific than a general "I need more explanation". The JUJU setting adds more advanced capabilities in which an explainee can make an own proposal that a model must be able to evaluate against its own beliefs. Finally, the WORDLE setting creates a solution space that is constrained by game rules (5 letters, a limited amount of steps) and dynamically changing context (letter-based feedback), as well as a visual input. This latter setting requires advanced reasoning skills and strategic game play in order to stay within the maximum number of allowed steps.

5 Related Work

A growing body of research investigates natural language explanations in the context of classification decisions. Wiegreffe and Marasovic (2021) have compiled an overview of datasets, 10 of which include free-text explanations in a variety of classification tasks that involve visual as well as textual input. Two of these datasets are extensions to the Visual QA task (Antol et al., 2015) in which a model must answer consecutive questions about an image (Li et al., 2018; Park et al., 2018), giving specific elaborations for an answer. Others include explanations for particular action decisions in a given context, e.g., for self-driving cars (Kim et al., 2018) or in a controlled game setting (Ehsan et al., 2019). However none of these datasets include multi-turn fine-grained negotiations and argumentation, or allow the original model decision to be changed during interaction.

There do exist datasets and analyses that put the focus on the process of collaboratively reaching a conclusion or agreement, rather than explaining a specific decision or output. For example, Myrendal (2019) details the collaborative process of negotiating word meanings using the example of online forums. In these interactions, the participants are not necessarily reaching agreement, but give detailed explanations for the aspects of a word meaning that they find relevant in a particular situation. The FORUM data we have shown in Section 3.1 is similar in this respect. Similarly, Madumal et al. (2019) have analyzed dialog data from different explanation settings, including settings in which either the explainer or the explainee is an automated agent, empirically affirming Walton (2009)'s theoretical formalization of dialog structure.

Madumal et al. (2019)'s work makes visible that explanatory interactions contain a component of argumentation, in which claims can be questioned and counter-arguments be proposed. Computational argumentation has its own line of research that incorporates aspects that are important for modelling an explainer, e.g., Wachsmuth and Alshomary (2022) build a model that can generate arguments that are user-dependent, taking into account the explainee's previous beliefs.

While these settings assume specialized roles for the interaction participants, these roles could change throughout a conversation. In the domains of negotiation and collaborative problem-solving, participants are both explaining and being explained to. Both types of dialog require participants to justify their proposals or make counter-proposals to achieve either a common or individual goal. Data exists based on in-lab data collections (Petukhova et al., 2016; Karadzhov et al., 2022a) and Wikipedia data (De Kock and Vlachos, 2021). DeliData (Karadzhov et al., 2022a) is very similar to the WORDLE game we are proposing here in that participants are explaining fine-grained options for a solution, but there is no image context.

6 Conclusion and Outlook

All three settings that we have shown elicit aspects of dialogical classification explanation. In the next step, we plan to collect a larger number of dialogs from each game and annotate them systematically in order to add a quantitative analysis to the qualitative examples we have shown here.

Acknowledgments

We are very thankful to our students, Saskia Rauhut, Elena Kröner, Mariia Kurilina, and Sebastiano Gigliobianco for implementing the tasks and carrying out the data collections. We also thank the anonymous reviewers for their valuable feedback.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA. Association for Computing Machinery.

Christine De Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 263–274.

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 Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022a. DeliData: A dataset for deliberation in multi-party problem solving. *arXiv:2108.05271*.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022b. What makes you change your mind? An empirical investigation in online group decision-making conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 552–563, Edinburgh, UK. Association for Computational Linguistics.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 577–593, Cham. Springer International Publishing.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 570–586, Cham. Springer International Publishing.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS

’19, pages 1033–1041, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38.

Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse Studies*, 21(3):317–339.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Volha Petukhova, Christopher Stevens, Harmen de Weerd, Niels Taatgen, Fokie Cnossen, and Andrei Malchanau. 2016. Modelling Multi-issue Bargaining Dialogues: Data Collection, Annotation Design and Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3133–3140, Portorož, Slovenia. European Language Resources Association (ELRA).

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. Slurk – A Lightweight Interaction Server For Dialogue Experiments and Data Collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / Semdial 2018)*, Aix-en-Provence, France.

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. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Douglas Walton. 2009. Objections, Rebuttals and Refutations. *OSSA Conference Archive*.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-201, Caltech.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

A Data usage

We have used material from a number of sources. For the SiJu and JuJu tasks, we have used texts and images from `https://forums.whatbird.com` with their explicit permission that was granted via email. Images for these tasks also come from the CUB dataset (Welinder et al., 2010). For the WORDLE tasks, we have used images from ImageNet ILSVRC2012 dataset (Russakovsky et al., 2015) and Wikimedia Commons (`https://commons.wikimedia.org/wiki/Main_Page`). The corresponding target words have been selected manually for now.

B Example dialogs

Figures 9, 10, and 11 show example dialogs for the SiJu, JuJu, and WORDLE games, respectively.

C Chat interfaces

Figure 12 shows the complete user interfaces for the three data collections SiJu, JuJu, and WORDLE at a larger scale.

#	A	Game Master	B
	[Task intro]		
			[B⟩A] Great! Let's begin.
1	[A⟩B] ACTION: BUTTON-AGREE		
2			[B⟩A] Really? I don't agree. Which attributes in the text make you think so? Convince me!
3	[A⟩B] The description matches the image - all components have a corresponding part to the image. You could disagree on the disambiguation of Warbler but the colour, pattern, plummage and characteristics of the bird match the description.		
4	[A⟩B] ACTION: BUTTON-NEXT		
5			[B⟩A] I'm not convinced yet. What do you think about the other attributes, such as the side and the feet?
6	[A⟩B] The description states that the sides have distinctive orange-red patches. The image shows that the sides have distinctive orange-red patches. The description states that the feet are black. The image shows that the feet are black.		
7	[A⟩B] ACTION: BUTTON-NEXT		
8			[B⟩A] Okay, let's get you to the next round. 2 to go!
9	[A⟩B] ACTION: BUTTON-AGREE		
10			[B⟩A] Really? I'm not so sure. Which attributes in the text make you think so? Convince me!
11	[A⟩B] The description matches the image - all components have a corresponding part to the image. The colour, pattern, plummage and characteristics of the bird match the description.		
12	[A⟩B] ACTION: BUTTON-NEXT		
13			[B⟩A] I'm still not sure. What about the other attributes, like the head and the eyes?
14	[A⟩B] You could disagree that the head has a purple sheen, as some individuals have a higher sensitivity to colour. However, the eyes are very dark. Unmistakeably so.		
15	[A⟩B] ACTION: BUTTON-NEXT		
16			[B⟩A] Okay, let's get you to the next round. 1 to go!
17	[A⟩B] ACTION: BUTTON-MAYBE		
18			[B⟩A] Alright. Are there any features that you ARE certain about? Which attributes in the text make you unsure?
19	[A⟩B] Small. Yellow Underparts. chestnut-brown ear patch, black crown.		
20	[A⟩B] ACTION: BUTTON-DONE		
21			[B⟩A] That was the last round. Thank you for participating!

Image			
Description	Medium, active warbler with black upperparts and hood, distinctive orange-red patches on wings, sides, and long, fanned tail, and white underparts. Bill, legs, and feet are black.	Small blackbird with purple sheen on head, back, breast. Eyes are dark. Black bill, legs, feet.	Small warbler, olive-yellow upperparts, thick, black streaks on yellow underparts. Bright yellow face, chestnut-brown ear patch, black crown. Wings are dark with large white patches.

Figure 9: A transcript from a SIJU interaction. Text and description match for the first two images. The mismatch in the third pair is a difficult instance as parts of the description are correct.

#	A	Game Master	B
	[Task intro]		
	[GM⟩A,B] Yay! The game will begin now.		
1	[A⟩B] This is a test dialog.		
2		[B⟩A] None of these is a gul, so A) doesn't go with any picture	
3	[A⟩B] B) neither		
4	[A⟩B] none of these match		
5		[B⟩A] DECISION: NO MATCH	
6		[GM⟩B] Are you sure? Please discuss some more!	
7	[A⟩B] DECISION: NO MATCH HERE		
8	[GM⟩A] Are you sure? Please discuss some more!		
9	[A⟩B] c could go with 3, but 3 doesn't have a black face		
10		[B⟩A] or a pink bill	
11		[B⟩A] it's more grayish	
12	[A⟩B] DECISION: NO MATCH		
13	[GM⟩A] Let's wait for your partner to also type /decision.		
14		[GM⟩B] Your partner thinks that the two of you have made a decision. Type /decision and a brief explanation if you agree.	
15		[B⟩A] DECISION: NO MATCH	
16	[GM⟩A,B] Ok, let's get the two of you the next level. 2 to go!		
17	[A⟩B] d goes with 4		
18		[B⟩A] the others dont seem to match	
19	[A⟩B] none of these have a white neck		
20		[B⟩A] and there's no description that's saying anything about a bird with a red head	
21	[A⟩B] nothing with bright yellow either		
22	[A⟩B] DECISION: 4D		
23	[GM⟩A] Are you sure? Please discuss some more!		
24	[A⟩B] yes I'm sure		
25	[A⟩B] DECISION: 4D		
26	[GM⟩A] Are you sure? Please discuss some more!		
27	[A⟩B] yes sure		
28		[B⟩A] I'm sure too	
29		[B⟩A] DECISION: 4D	
30		[GM⟩B] Let's wait for your partner to also type /decision.	
31	[GM⟩A] Your partner thinks that the two of you have made a decision. Type /decision and a brief explanation if you agree.		
32	[A⟩B] DECISION: 4D		
33	[GM⟩A,B] Ok, let's get the two of you the next level. 1 to go!		
34		[B⟩A] ooh tricky	
35		[B⟩A] c3?	
36	[A⟩B] yes agree		
37	[A⟩B] and 1a		
38		[B⟩A] test	
39		[B⟩A] DECISION: 1A 3C	
40		[GM⟩B] Are you sure? Please discuss some more!	
41	[A⟩B] yes sure		
42	[A⟩B] sure		
43	[A⟩B] DECISION: 1A 3C		
44	[GM⟩A] Let's wait for your partner to also type /decision.		
45		[GM⟩B] Your partner thinks that the two of you have made a decision. Type /decision and a brief explanation if you agree.	
46		[B⟩A] DECISION: 1A 3C	
47	[GM⟩A,B] The game is over! Thank you for participating!		

Figure 10: A transcript from a JUJU interaction.

#	A	Game Master	B
1		[GM⟩A,B] Welcome to Wordle with Images	
2		[GM⟩A,B] Let's start with the first of 1 images	
3	[A⟩B] Hello!		
4			[B⟩A] hello
5			[B⟩A] my guess is the word music
6	[A⟩GM] GUESS: music		
7		[GM⟩A] Let's wait for your partner to also enter a guess.	
8		[GM⟩B] Your partner thinks that you have found the right word. Enter your guess.	
9			[B⟩GM] GUESS: music
10		[GM⟩A,B] SUBMIT GUESS: music - FEEDBACK: M U S I C	
11	[A⟩B] Do you have an image? I'm just seeing normal wordle		
12			[B⟩A] yes i have an image, its of a homeless looking guy playing a flute with a dog
13	[A⟩B] Audio?		
14	[A⟩B] Wait no I		
15			[B⟩A] no aduio
16	[A⟩B] I've got no clue		
17	[A⟩B] Any other details in the image?		
18			[B⟩A] they arestanding on a crumpled blanket on a cobblestone like sidewalk the dog is a dalmation
19			[B⟩A] the guy has a sleevless tank top and shaggy hair
20	[A⟩B] Maybe buddy? Because of his dog?		
21			[B⟩A] yes probably buddy
22			[B⟩GM] GUESS: buddy
23		[GM⟩B] Let's wait for your partner to also enter a guess.	
24		[GM⟩A] Your partner thinks that you have found the right word. Enter your guess.	
25	[A⟩GM] GUESS: buddy		
26		[GM⟩A,B] SUBMIT GUESS: buddy - FEEDBACK: B U D D Y	
27	[A⟩B] Okay so we've got U and Y		
28	[A⟩B] B and D are wrong		
29			[B⟩A] i typed in funky before i switched to buddy
30	[A⟩B] Oooo funky could work, because of the music		
31	[A⟩GM] GUESS: funky		
32		[GM⟩A] Let's wait for your partner to also enter a guess.	
33		[GM⟩B] Your partner thinks that you have found the right word. Enter your guess.	
34			[B⟩GM] GUESS: funky
35		[GM⟩A,B] SUBMIT GUESS: funky - FEEDBACK: F U N K Y	
36	[A⟩B] Let's guess Funky		
37	[A⟩B] Lmao I'm a bit confused now		
38	[A⟩B] Perhaps the word is unrelated to the picture?		
39			[B⟩A] maybe
40	[A⟩B] Puppy? Yuggy?		
41	[A⟩B] Ruggy*		
42	[A⟩B] Not even sure Ruggy is a word		
43			[B⟩A] puppy or pushy?
44	[A⟩B] Can't be S, used that in Music		
45	[A⟩B] Wanna try Puppy?		
46			[B⟩A] sure
47	[A⟩GM] GUESS: puppy		
48		[GM⟩A] Let's wait for your partner to also enter a guess.	
49		[GM⟩B] Your partner thinks that you have found the right word. Enter your guess.	
50			[B⟩GM] GUESS: puppy
51		[GM⟩A,B] SUBMIT GUESS: puppy - FEEDBACK: P U P P Y	
52		[GM⟩A,B] YOU WON! For this round you get 10 points. Your total score is: 10	
53		[GM⟩A,B] The game is over! Thank you for participating!	

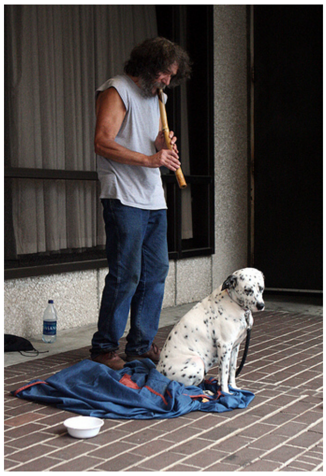


Figure 11: A shortened transcript from a Wordle interaction. The image is taken from the ImageNet ILSVRC2012 dataset (Russakovsky et al., 2015).

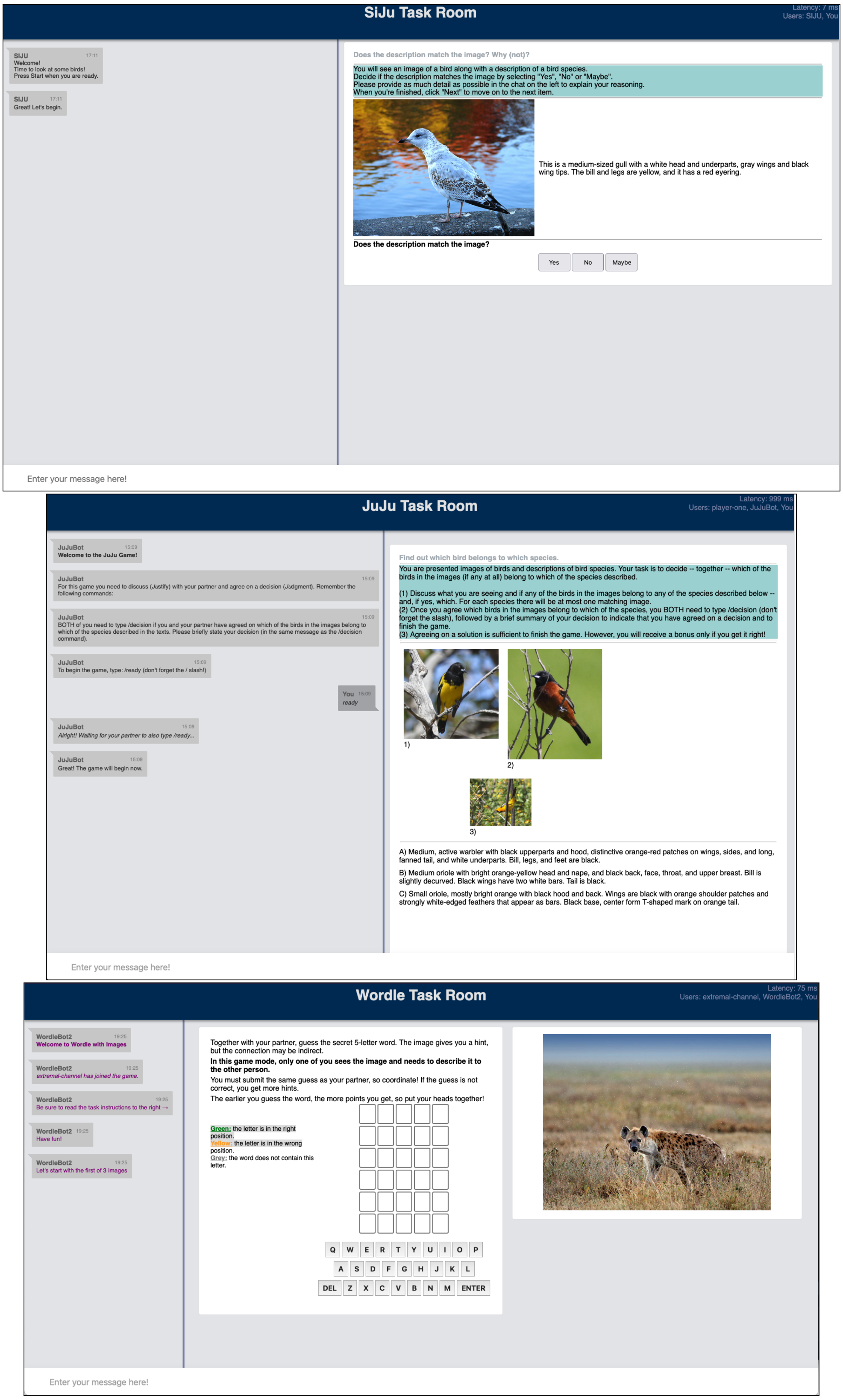


Figure 12: The visual interfaces for the SIJU (top), JUJU (middle), and WORDLE (bottom) games.

Lexical Level Alignment in Dialogue

Yikai Tseng **Takenobu Tokunaga**
Tokyo Institute of Technology / Tokyo, Japan
tseng.y.ab@m.titech.ac.jp take@c.titech.ac.jp

Hikaru Yokono
Meisei University / Tokyo, Japan
hikaru.yokono@meisei-u.ac.jp

Abstract

Alignment in dialogue is believed to make communication progress smoothly. Lexical alignment has been particularly well studied. However, we hypothesise that it is not just words that get aligned but also the overall difficulty level of the vocabulary used. For instance, when talking to children or non-native speakers, one chooses familiar words to ensure their partner can understand their utterances. We call this phenomenon "lexical level alignment (LLA)". This study investigates whether LLA occurs in natural dialogues and the factors influencing LLA by analysing an existing Japanese dialogue corpus. The analysis revealed that LLA occurs in dialogues between firstly-encountered native and non-native speakers.

1 Introduction

It is well known that alignment at various levels occurs between interlocutors in dialogue for successful communication (Pickering and Garrod, 2006). Pickering and Garrod (2004) proposed the interactive alignment account of dialogue, which assumes that the linguistic representations employed by the interlocutors become aligned at various levels as a result of a largely automatic process. However, a single-level alignment does not necessarily lead to a successful dialogue. The alignment at different levels depends on each other, i.e., alignment at one level leads to those of other levels, and the alignment in total leads to a successful dialogue (Pickering and Garrod, 2006).

Lexical alignment is a typical alignment phenomenon where linguistic descriptions by interlocutors converge during the course of dialogue, and they gradually use the same expression referring to an object (Garrod and Anderson, 1987). Lexical alignment is also attracting attention in the context of human-computer interaction, conversational agents and explainable artificial intelligence (Branigan et al., 2010; Srivastava et al., 2023).

Sys 1:	Is your *router* connected to the computer?
Usr 1:	Uh. What's a router?
Sys 2:	It's *the big black box*.
Usr 2:	Ok.. yes.
Sys 3:	Do you see a small white box connected to the router?
Usr 3:	Yes.
Sys 4:	Ok. Is there *a flashing monitor symbol at the bottom right of the screen*?
Usr 4:	*The network icon*?
Sys 5:	Yes. Is it flashing?
Usr 5:	Yes. It is flashing.
Sys 6:	Ok. Please open your *browser*.

Figure 1: Dialogue example (Janarthanam and Lemon, 2009)

Janarthanam and Lemon (2009) proposed a dialogue system for troubleshooting which can choose referring expressions depending on the lexical knowledge of the user. Figure 1 shows an example of their dialogue data. In Sys 1, the system uses the term "router", but the user does not understand the word and clarifies what it is in Usr 1. This clarification makes the system rephrase "router" with a simple expression "the big black box" in Sys 2, assuming that the user has little lexical knowledge in the network domain. The system continues to use simpler expressions like "a small white box" and "a flashing monitor symbol at the bottom right of the screen". However, once the user rephrases "a flashing monitor symbol" with "the network icon" in Usr 4, the system updates the user's lexical knowledge again. It starts to use technical terms like "browser" in Sys 6. Janarthanam and Lemon (2009) aimed to dynamically adapt the lexical choice to the user's lexical knowledge, as this example illustrates. Although they call this phenomenon lexical alignment as well, we claim it should be distinguished from conventional lexical alignment.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

Besides troubleshooting dialogues, when adults talk to children or native speakers to non-native speakers, they try to avoid difficult words in the first place and use easier words if their partner cannot understand their utterances. Namely, the native speaker aligns the lexical level of words in their utterance to their partner's. This phenomenon is different from well-known lexical alignment, where the lexical choice of interlocutors converges to align during the progress of dialogue. We call this phenomenon "lexical level alignment (LLA)". We expect LLA occurs in natural dialogue as likely as lexical alignment does. The system by Janarthanam and Lemon (2009) above can be considered to aim at realising LLA.

In this study, we investigate the phenomenon of LLA by analysing an existing Japanese dialogue corpus. Our research question is twofold.

RQ1: Does LLA occur in natural dialogue?

RQ2: What factors affect LLA?

Examining RQ2, we consider the following two factors: firstly, the intimacy between two interlocutors, whether friends or first-encounters; secondly, the language proficiency level of the interlocutors, whether a pair of native speakers or a pair of a native speaker and a non-native speaker.

2 Related Work

Lexical alignment, the alignment of words, has been widely studied and confirmed in various dialogues. Campano et al. (2014) confirmed that lexical alignment occurs in human-human dialogues both in natural settings and in Wizard of Oz settings, where one of the interlocutors plays the role of the virtual agent using limited utterances. Sinclair et al. (2018) analysed dialogues between second language (L2) learners and tutors and confirmed lexical priming, which indicates lexical alignment. They observed that alignment increases according to the ability of the L2 learners and the word complexity, and student-to-tutor alignment has a stronger priming effect than tutor-to-student alignment. Misiek et al. (2020) analysed child-adult dialogues and confirmed that lexical alignment occurs in both directions. In addition, they observed that adults align with children more than vice versa, even if the factor of language production ability was controlled. Although both Sinclair et al. (2018) and Misiek et al. (2020) consider the difference in lexical knowledge between interlocutors, their interest remains in lexical alignment. Wang et al. (2014) analysed multi-party conversations in online health communities and observed a strong lexical alignment effect.

Xu and Reitter (2015) compared three metrics for measuring linguistic alignment: indiscriminate local linguistic alignment, repetition decay, and Spearman's correlation coefficient. The indiscriminate local linguistic alignment has the overall best performance; it is especially favourable concerning individuals' inherent propensity of alignment. The repetition decay is favourable for exploring the correlations between alignments at different linguistic levels. Spearman's correlation coefficient has poorer normality and consistency than the other two. These metrics are developed for lexical alignment. We need to develop metrics to measure LLA.

Buschmeier et al. (2009) presented an alignment-capable micro planner, SPUD prime, which uses a priming-based interactive alignment model to model human speakers' alignment behaviour. Hu et al. (2018) proposed the Dialog Adaptation Score (DAS) measure to evaluate the adaptation in generated dialogues.

While the past lexical alignment research focuses on individual words from a microscopic viewpoint, we look at the alignment of a macroscopic property, i.e., the lexical level of interlocutor utterances.

3 Data

3.1 Dialogue corpus

We use the BTSJ[1] Japanese 1000-person natural conversation corpus[2] (BTSJ-1000 corpus hereafter) (USAMI, 2023) for analysis. The BTSJ-1000 corpus contains 514 dialogues in various settings totalling 127 hours. The interlocutors have various demographic properties regarding gender, age, first language, and professions. Relations between interlocutors also vary. This demographic information is helpful for us to investigate the factors that affect LLA. The BTSJ-1000 corpus contains dialogues in various situations, such as paper writing, interview, role-play of apology dialogues, and so on. Since we want to analyse the phenomenon of LLA in natural dialogues, we consider only 368 chat dialogues of general topics such as travel and school life in this study. Most of the themes of these chat dialogues are left to the interlocutors.

[1]Basic Transcription System for Japanese
[2]https://isplad.jp/lab/btsj_corpus_2023/

Table 1: Number of dialogues in the BTSJ-1000 corpus

	N-N	N-L
friend	141	43
first-encounter	125	59

Furthermore, we categorise these dialogues regarding two factors: intimacy between interlocutors and the language proficiency level of interlocutors. We have two cases for each factor, i.e., “friend” vs. “first-encounter” for intimacy, and “N-N” and “N-L” for proficiency level, where N-N stands for a pair of native Japanese speakers, while N-L means a pair of a native speaker and a Japanese learner, i.e., a non-native speaker. Table 1 shows the number of dialogues in each category.

3.2 Metric of lexical level

We need a metric to measure lexical level in Japanese (Tellols et al., 2023) to assess LLA in Japanese dialogues. One common metric of Japanese lexical level is the JLPT[3] level, which classifies the vocabulary into five discrete levels from N5 to N1, with N5 being the easiest and N1 the hardest. However, there is no available official vocabulary list for the JLPT level, and the coverage is lower, with only less than 10,000 words in total in an unofficial vocabulary list[4]. Another common metric is the occurrence frequency of BCCWJ (Balanced Corpus of Contemporary Written Japanese) (Maekawa et al., 2014). However, the frequency-based metric heavily depends on the corpus. In addition, since BCCWJ collects written text, it is unsuitable for dialogue analysis.

To remedy these problems, we adopt the WLSP[5] familiarity rate as the metric of lexical level. WLSP is a popular Japanese thesaurus, including 96,557 words with four syntactic categories (nominal, verbal, modifier, and other) and hypernymy and synonymy relations among them (National Institute for Japanese Language and Linguistics, 2004). Asahara (2019) collected familiarity ratings of words in WLSP through the Yahoo! crowdsourcing platform with 3,392 participants. The participants were asked to answer the familiarity of words regarding the five perspectives: KNOW, WRITE, READ, SPEAK, and LISTEN. To remove individual participant bias, a Bayesian linear mixed model was employed to estimate the familiarity rate for each word. Familiar words are assigned a higher value.

As we are focusing on dialogue, we use the LISTEN familiarity as the metric for lexical level, which represents how often one listens to the word. Low-familiarity words would be difficult to understand for the listener.

4 Analysis

4.1 Preprocessing

4.1.1 Dialogue data

The transcribed text of the BTSJ-1000 corpus contains annotation of paralinguistic information, such as filler, intonation and interruption. Since we are measuring lexical level, we remove this paralinguistic annotation in utterances and leave the content of the utterances.

4.1.2 Extracting WLSP words from utterances

To measure the lexical level in terms of WLSP familiarity, we need to extract WLSP words from the utterances. Since words are not separated by whitespaces in Japanese sentences, we first conduct the segmentation of utterances into tokens by a morphological analyser MeCab[6] with UniDic[7] v3.1.0 as the dictionary.

To convert the tokens in utterances into WLSP words, we use the WLSP2UniDic[8] list, which provides the association between WLSP words and UniDic tokens. However, this list only covers WLSP words corresponding to a single UniDic token. Covering multi-token words is essential since a token may occur more often with other tokens than occurs alone, resulting in a higher familiarity for the multi-token word than the token itself. For instance, the WLSP word “*gozaimasu* (an auxiliary verb for a polite form)” with the familiarity rate (LISTEN) of 1.48 is tokenised into two UniDic tokens “*gozaru*” and “*masu*”. Although the token “*gozaru*” is also a WLSP word, it does not frequently occur and its familiarity rate (LISTEN) is -0.51, being less familiar than that of “*gozaimasu*”.

To ensure the validity of word familiarity, we extend the WLSP2UniDic list to cover multi-token WLSP words as follows. First, we tokenise the

[3] Japanese Language Proficiency Test
[4] http://www7a.biglobe.ne.jp/nifongo/data/noryoku.html
[5] Word List by Semantic Principles
[6] https://taku910.github.io/mecab/
[7] https://clrd.ninjal.ac.jp/unidic/en/
[8] https://github.com/masayu-a/wlsp2unidic

WLSP words not in the original WLSP2UniDic list. However, we cannot simply apply MeCab to the unlisted WLSP words since MeCab cannot accurately tokenise them without a surrounding context. Therefore, we conduct a string-based search for utterances that include the unlisted WLSP words in the dialogue corpus and tokenise the utterances with MeCab. After confirming the consistency of the token boundary and readings between the unlisted word and the MeCab output, the corresponding token sequence for the unlisted WLSP word is added to the extended WLSP2UniDic list. In addition, we ignore the 19 unlisted words consisting of a single *Hiragana*[9] since they are not commonly used and cause many false matches.

We then construct a UniDic-to-WLSP list by inverting the extended WLSP2UniDic list, which maps UniDic token sequences to WLSP words. If multiple WLSP words in the extended WLSP2UniDic list correspond to a token sequence, we select the WLSP word with the highest familiarity, assuming that words with higher familiarity are more likely to occur.

Finally, with the tokenised utterances and the UniDic-to-WLSP list, we extract WLSP words from the utterances using a dynamic programming algorithm. Specifically, we compare the tokenised utterances with the UniDic-to-WLSP list to find a WLSP word sequence that minimises the number of unmatched tokens and the number of extracted words.

4.2 Method

As with lexical alignment, LLA is expected to occur as the dialogue progresses. When LLA occurs, the difference in the lexical level of the words used by the two interlocutors becomes smaller. To capture LLA, we divide each dialogue into two halves with the same length (in terms of the number of turns) and measure the lexical level of the utterances by each interlocutor in each dialogue segment. Consider a dialogue between A and B. Let $LL_p^{(j)} (p \in \{A, B\}, j \in \{1, 2\})$ be the lexical level of the utterances by p in the j-th half of the dialogue. When LLA occurs, we have

$$\Delta := |LL_A^{(2)} - LL_B^{(2)}| - |LL_A^{(1)} - LL_B^{(1)}| < 0.$$

That is, the difference in the lexical level of the interlocutors' utterances decreases in the later half of the dialogue. We calculate $LL_p^{(j)}$ based on the word types used in all p's utterances in the j-th dialogue segment.

We classify the dialogues into four groups by the two factors described in the research questions and perform a hypothesis test to check whether LLA occurs in each group of dialogues.

4.2.1 Lexical level of utterances

We assume that each interlocutor has their lexical level, representing that they understand all words with familiarity at least this level. Under this assumption, we define the lexical level of a word set as follows. After arranging the words in ascending order of familiarity, i.e., less familiar to more familiar, we assume that interlocutors can communicate even though they do not know the first $q\%$ of the difficult words in the list. Then, we define the lexical level of utterances $LL_p^{(j)}$ as the lexical level of the word set used in the utterances by p in the j-th dialogue segment. In this study, we consider 25 and 50 (the first and second quartiles) for q. That is, we assume that the interlocutors understand 75% and 50% of the words used by their partners.

Since the lexical alignment implies the interlocutors use the same words, it automatically induces LLA. To ensure that LLA is not just a by-product of lexical alignment, we exclude those words used by both interlocutors when calculating $LL_p^{(j)}$.

4.2.2 Hypothesis test

We conduct a hypothesis test to show that LLA occurs. The null hypothesis (H_0) assumes that LLA does not occur; in this case, there is no change in the lexical-level difference of the interlocutors' utterances between the first and second half of the dialogues. The alternative hypothesis (H_1) assumes that LLA occurs; in this case, the difference change Δ is negative, meaning the difference of the lexical level becomes smaller as the dialogue progresses.

$$\begin{aligned} H_0 &: \quad \Delta = 0 \\ H_1 &: \quad \Delta < 0 \end{aligned}$$

Since the distribution of the lexical level of utterances is unknown, we test the hypothesis with a one-sided permutation test with the resample count set to 100,000.

4.3 Result

Table 2 shows the result of the permutation test. The # column shows the number of dialogues in each group, and the "$q = N$" columns show the mean values of the lexical level difference change

[9] One of the Japanese writing scripts, a phonogram.

Table 2: Result of the permutation test. The numbers outside and inside the parentheses are the mean values of Δ and the P-values, respectively. The asterisk ($*$) indicates statistical significance at $p < .05$.

Dialogue group	#	$q = 25$	$q = 50$
N-L first-encounter	59	$-.040^*(.026)$	$.003\ (.569)$
N-L friend	43	$.005\ (.601)$	$-.002\ (.441)$
N-N first-encounter	125	$-.013\ (.103)$	$-.005\ (.247)$
N-N friend	141	$.020\ (.932)$	$.015\ (.974)$

Δ and their P-values in parentheses. From this result, we can see that while there is no significant change in lexical level in the N-L friend dialogues and N-N dialogues, there is a significant decrease in lexical level difference in the N-L first-encounter dialogues. This result suggests that LLA occurs in the N-L first-encounter dialogues even if we eliminate the effect of lexical alignment.

5 Discussion

5.1 Factors that affect LLA

LLA occurs in the first-encounter dialogues but not in the friend dialogues. In first-encounter dialogues, the interlocutors initially do not know their partners' lexical level but can estimate the level as the dialogue progresses. Therefore, they try to align their lexical level later in the dialogue. On the other hand, in the friend dialogues, the interlocutors already know their partners' lexical level before the dialogue. Therefore, their lexical levels can be aligned from the beginning. We cannot observe LLA during dialogue in this case.

LLA occurs in the N-L dialogues but not in the N-N dialogues. In the N-L dialogues, the native speaker might consider that their partner may not be able to understand difficult words and try to estimate and align the lexical level to their partner. In addition, the non-native speaker might also try to align their lexical level to their partner, which might be considered a language learning process. On the other hand, in the N-N dialogues, the interlocutors might assume that their partners would understand most of the words. It is unnecessary to consider the lexical level, so LLA does not occur.

LLA is not observed when measuring the lexical level of utterances calculated with $q = 50$. We do not observe LLA when calculating the lexical level of utterances with $q = 50$. Knowing more than 50% of the words might be needed to understand their partners' utterances.

Table 3: Number of dialogues with and without LLA when using the entire and the difference word sets. "O" and "X" indicate whether LLA is observed ($\Delta < 0$) or not ($\Delta \geq 0$), respectively.

Entire	Diff.	#	(%)
O	O	135	(34.5)
X	X	127	(36.7)
O	X	48	(13.0)
X	O	58	(15.8)

5.2 The validity of using difference word set

To ensure that LLA is not just a by-product of lexical alignment, we excluded the words used by both interlocutors, i.e., we excluded the intersection of the interlocutor word sets and considered the difference word sets when calculating the lexical level of utterances. To assess the validity of using the difference word sets, we compare Δ of each dialogue when using the entire and difference word sets in the calculation of lexical level.

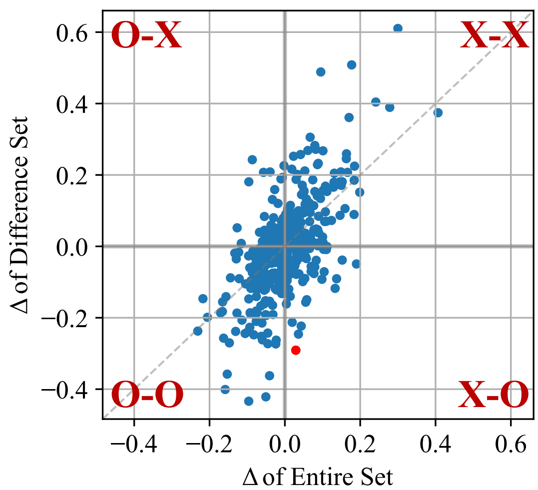


Figure 2: Δ pairs when using the entire and the difference word sets

Table 3 shows the number of dialogues which

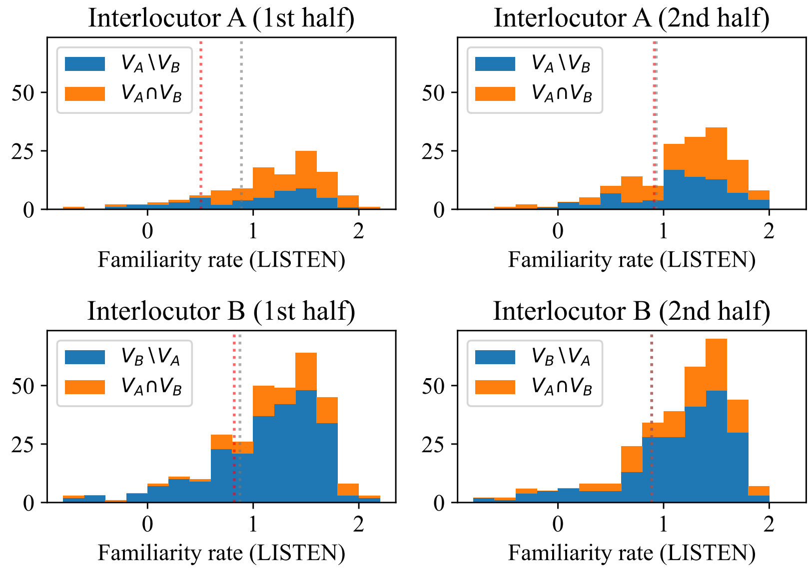


Figure 3: Familiarity rate distribution of word sets in the dialogue denoted by the red data point in Figure 2. V_A and V_B represent the word set of utterances by interlocutors A and B, respectively. The blue and orange bars represent the distribution of word familiarity rate for the difference set and the intersection of the word sets of utterances by both interlocutors, respectively. The orange bars are stacked on the blue bars; therefore, the sum of both bars represents the distribution of word familiarity rate for the entire set. The grey and red dashed lines represent the lexical level of the entire and the difference word sets of utterances, respectively.

are divided into categories according to whether LLA is observed ("O") using the entire and the difference word sets. For instance, we have 135 dialogues where we observe LLA using both the entire and the difference word sets (the O-O column). These dialogues indicate that LLA is not a by-product of lexical alignment. On the other hand, in the dialogues of the O-X column, LLA observed using the entire word set can be considered a spurious one induced by lexical alignment.

However, 15.8% of the dialogues (the X-O column) are unexpected since it indicates that LLA is not observed when considering the entire set while observed after the intersection is excluded. To investigate the reason for these unexpected cases, we analyse the familiarity rate distribution of word sets in these dialogues. Figure 2 shows the scatter diagram plotting Δ using the entire word set on the x-axis and Δ for the difference word set on the y-axis. Each quadrant corresponds to each column in Table 3. We pick up the red data point, which has the largest difference in Δ between the entire and the difference word set, and calculate the familiarity rate distribution of the words in the dialogue.

Figure 3 shows the distribution of the familiarity rate for the dialogue indicated by the red point in Figure 2. Comparing the distribution of the interlocutor A and B, while B's distributions are similar between the entire and the difference sets, A's distributions are less similar. As the distribution determines the lexical levels (grey and red dashed lines), its shape directly affects the lexical level value. In the second half of the dialogue, since the proportion of words with a familiarity rate higher than 1.0 is almost the same for the entire and the difference sets, the lexical levels of the two sets are similar for A, even though their distributions are quite different. On the other hand, in the first half of the dialogue, there is a non-negligible peak at $0.4 \sim 0.6$ for the difference set, which deviates the lexical level of the difference set from that of the entire set. Considering that the lexical level of the entire set tends to be lower than that of the difference set, we observe LLA only for the difference set. This example reveals the limitation of using a fixed q value for measuring the lexical level of utterances regardless of the familiarity distribution.

5.3 LLA patterns

In 4.2.2, we analysed LLA from a macroscopic viewpoint with a hypothesis test. We can also analyse it from a microscopic viewpoint by investigating the lexical level change of the interlocutors between the first and the second half of individual dialogues.

First, we consider whether both interlocutors contribute to LLA (two-way alignment) or only one of them does (one-way alignment). For dialogues between interlocutor A and B, let $\Delta_p := LL_p^{(2)} - LL_p^{(1)} (p \in \{A, B\})$. When LLA occurs ($\Delta < 0$), the two-way and one-way alignments are formulated as follows.

Two-way alignment: $\Delta_A \Delta_B < 0$

One-way alignment: $\Delta_A \Delta_B > 0$

Figure 4 illustrates the corresponding alignment patterns[10].

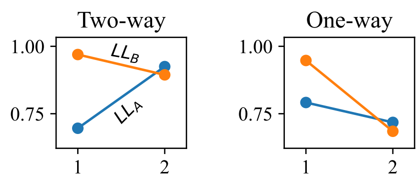


Figure 4: Example of the two LLA patterns regarding contributors. The "1" and "2" on the horizontal axis represent the first and the second half of the dialogue. The vertical axis represents the lexical level of utterances used by each interlocutor in each half.

Table 4 shows the number of each pattern. We can see that regardless of the intimacy of interlocutors, the two-way alignment occurs more than the one-way alignment in the N-L dialogues, but the opposite happens in the N-N dialogues. This suggests that native and non-native speaker pairs jointly tend to align their lexical level with each other, but it is not the case for native speaker pairs.

Regarding LLA in the N-L dialogues, we also consider the direction of alignment. Specifically, we focus on the absolute lexical level change of the interlocutor utterances from the first half to the second half of the dialogue $|\Delta_p|$, and assume that the interlocutor with larger $|\Delta_p|$ aligns to their partner. We have the following two patterns of lexical level change (Figure 5).

- N-to-L alignment: $|\Delta_N| > |\Delta_L|$
- L-to-N alignment: $|\Delta_N| < |\Delta_L|$

[10]The one-way alignment example in figure 4) shows the case where both interlocutors use more difficult words in the second half (i.e., $\Delta_A, \Delta_B < 0$). However, it is also possible that both of them use easier words (i.e., $\Delta_A, \Delta_B > 0$).

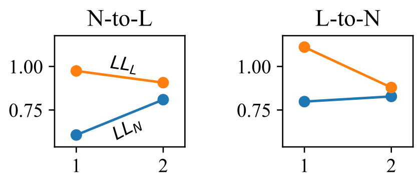


Figure 5: Example of the two LLA patterns regarding alignment direction. The "1" and "2" on the horizontal axis represent the first and the second half of the dialogue. The vertical axis represents the lexical level of utterances used by each interlocutor in each half.

Table 5 shows the number of each pattern. We can see that the L-to-N alignment occurs more than the N-to-L alignment, especially in the first-encounter dialogues. This result indicates that non-native speakers try to align their lexical level to native speakers as a part of the language learning process.

5.4 Alignment and dialogue quality

Alignment contributes to a successful dialogue. Here, we investigate the relationship between LLA and dialogue quality. It is, however, difficult to define dialogue quality in general. We focus on an aspect of to what extent both interlocutors speak equally to assess dialogue quality. Specifically, we consider the ratio of the WLSP word count per turn[11] between two interlocutors as the metric for dialogue quality. We take the larger word count as the denominator to make the metric range between 0 and 1. Therefore, a larger value means both interlocutors speak equally, and hence the dialogue has higher quality.

Figure 6 shows the relation between Δ and dialogue quality. While there is no correlation between Δ and dialogue quality for all dialogues, there is a weak tendency that smaller Δ, i.e. high LLA, leads to higher dialogue quality for the N-L first-encounter dialogues only, with Pearson correlation coefficient being -0.343 and P-value being 0.008. Our metric for dialogue quality is a rough approximation and captures only one of many other aspects. We need to investigate further the relationship between LLA and other aspects of dialogue quality more precisely.

[11]We also tried "ratio of UniDic token count per turn" and "ratio of vocabulary set size" and obtained similar results.

Table 4: Distribution of LLA patterns regarding contributors

Dialogue group	#	Two-way	One-way	No alignment
N-L first-encounter	59	20 (34%)	14 (24%)	25 (42%)
N-L friend	43	13 (30%)	7 (16%)	23 (54%)
N-N first-encounter	125	30 (24%)	47 (38%)	48 (38%)
N-N friend	141	30 (21%)	32 (23%)	79 (56%)
Total	368	93 (25%)	100 (27%)	175 (48%)

Table 5: Distribution of LLA patterns regarding alignment direction

Dialogue group	#	N-to-L	L-to-N	No alignment
N-L first-encounter	59	6 (10%)	28 (47%)	25 (42%)
N-L friend	43	9 (21%)	11 (26%)	23 (53%)

6 Conclusion

This study discussed lexical level alignment (LLA) in dialogue, which has not received explicit attention in past research. Analysing a Japanese dialogue corpus, we showed that LLA is observed (RQ1) when the interlocutors' lexical levels differ, and they do not know their partner's lexical level (RQ2).

We used WLSP familiarity rate (LISTEN) as the metric of lexical level and defined the lexical level of utterances as the required lexical level for the interlocutor to communicate without knowing the most difficult $q\%$ (we used 25 and 50 for q in this study) of the words used in the utterances. Specifically, we excluded those words used by both interlocutors when calculating their lexical level to ensure that LLA is not just a by-product of lexical alignment.

We classified the dialogues into four groups by the familiarity between interlocutors (friend or first-encounter) and their language proficiency level (N-N or N-L). We performed a permutation test to see if LLA occurs in each group. Specifically, we considered the change of lexical level difference between the utterances by the two interlocutors from the first half to the second half of the dialogue, and verified whether the difference decreased. As a result, we confirmed that LLA occurs in first-encounter dialogues between a native speaker and a non-native speaker when q is set to 25.

In addition, we checked the validity of using the difference word set when calculating lexical level and confirmed that 71.2% of the dialogues have the same result after excluding the words used by both interlocutors; 13.0% of the dialogues have spurious LLA; 15.8% of the dialogues are unexpected, which suggests the limitation of using the fixed q value for measuring the lexical level of utterances.

We also analysed the LLA patterns. We first analysed whether both interlocutors contribute to LLA or only one of them does. We found that the two-way alignment occurs more than the one-way alignment in the N-L dialogues but not in the N-N dialogues. This tendency indicates that native and non-native speaker pairs jointly try to align with each other, but it is not the case for native speaker pairs. We then analysed the direction of alignment in the N-L dialogues. We found that the L-to-N alignment occurs more than the N-to-L alignment, especially in the first-encounter dialogues. This indicates that non-native speakers try to align with native speakers, which might be considered a part of the language-learning process.

Finally, we investigated the relationship between LLA and dialogue quality. We considered the word count ratio per turn between interlocutors as the metric of dialogue quality, assuming that interlocutors speak equally in successful dialogues. We observed a weak tendency that LLA leads to higher dialogue quality in the N-L first-encounter dialogues.

7 Future Work

As we discussed in 5.2, we had unexpected dialogues where LLA was observed only using the difference word sets of interlocutors. The detailed analysis suggests that such an anomaly is caused by calculating the lexical level of utterances without considering the word familiarity distribution in utterances. More sophistication is needed in

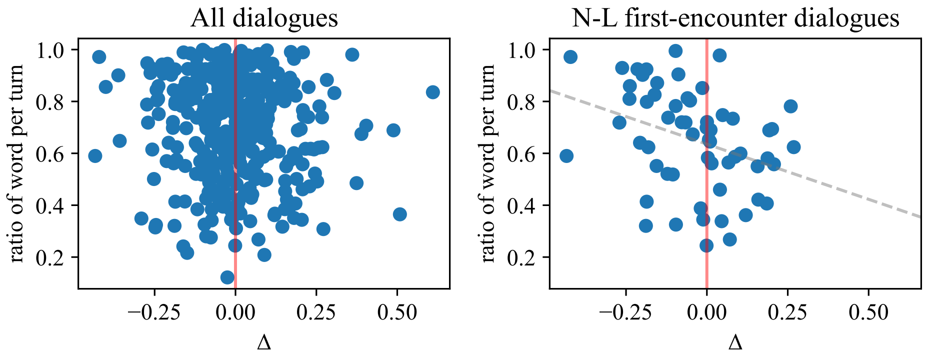


Figure 6: The relation between Δ (the metric of LLA) and "ratio of word count per turn" (the metric of dialogue quality). The grey dashed line represents the linear regression line.

measuring the lexical level of utterances.

We considered to what extent both interlocutors speak equally as a metric of dialogue quality in analysing the relation between LLA and dialogue quality. As we already mentioned, there are many other aspects of dialogue quality. For instance, lexical level gaps might cause frequent clarification, misunderstanding, or even dialogue breakdown. We would like to shed light on other aspects of dialogue quality and investigate their relation to LLA in future.

In addition, word difficulty is likely affected by the topic in dialogues. Therefore, the distribution of the lexical level of utterances can be unstable when the topic changes in the dialogue. We need to investigate the influence of the topic and how the lexical level aligns as it changes. In this study, we adopted a popular Japanese thesaurus WLSP, which assigns semantic categories to each word. We would also look at the relationship between the dialogue topic (change) and the distribution of the word categories in the utterances for investigating the LLA process.

Limitations

This study uses the WLSP familiarity rate for lexical level measurement, which might not be available for other languages. Besides, since we capture LLA from a macroscopic viewpoint, even though we confirmed that LLA occurs in the N-L first-encounter dialogues, the alignment process is still an open question. Further study is necessary for the dynamic nature of LLA.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K18358.

References

Masayuki Asahara. 2019. Word familiarity rate estimation using a Bayesian linear mixed model. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 6–14, Hong Kong. Association for Computational Linguistics.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for Natural Language Generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 82–89, Athens, Greece. Association for Computational Linguistics.

Sabrina Campano, Jessica Durand, and Chloé Clavel. 2014. Comparative analysis of verbal alignment in human-human and human-agent interactions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4415–4422, Reykjavik, Iceland. European Language Resources Association (ELRA).

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Zhichao Hu, Jean Fox Tree, and Marilyn Walker. 2018. Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 20–31, Melbourne, Australia. Association for Computational Linguistics.

Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 74–81, Athens, Greece. Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. 2020. Development of multi-level linguistic alignment in child-adult conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 54–58, Online. Association for Computational Linguistics.

National Institute for Japanese Language and Linguistics. 2004. *Bunrui goihyo zôho kaitei-ban* (Word List by Semantic Principles Revised and Enlarged Edition). Dainippon Tosyo, Tokyo.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Martin J. Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation*, 4:203–238.

Arabella Sinclair, Adam Lopez, C. G. Lucas, and Dragan Gasevic. 2018. Does ability affect alignment in second language tutorial dialogue? In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 41–50, Melbourne, Australia. Association for Computational Linguistics.

Sumit Srivastava, Mariët Theune, and Alejandro Catala. 2023. The role of lexical alignment in human understanding of explanations by conversational agents. In *The 28th International Conference on Intelligent User Interfaces (IUI '23),*, pages 423–435.

Dolça Tellols, Takenobu Tokunaga, Hilofumi Yamamoto, and Hikaru Yokono. 2023. Pic2PLex: A test for assessing free productive vocabulary. *Transactions of the Japanese Society for Artificial Intelligence*, 38(1):B–M52 1–11.

Mayumi (ed.) USAMI. 2023. Building of a Japanese 1000 person natural conversation corpus for pragmatic analyses and its multilateral studies, and ninjal institute-based projects: Multiple approaches to analyzing the communication of japanese language learners.

Yafei Wang, David Reitter, and John Yen. 2014. A model to qualify the linguistic adaptation phenomenon in online conversation threads: Analyzing priming effect in online health community. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 55–62, Baltimore, Maryland, USA. Association for Computational Linguistics.

Yang Xu and David Reitter. 2015. An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–67, Denver, Colorado. Association for Computational Linguistics.

Learning to generate and `corr-` uh I mean *repair* language in real-time

Arash Eshghi*† Arash Ashrafzadeh*
*Heriot-Watt University, Edinburgh, United Kingdom
†AlanaAI, Edinburgh, United Kingdom
`{a.eshghi, aa2070}@hw.ac.uk`

Abstract

In conversation, speakers produce language *incrementally*, word by word, while continuously monitoring the appropriateness of their own contribution in the dynamically unfolding context of the conversation; and this often leads them to repair their own utterance on the fly. This real-time language processing capacity is furthermore crucial to the development of fluent and natural conversational AI. In this paper, we use a previously learned Dynamic Syntax grammar and the CHILDES corpus to develop, train and evaluate a probabilistic model for incremental generation where input to the model is a purely *semantic generation goal concept* in Type Theory with Records (TTR)[1]. We show that the model's output exactly matches the gold candidate in 78% of cases with a ROUGE-l score of 0.86. We further do a zero-shot evaluation of the ability of the same model to generate *self-repairs* when the generation goal changes mid-utterance. Automatic evaluation shows that the model can generate self-repairs correctly in 85% of cases. A small human evaluation confirms the naturalness and grammaticality of the generated self-repairs. Overall, these results further highlight the generalisation power of grammar-based models and lay the foundations for more controllable, and naturally interactive conversational AI systems.

1 Introduction

People process language incrementally, in real-time (see Crocker et al. (2000); Ferreira (1996); Kempson et al. (2016) among many others), i.e. both language understanding and generation proceed on a word by word rather than a sentence by sentence, or utterance by utterance basis. This real-time processing capacity underpins participant coordination in conversation (Gregoromichelaki et al., 2012, 2020) and leads to many characteristic phenomena such as split-utterances (Poesio and Rieser, 2010; Purver et al., 2009), mid-utterance feedback in the form of backchannels (Heldner et al., 2013) or clarification requests (Healey et al., 2011; Howes and Eshghi, 2021), hesitations, self-repairs (Schegloff et al., 1977) and more.

Language generation – our focus here – is just as incremental as language understanding: speakers normally do not have a fully formed conceptualisation or plan of what they want to say before they start articulating, and conceptualisation needs only to be one step ahead of generation or articulation (Guhe, 2007; Levelt, 1989). This is possible because speakers are able to continuously monitor the syntax, semantics, and the pragmatic appropriateness of their own contribution (Levelt, 1989) in the fast, dynamically evolving context of the conversation. In turn this allows them to pivot or correct themselves on the fly if needed, e.g. because they misarticulate a word, get feedback from their interlocutors (Goodwin, 1981), or else the generation goal changes due to the dynamics of the environment.

Real-time language processing is likewise crucial in designing dialogue systems that are more responsive, more naturally interactive (Skantze and Hjalmarsson, 2010; Aist et al., 2006), and are more accessible to people with memory impairments (Addlesee et al., 2019; Addlesee and Damonte, 2023; Nasreen et al., 2021). Despite this importance, relative to turn-based systems, it has received little attention from the wider NLP community; perhaps because it has deep implications for the architecture of such systems (Schlangen and Skantze, 2009; Skantze and Schlangen, 2009; Kennington et al., 2014), which make them much harder to build and maintain.

In this paper, we extend the work of Purver and Kempson (2004); Hough and Purver (2012);

[1] All relevant code, models, and data are available at `https://bitbucket.org/dylandialoguesystem/dsttr/src/dsttr_arash_a/`

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

Hough (2015), who lay the theoretical foundations for incremental generation and later the processing of self-repairs in Dynamic Syntax (Kempson et al., 2001, 2016, Sec. 2.3). For the first time, we develop a probabilistic model for incremental generation (Sec. 3) that conditions next word selection on the current incrementally unfolding context of the conversation, and also on features of a *purely semantic generation goal concept*, expressed as a Record Type (RT) in Type Theory with Records (Cooper, 2012; Cooper and Ginzburg, 2015). The model is trained and evaluated on part of the CHILDES corpus (MacWhinney, 2000) using an extant grammar that was learned by Eshghi et al. (2013) from the same data. Results show that in the best case, the model output matches the gold generation test candidate in 83% of cases (Sec. 4.2). We then go on to experiment with and evaluate the ability of the same model to generate self-repairs in a zero-shot setting in the face of *revisions to the goal concept RT* under various conditions (Sec 4.3): viz. for forward-looking and backward-looking repair and at different distances from the reparandum. Automatic evaluation shows that it can generate self-repairs correctly in 85% of cases. A small human evaluation confirms the overall naturalness and grammaticality of the generated repairs. Overall, these results further highlight the generalisation power of grammar-based models (see also Mao et al. (2021); Eshghi et al. (2017) and lay the foundations for more controllable, and naturally interactive conversational AI systems.

2 Dynamic Syntax and Type Theory with Records (DS-TTR)

Dynamic Syntax (DS, Kempson et al., 2016; Cann et al., 2005; Kempson et al., 2001) is a process-oriented grammar formalism that captures the real-time, incremental nature of the dual processes of linguistic comprehension and production, on a word by word or token by token basis. It models the time-linear construction of *semantic* representations (i.e. *interpretations*) as progressively more linguistic input is parsed or generated. DS is idiosyncratic in that it does not recognise an independent level of structure over words: on this view syntax is sets of constraints on the incremental processing of semantic information.

The output of parsing any given string of words is thus a *semantic tree* representing its predicate-argument structure (see Fig. 1). DS trees are always binary branching, with argument nodes conventionally on the right and functor nodes to the left; tree nodes correspond to terms in the lambda calculus, decorated with labels expressing their semantic type (e.g. $Ty(e)$) and formulae – here as record types of Type Theory with Records (TTR, see Sec. 2.1 below); and beta-reduction determines the type and formula at a mother node from those at its daughters (Fig. 1). These trees can be *partial*, containing unsatisfied *requirements* potentially for any element (e.g. $?Ty(e)$, a requirement for future development to $Ty(e)$), and contain a *pointer*, $\diamond$, labelling the node currently under development.

Grammaticality is defined as parsability in a context: the successful incremental word-by-word construction of a tree with no outstanding requirements (a *complete* tree) using all information given by the words in a string. We can also distinguish *potential grammaticality* (a successful sequence of steps up to a given point, although the tree is not complete and may have outstanding requirements) from *ungrammaticality* (no possible sequence of steps up to a given point).

Fig. 1 shows "John arrives", parsed incrementally, starting with the axiom tree with one node ($?Ty(t)$), and ending with a complete tree. The intermediate steps show the effects of: (i) DS Computational Actions (e.g. Completion which moves the pointer up and out of a complete node or Anticipation which moves the pointer down from the root to its functor daughter.) which are language-general and apply without any lexical input whenever their preconditions are met; and (ii) Lexical Actions which correspond to words and are triggered when a word is parsed.

Context In DS, context, required for processing various forms of context-dependency – including pronouns, VP-ellipsis, and short answers, as well as self-repair – is the parse search Directed Acyclic Graph (DAG), and as such, is also process-oriented. Edges correspond to DS actions – both Computational and Lexical Actions – and nodes correspond to semantic trees after the application of each action (Sato, 2011; Eshghi et al., 2012; Kempson et al., 2015). Here, we take a coarser-grained view of the DAG with edges corresponding to words (sequences of computational actions followed by a single lexical action) rather than single actions, and we drop abandoned parse

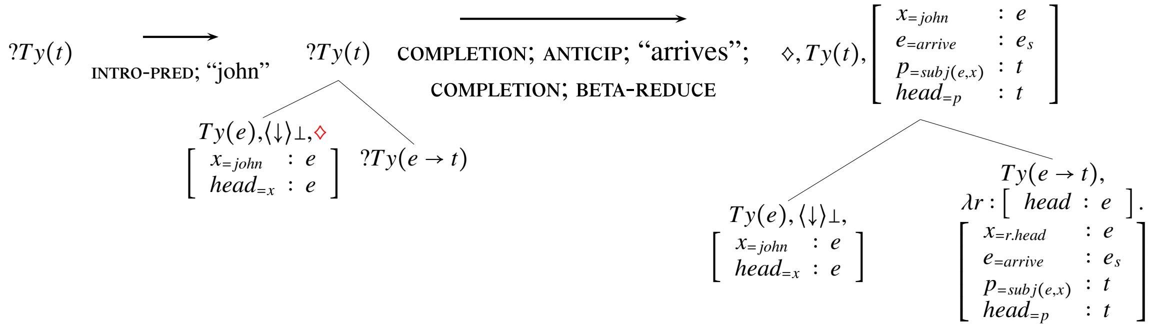


Figure 1: Incremental parsing in DS-TTR: *“John arrives”*

paths (see Eshghi et al., 2015; Howes and Eshghi, 2021, for details) – Fig. 4 shows an example.

2.1 Type Theory with Records (TTR)

Dynamic Syntax is currently integrated with TTR (Cooper, 2012, 2005) as the semantic formalism in which meaning representations are couched (Eshghi et al., 2012; Purver et al., 2011, 2010)[2].

TTR is an extension of standard type theory, and has been shown to be useful in contextual and semantic modelling in dialogue (see e.g. Ginzburg, 2012; Fernández, 2006; Purver et al., 2010, among many others), as well as the integration of perceptual and linguistic semantics (Larsson, 2013; Dobnik et al., 2012; Yu et al., 2017). With its rich notions of underspecification and subtyping, TTR has proved crucial for DS research in the incremental specification of content (Purver et al., 2011; Hough, 2015); specification of a richer notion of dialogue context (Purver et al., 2010); models of DS grammar learning (Eshghi et al., 2013); and models for learning dialogue systems from data (Eshghi et al., 2017; Kalatzis et al., 2016; Eshghi and Lemon, 2014).

In TTR, logical forms are specified as *record types*, which are sequences of *fields* of the form $\left[\ l : T\ \right]$ containing a label l and a type T. Record types can be witnessed (i.e. judged true) by *records* of that type, where a record is a sequence of label-value pairs $\left[\ l = v\ \right]$. We say that $\left[\ l = v\ \right]$ is of type $\left[\ l : T\ \right]$ just in case v is of type T. Fields can be *manifest*, i.e. given a singleton type e.g. $\left[\ l : T_a\ \right]$ where T_a is the type of which only a is a member; here, we write this as $\left[\ l_{=a} : T\ \right]$. Fields can also be *dependent* on fields preceding them (i.e. higher) in the record type (see Fig. 2).

[2]DS models the structural growth of representations and is agnostic to the formalism for semantic representation. As such, it has also been combined with RDF (Addlesee and Eshghi, 2021) and with vector-space representations (Purver et al., 2021)

$$R_1 : \begin{bmatrix} l_1 & : & T_1 \\ l_{2=a} & : & T_2 \\ l_{3=p(l_2)} & : & T_3 \end{bmatrix} \quad R_2 : \begin{bmatrix} l_1 & : & T_1 \\ l_2 & : & T_{2'} \end{bmatrix} \quad R_3 : [\,]$$

Figure 2: Example TTR record types

The standard subtype relation $\sqsubseteq$ can be defined for record types: $R_1 \sqsubseteq R_2$ if for all fields $\left[\ l : T_2\ \right]$ in R_2, R_1 contains $\left[\ l : T_1\ \right]$ where $T_1 \sqsubseteq T_2$. In Fig. 2, $R_1 \sqsubseteq R_2$ if $T_2 \sqsubseteq T_{2'}$, and both R_1 and R_2 are subtypes of R_3. This subtyping relation allows semantic information to be incrementally specified, i.e. record types can be indefinitely extended with more information and/or constraints.

Additionally, Larsson (2010) defines the meet ($\curlywedge$) operation of two (or more) RTs as the union of their fields; the equivalent of conjunction in FoL; see figure 3 for an example. We will need this below (Sec.3) where we define our probabilistic model.

$$\begin{bmatrix} l_1 & : & T_1 \\ l_2 & : & T_2 \end{bmatrix} \curlywedge \begin{bmatrix} l_2 & : & T_2 \\ l_3 & : & T_3 \end{bmatrix} = \begin{bmatrix} l_1 & : & T_1 \\ l_2 & : & T_2 \\ l_3 & : & T_3 \end{bmatrix}$$

Figure 3: Example of merge operation between two RTs

2.2 Generation in DS-TTR

As alluded to in the introduction, to handle typical incremental phenomena in dialogue such as split utterances, interruptive clarification requests or self-repair, any generation model must be as incremental as interpretation: full syntactic and semantic information should be available after generating every word with continual access to the incrementally unfolding context of the conversation (Hough and Purver, 2012; Eshghi et al., 2015).

In generation, there is an extra requirement on models, namely *representational interchangability* (Eshghi et al., 2011): parsing and generation should employ the same mechanisms and use the same kind of representation so that parsing can pick up where generation left off, and vice versa.

DS-TTR can meet these requirements, because generation employs exactly the same mechanisms as in parsing (Purver and Kempson, 2004) with the simple addition of a *subsumption check* against a *generation goal concept*, expressed as a Record Type (RT) in TTR (see Sec. 2.1); and where this goal concept can be partial (does not need to correspond to a complete sentence), and need only to be one step ahead of the generated utterance so far. This ease of matching incrementality in both generation and parsing is not matched by other models aiming to reflect incrementality in the dialogue model while adopting relatively conservative grammar frameworks, some matching syntactic requirements but without incremental semantics (Skantze and Hjalmarsson, 2010), others matching incremental growth of semantic input but leaving the incrementality of structural growth unaddressed (Guhe, 2007).

As such, generation involves *lexical search* whereby at every step, words from the lexicon are test-parsed in order to find words that (i) are parsable in the current context; and (ii) the resulting TTR semantics of the current DS tree subsumes or is monotonically extendable the generation goal. The subsumption relation is the inverse of the subtype relation defined above (see Sec. 2.1; i.e. R_1 *subsumes* R_2 iff $R_2 \sqsubseteq R_1$).

Without a probabilistic model for word selection at each step of generation, this process is effectively brute-force, computationally very inefficient, and therefore simply impractical, especially with large lexicons. This is the shortcoming that we address here for the first time by conditioning word selection on the generation goal RT. This involves learning, through Maximum Likelihood Estimation from data, $P(w|T, R_g)$, where w ranges over the lexicon, T is the current DS tree including its maximal semantics, and R_g is the generation goal. This parametrisation is described in full below in Sec. 3.

2.3 Processing Self-repair in DS-TTR

In this section, we briefly introduce the DS model of self-repair from (Hough and Purver, 2012): there are two types of self-repair that are addressed: *backward-looking repair* (aka. overt repair), where the repair involves a local, and partial restart of the reparandum, as in (1) and forward-looking repair (aka. covert repair) where the repair is simply a local extension, i.e. a further specification of the reparandum as in (2).

(1) “Sure enough ten minutes later the bell r-the doorbell rang" *(Schegloff et al., 1977)*

(2) “I-I mean the-he-they, y'know the guy, the the pathologist, looks at the tissue in the microscope. . . ” *(Schegloff et al., 1977)*

In the model set out above, a backward-looking repair arises due to an online revision of a generation goal RT, whereby the new goal is not a subtype of the one the speaker (or the dialogue manager) had initially set out to realise. We model this via backtracking along the incrementally available context DAG as set out above. More specifically, repair is invoked if there is no possible DAG extension after the test-parsing and subsumption check stage of generation (resulting in no candidate succeeding word edge).

The repair procedure proceeds by restarting generation from the last realised (generated) word edge. It continues backtracking by one DAG vertex at a time until the root record type of the current partial tree is a subtype of the new goal concept. Generation then proceeds as usual by extending the DAG from that vertex. The word edges backtracked over are not removed, but are simply marked as repaired (see also Eshghi et al. (2015) for a fuller account), following the principle that the revision process is on the public conversational record and hence should still be accessible for later anaphoric reference (see Fig. 4).

Forward-looking repairs on the other hand, i.e. *extensions*, where the repair effects an “afterthought” are also dealt with straightforwardly by the model. The DS-TTR parser simply treats these as monotonic extensions of the current tree, resulting in subtype extension of the root TTR record type. Thus, a change in goal concept during generation will not always put demands on the system to backtrack, such as in generating the fragment after the pause in “I go to Paris . . . from London". Backtracking only operates at a semantics-syntax mismatch where the revised goal concept is no longer a subtype of the root record type for the (sub-)utterance so far realised, as in Figure 4.

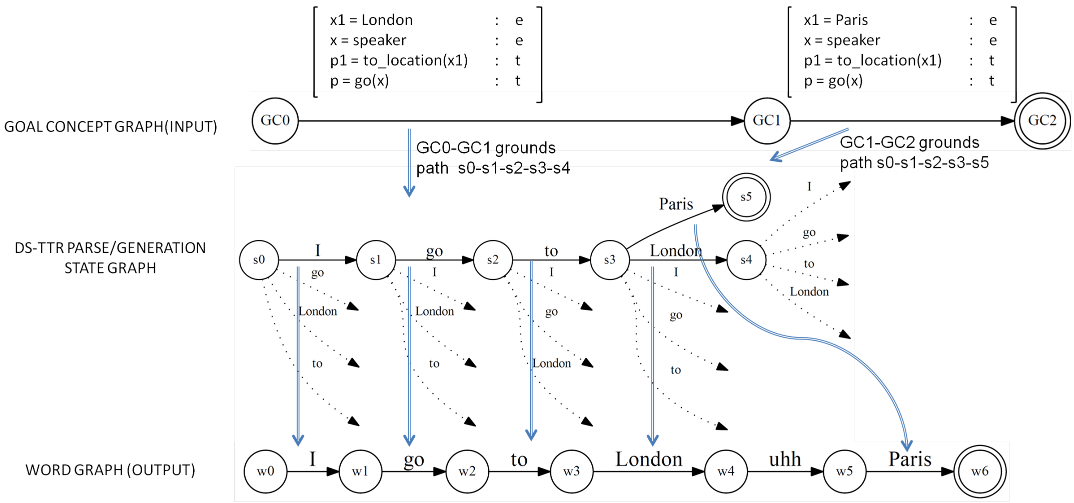


Figure 4: Incremental DS-TTR generation of a self-repair upon change of goal concept. Type-matched record types are double-circled nodes and edges indicating failed paths are dotted.

3 Probabilistic Model of Generation

In this section, we follow on from Sec. 2.3 above and describe the probabilistic model that we have developed for incremental probabilistic generation. First we describe the model itself, its parameters, and how these are estimated from data. Then we describe how the model is used at inference time to generate.

Model and Parameter Estimation As noted, generation in Dynamic Syntax is defined in terms of parsing. Specifically, it proceeds via lexical search, i.e. test-parsing (all) words from the lexicon while checking for *subsumption* against the *goal concept*: a record type (RT) in TTR; henceforth R_g. Words that parse successfully with a resulting (partial) semantics that subsume the goal concept are successfully generated. This process goes on until the semantics of the generated sentence equals the goal. This process is highly inefficient and impractical for larger lexicons.

On a high level, we solve this problem by building a probabilistic model which conditions the probability of generating the next word, w, on: (i) R_{cur}: the semantics of the generated utterance thus far; (ii) R_g, the goal concept; and (iii) the current DS tree (henceforth T_{cur}). We condition on (i) to allow the model to keep track of the semantics of what's already been generated, i.e. the left semantic context of generation; on (ii) to aid the model in selecting words that contribute the correct semantic increments to approach the goal concept; and on (iii) to capture the syntactic constraints on what words can grammatically follow. In sum, we need to compute $P(w|T_{cur}, R_{cur}, R_g)$ for all the words w in the lexicon.

As you will see below, we learn to generate by parsing, and therefore we the use Bayes rule in Eq. 3 to cast probabilistic generation roughly in probabilistic parsing terms:

$$\underbrace{P(w|T_{cur}, R_{cur}, R_g)}_{\text{probabilistic generation}} \overset{\text{Bayes Rule}}{=} \frac{\overbrace{P(T_{cur}, R_{cur}, R_g|w)}^{\text{probabilistic parsing}} P(w)}{P(T_{cur}, R_{cur}, R_g)} \tag{3}$$

On the right hand side of Eq. 3, $P(w)$ is the prior probability of w, which we obtain from the frequency of w in our training data; and $P(T_{cur}, R_{cur}, R_g)$ a normalisation constant which we do not need to estimate.

We learn $P(T_{cur}, R_{cur}, R_g|w)$ from gold data in the form of $\langle Utt = \langle w_1, \ldots, w_N \rangle, R_g \rangle$, where Utt is the utterance to be generated, and R_g is its gold semantics. To do this, we use the DS parser to parse Utt yielding a parse path (see e.g. Fig. 4) that starts with the DS axiom tree (empty tree) to the tree whose semantics is R_g together with all the DS trees produced after parsing each w_i in between; viz. a sequence $S_p = \{\langle T_1, w_1 \rangle, \ldots, \langle T_N, w_N \rangle\}$, where T_i are the DS trees in the context of which the w_i's were parsed. This sequence constitutes the observations from which we estimate $P(T_{cur}, R_{cur}, R_g|w)$ by Maximum Likelihood Estimation (MLE).

T_{cur}, R_{cur} and R_g are all composed of many individual features, and as a whole, would be observed very rarely. Therefore, for generalisation, we need to decompose them and compute the probability of the whole as the conjunction (product) of the probabilities of their individual atomic features.

For T_{cur} we follow Eshghi et al. (2013) and consider only one feature of T_{cur}: that of the type of the pointed node, or a requirement for a type (e.g. $Ty(e)$, $?Ty(e \rightarrow t)$, etc) – call this Ty_p. This simplifies the model considerably, but has the downside of not capturing all grammatical constraints (e.g. *person* constraints in English verbs will not be captured this way), and leading to some overgeneration.

We also simplify the model by conditioning on the semantics that *remains to be generated* – call it R_{inc} – rather than conditioning on both R_{cur} and R_g. We can compute R_{inc} each time through the well-defined *record type subtraction* operation in TTR where: $R_{inc} = R_g \backslash R_{cur}$.

With these simplifications, what we need to estimate by MLE from each sequences S_p (see above) is: $P(Ty_p, R_{inc}|w)$.

As noted, for any generalisation at all, R_{inc} now needs to be decomposed into its individual atomic features so that we can compute the probability of each of these features individually, rather than that of R_{inc} as a whole. We decompose R_{inc} as follows: $R_{inc} = \bigwedge_k(R_k)$, where $\bigwedge$ is the TTR equivalent of the conjunction operation in FoL (see above, Sec. 2.1); and each R_k is potentially *dependent* on R_j where $j < k$.

Using the probabilistic variant of TTR (Cooper et al., 2013, 2014), we can use the chain rule to then derive Eq. 4:

$$P(\bigwedge_k R_k|w) = \Pi_k P(R_k|w, R_1 \bigwedge \dots \bigwedge R_{k-1}) \quad (4)$$

This then allows us to express the probability of a complex record type in terms of the product of its potentially *dependent*, atomic supertypes. This, finally, puts us in a position to compute $P(Ty_p, R_{inc}|w)$ as follows:

$$P(R_{inc}, Ty_p|w) \overset{independence}{=} P(R_{inc}|w) \cdot P(Ty_p|w)$$
$$\overset{decompose\ R_{inc}}{=} P(\bigwedge_k R_k|w) \cdot P(Ty_p|w)$$

We implement the above procedure by constructing a 2D conditional count table where the rows are the words, and the columns are all the atomic semantic features observed during learning by parsing: effectively the result of decomposing all the R_g's in our data; this, in addition to all the Ty_p features we have observed on all the DS trees encountered in the S_p sequences above. Then, each time we observe an atomic semantic feature of R_{inc}, say, R_k, in the context of a word, w, we increment the cell (R_k, w) by 1. After learning, we normalise the columns of the table to obtain all $P(F|w)$ where F ranges over all semantic features and pointed node types, and w over all words in the lexicon.

Inference At inference time, we need to estimate $P(w|T_{cur}, R_{cur}, R_g)$: a probability distribution over all the words in the lexicon, given the current context of generation, T_{cur} including the current semantics so far generated, R_{cur}, and the goal concept, R_g. Given the above we take the following steps to *populate a beam* for generating the next word: (i) compute $R_{inc} = R_g \backslash R_{cur}$; (ii) compute all the atomic semantic features, R_k – the headings in the columns in our conditional probability table – that R_{inc} triggers or 'turns on'. This can be done by checking whether $R_{inc} \sqsubseteq R_k$; (iii) compute the single Ty_p (type of pointed node) feature by observing the type of the pointed node on T_{cur}; (iv) for each row (i.e. each word) take the product (or sum of log probabilities) of all the column features thus triggered in steps (ii) and (iii); (v) sort the words in the lexicon by their probability from (iv) and have the top N fill the beam of size N.

Once the beam is thus populated, we use the DS grammar to parse each word in the beam in turn; upon success, that is, if the word is parsable, and the resulting semantics subsumes the goal concept, R_g, we move on to generate the next word incrementally until we reach the goal concept, that is, until $R_g \sqsubseteq R_{cur} \wedge R_{cur} \sqsubseteq R_g$.

Repair mechanism The DS repair mechanism, i.e. that of backtrack and parse / generate (see above Sec. 2.3), is triggered when none of the words in the beam successfully generate; either because neither are parsable, or else their resulting semantics don't subsume R_g (because it may have been revised). When triggered, the model backtracks over the context DAG path (see above), and, following the same inference process, attempts to (re-)populate the beam and generate from there. Backtracking continues until generation is successful, with the model having generated the interregnum (e.g. "I mean", "sorry I mean", "uh", "no", etc.) before it generates the first repair word.

Generation continues normally from that point until the (potentially new) goal concept is reached.

4 Evaluation

4.1 Data

The data to train and test our model comes from the Eve section of the CHILDES corpus (MacWhinney, 2000). This section was annotated with logical forms (LF) by Kwiatkowski et al. (2012). The LFs were then converted to TTR record types (RT) by Eshghi et al. (2013). This dataset consists of utterances towards children from parents, therefore the sentences have a relatively simple structure than adult language. We will use it in the shape of ⟨Utterance, Goal Concept⟩ pairs to train and test our model.

For training our generator, we test-parsed the dataset using two versions of the grammar learned by Eshghi et al. (2013): the grammar containing the top 1 hypothesis and the grammar containing the top 3. This resulted in two subsets of the data that could be parsed and in which the produced RT semantics matched the gold semantics exactly. Let's call these `top-1` and `top-3` respectively. We report their relevant statistics in Table 1.

dataset	total samples	total words	mode length	max length	type / token ratio
top-1	729	2152	3	7	18.08
top-3	1361	4194	3	7	21.96

Table 1: Filtered Dataset Statistics

However, even as the top-3 grammar from Eshghi et al. (2013) gives wider parsing coverage, it included many erroneously learned lexical actions. We therefore decided to carry out our experiments below on the `top-1` dataset filtered using the top-1 grammar. This is at the expense of not generating sentences that we'd otherwise be able to generate since the overall distribution of the two datasets are similar. Therefore, the results we report below are more conservative (i.e. lower) than those we'd have been able to achieve if we'd manually cleaned up the top-3 grammar and applied it to learning and generation.

4.2 Model Evaluation

We evaluate our generation model on the `top-1` set in two ways: (i) standard evaluation of generation without any mid-generation revisions to the goal; (ii) we evaluate the capability of the same model to generalise to cases where the goal concept is revised mid-generation, i.e. to cases where the model needs to produce *self-repairs*.

Standard evaluation For this, we report percentage of exact match (EM), ROUGE-1, Rouge-2, and ROUGE-l between the gold sentences in the dataset and the output sentences from the model. On the training set, we could observe that out of 656 training samples, we can generate 597 utterances (91.01%) whose semantics exactly matches the generation goal concept; 416 of these fully match the gold sentence, yielding an EM score of 0.6341 (meaning 63.41% of the output sentences fully match the gold sentences). For the test set, out of 73 total samples, 64 sentences were generated fully to the goal concept (87.67%), and 46 of these (63.01%) completely matched the gold sentence in the dataset. Among the outputs not fully match by the gold sentences a large portion of them are very close to an exact match. For example the generated sample, "what is that", where the gold sentence is "what's that": such samples were not counted initially among the exact matches. We then took these to be exact matches and recomputed evaluation scores. The final results are summarised in Table. 2.

	EM	ROUGE-1	ROUGE-2	ROUGE-l
Train	0.84	0.94	0.71	0.92
Test	0.78	0.88	0.67	0.86

Table 2: Evaluation results for generation without any goal concept revisions

4.3 Generating self-repairs: a zero-shot evaluation

To evaluate the ability of the model to generate self-repairs in a zero shot setting, we generate a dataset of *semantic revisions* to the goal concept using the original `top-1` data. We use the Stanford POS tagger to automatically generate a set of revisions, where each revision is a tuple, $\langle R_g, index, R_r, Utt_r, forward\rangle$: R_g: is the original goal concept; *index*: is the position along the generation path where the revision takes place; R_r: is the revised goal; Utt_r: is the result of replacing a single word in the *original* gold utterance with a word from our data of the same POS – R_r now corresponds to the (revised goal) semantics of Utt_r; and, finally: $forward$: is either true or false, marking whether the revised semantic material has already been contributed before *index* or not; if true, we would expect a *forward*-

looking self-repair, and otherwise a *backward-looking* one (see Sec. 2.3 above). We derive these revision tuples for every utterance in the dataset with length greater than 4, and on the following Parts of Speech: {NOUN, ADJ, PROPN, ADP, ADV}. These tuples therefore give us 4 experimental conditions, across two binary factors: (i) locality: is the point at which the revision is made strictly local to the repairandum; or does it have a distance of more than 1; (ii) Is the revision after or before the corresponding semantic contributions have been made?

We then run the revisions through the model and evaluate the output automatically as follows: we use a simple rule-based algorithm to 'clean out' the self-repair from the model output, and compare this to the revised utterance, Utt_r. For this comparison, we only report EM – see Table 3. We observed 641 of the generatable revisions in total are an exact match.

	forward-looking	backward-looking
local	0.93	0.89
distant	0.73	0.82

Table 3: EM for zero-shot evaluation of repairs

Since we do not have gold data for self-repairs, we did a small human evaluation on the model output: the authors each independently annotated a subset of 30 examples, assigning scores on a Likert scale from 1 to 3 for: (a) grammaticality of the self-repairs; and (b) their human-likness or naturalness, which initially led to a low agreement. They then met to discuss the disagreements in order to iron out the differences between the criteria they had applied. They then continued to annotate 70 additional system outputs. This led to a Krippendorff's alpha score of 0.88 for grammaticality and 0.82 for naturalness, demonstrating very high agreement. To then report the average scores given by the human annotators, the lower score was chosen when there was a disagreement, resulting in 2.72 and 2.28 mean scores for grammaticality and naturalness respectively, confirming the quality of the generated output.

5 Discussion

During the error analysis we observed the following error patterns: In the standard evaluation of generation, there were 199 instances where the model had fully generated to the goal concept, while the generated output did not match the gold utterance. Many were cases where the model had generated a statement instead of a question or vice versa (e.g. "I may see them" is generated over "may I see them"). In a few cases, the generated output was ungrammatical with the wrong word order: both of these are caused by the original grammar from Eshghi et al. (2013) overgenerating – this is acknowledged by the authors, and it is due to the fact that their induced grammar did not capture the full set of syntactic constraints present in their data. This is in turn because they were only conditioning their search on the type of the pointed node, like we do here. Inducing the full set of syntactic constraints was left to future work, as it is here.

5.1 Limitations

Our evaluation in this paper has at least two important **limitations**:

(1) We evaluate our incremental generation model on a small, and relatively simple dataset (leading to high ROUGE scores because of the little variation in data and relative similarity between training and testing sets) due to the fact that we currently do not have access to a wider coverage grammar. However, this was a conscious choice on the authors' part: we used a learned grammar to induce our probabilistic generation model and evaluated it on exactly the same dataset from which the grammar was learned (Eshghi et al., 2013). This was deemed to be methodologically both sounder and cleaner than, say, use of a manually constructed grammar. We also believe that the probabilistic model we have contributed here will generalise to larger, more complex datasets when wider-coverage grammars becomes available. We leave this for future work.

(2) Perhaps more importantly, we have no comparative evaluation, and this in a climate where neural NLG has seen astonishing advances in the work on Transformer-based (large) Language Models. To carry out this comparative evaluation, we need to integrate our model with a downstream, and, ideally, multimodal dialogue task (see e.g. Yu et al. (2016, 2017) for how DS-TTR can be integrated within a visually grounded task). This requires substantial further work which is our next step.

5.2 Why a grammar-based approach?

It might reasonably be asked why we are using a grammar-based approach in the age of Large Language Models (LLM) such as GPT-4 and a large number of other, open source models following. These models are astonishing few-shot learners, and have recently achieved great successes that few thought possible (e.g. in open-domain dialogue, conversational question answering, essay writing, summerisation, translation etc), and are changing the human world in ways that we have not yet had time to grasp.

Nevertheless, for the moment, the fact remains that: (a) these models are extremely costly to train and run due their sheer size and the amount of resources (data, compute power, energy) needed to train them; it's also been demonstrated, time and again, that they have poor compositional generalisation properties (see Pantazopoulos et al. (2022); Nikolaus et al. (2019) among others), which explains much of their characteristic data inefficiency; (b) they are very difficult to *control* and/or adapt while often producing factually incorrect statements, commonly referred to as hallucinations (Rashkin et al., 2021; Dziri et al., 2022) using very convincing language – this extends to confident prediction of erroneous actions or plans in multi-modal, embodied settings; (d) they are very hard to sufficiently *verify*, making them unsuitable for use in safety-critical domains such as healthcare; (e) particularly important for us here, unlike recurrent models such as RNNs and LSTMs, standard Transformer-based neural architectures (Vaswani et al., 2017) are not properly incremental – even the auto-regressive variants such as GPT – in the sense that they process word sequences as whole, rather than word by word; they can be run under an 'incremental interface' (Madureira and Schlangen, 2020; Rohanian and Hough, 2021) where input is reprocessed from the beginning with every new token, but even then, they exhibit poor incremental performance with unstable output compared to e.g. LSTMs (Madureira and Schlangen, 2020). Interesting recent work has explored using Linear Transformers (Katharopoulos et al., 2020) with recurrent memory to properly incrementalise LMs (Kahardipraja et al., 2021a), but this work is as yet in its infancy, and we do not yet know of any work that integrates LMs end to end within a real-time, incremental dialogue system.

On the other hand, grammar-based approaches have the advantage of being highly controllable and transparent; but crucially, they incorporate the very large wealth of linguistic knowledge that has arisen from decades of linguistics and semantics research. This knowledge has been demonstrated to be a very effective source of inductive bias in grammar-based models which in turn translates to remarkable generalisation potential, and thus also data efficiency (see e.g. Mao et al. (2021) for a CCG-based multi-modal model, and Eshghi et al. (2017) for a DS-TTR-based one) – see Eshghi et al. (2022) for an extended discussion. One common criticism is that grammar-based models are brittle. This is often true, but we do not believe this to be a fundamental property, and think that specific grammars of a language are adaptable and learnable from interaction. But much work remains to be done to demonstrate this property.

For these reasons, we believe that grammar-based approaches hold promises that are as yet unfulfilled, and are therefore still worth exploring in parallel to the much needed work on making LM architectures and training regimes more incremental (see Kahardipraja et al. (2021b, 2023)).

6 Conclusion

We developed the first semantic, probabilistic model of real-time language generation using the Dynamic Syntax framework. The results show that the model performs well, even though we evaluated it only on a small dataset. We also demonstrated the zero-shot generalisation ability of the model to generate self-repairs where none were observed during training. To our knowledge, this is the first model capable of reacting to real-time changes to the generation goal by generating suitable self-corrections. This ability is essential in dialogue systems in highly dynamic contexts or environments. Our generation model can be seamlessly integrated into incremental dialogue system architectures (e.g. based on Schlangen and Skantze (2009)). This work further highlights the generalisation power of grammar-based approaches, and lays the foundations for creating conversational AI systems that are controllable, data-efficient, more naturally interactive, and more accessible to people with cognitive impairments.

Acknowledgements

We are very grateful to Tanvi Dinkar and Julian Hough for some of the ideas in this paper and subsequent discussion. We would also like to thank the SemDial reviewers whose constructive critique lead to further changes and elaboration.

References

Angus Addlesee and Marco Damonte. 2023. Understanding and answering incomplete questions. In *Proceedings of the 5th Conference on Conversational User Interfaces*.

Angus Addlesee and Arash Eshghi. 2021. Incremental graph-based semantics and reasoning for conversational AI. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 1–7, Gothenburg, Sweden. Association for Computational Linguistics.

Angus Addlesee, Ioannis Konstas, and Arash Eshghi. 2019. Current challenges in spoken dialogue systems and why they are critical for healthcare applications.

G.S. Aist, J. Allen, E. Campana, L. Galescu, C.A. Gomez Gallo, S. Stoness, M. Swift, and M. Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh.

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

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

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics, pages 271–323. North Holland.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2013. Probabilistic Type Theory and Natural Language Semantics. *Unpublished Manuscript, University of Gothenburg and King's College London*.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*, Gothenburg, Sweden. Association for Computational Linguistics.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. *The Handbook of Contemporary Semantic Theory*, pages 375–407.

Matthew Crocker, Martin Pickering, and Charles Clifton, editors. 2000. *Architectures and Mechanisms in Sentence Comprehension*. Cambridge University Press.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing (CSLP?Äô12)*, pages 51–63.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguisitics.

A. Eshghi, M. Purver, and Julian Hough. 2011. Dylan: Parser for dynamic syntax. Technical report, Queen Mary University of London.

Arash Eshghi, Eleni Gregoromichelaki, and Christine Howes. 2022. *Action Coordination and Learning in Dialogue*. CSLI.

Arash Eshghi, Julian Hough, and Matthew Purver. 2013. Incremental grammar induction from child-directed dialogue utterances. In *Proceedings of the 4th Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 94–103, Sofia, Bulgaria. Association for Computational Linguistics.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.

Arash Eshghi and Oliver Lemon. 2014. How domain-general can we be? Learning incremental dialogue systems without dialogue acts. In *Proceedings of Semdial 2014 (DialWatt)*.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017. Bootstrapping incremental dialogue systems from minimal data: linguistic knowledge or machine learning? In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Victor Ferreira. 1996. Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, 35:724–755.

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

C. Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York.

Eleni Gregoromichelaki, Ronnie Cann, and Ruth Kempson. 2012. Language as tools for interaction: Grammar and the dynamics of ellipsis resolution. *The Linguistic Review*, 29(4):563–584.

Eleni Gregoromichelaki, Gregory James Mills, Christine Howes, Arash Eshghi, Stergios Chatzikyriakidis, Matthew Purver, Ruth Kempson, Ronnie Cann, and Patrick G. T. Healey. 2020. Completability vs (in)completeness. *Acta Linguistica Hafniensia*, 52(2):260–284.

Markus Guhe. 2007. *Incremental Conceptualization for Language Production*. NJ: Lawrence Erlbaum Associates.

P. G. T. Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, Poitiers.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Nordic Prosody: Proceedings of XIth Conference, Tartu 2012*, pages 137–146.

Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Julian Hough and Matthew Purver. 2012. Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the 16th SemDial Workshop on the Semantics and Pragmatics of Dialogue (SeineDial)*, pages 136–144, Paris, France.

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.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2021a. Towards incremental transformers: An empirical analysis of transformer models for incremental nlu. pages 1178–1189. Association for Computational Linguistics.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2021b. Towards incremental transformers: An empirical analysis of transformer models for incremental NLU. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1178–1189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2023. TAPIR: Learning adaptive revision for incremental natural language understanding with a two-pass model. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4173–4197, Toronto, Canada. Association for Computational Linguistics.

Dimitrios Kalatzis, Arash Eshghi, and Oliver Lemon. 2016. Bootstrapping incremental dialogue systems: using linguistic knowledge to learn from minimal data. In *Proceedings of the NIPS 2016 workshop on Learning Methods for Dialogue*, Barcelona.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention.

Ruth Kempson, Ronnie Cann, Arash Eshghi, Eleni Gregoromichelaki, and Matthew Purver. 2015. Ellipsis. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantics*. Wiley-Blackwell.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikiriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–275.

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

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: A Toolkit for Incremental Situated Processing. In *Proceedings of SIGdial 2014: Short Papers*, pages 84–88.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings.

Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. *Proc. of Londial, SemDial Workshop*, pages 83–90.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of logic and computation*.

W.J.M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, third edition. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental nlu.

Jiayuan Mao, Freda H. Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. 2021. Grammar-based grounded lexicon learning. In *Advances in Neural Information Processing Systems*.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimers dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, 3.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

George Pantazopoulos, Alessandro Suglia, and Arash Eshghi. 2022. Combine to describe: Evaluating compositional generalization in image captioning. pages 115–131. Association for Computational Linguistics.

Massimo Poesio and Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1:1–89.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the 'I's and crossing the 'You's: Context, speech acts and grammar. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, Poznań. Polish Society for Cognitive Science.

Matthew Purver, Christine Howes, Eleni Gregoromichelaki, and Patrick G. T. Healey. 2009. Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271, London, UK. Association for Computational Linguistics.

Matthew Purver and Ruth Kempson. 2004. Incremental context-based generation for dialogue. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG04)*, number 3123 in Lecture Notes in Artifical Intelligence, pages 151–160, Brockenhurst, UK. Springer.

Matthew Purver, Mehrnoosh Sadrzadeh, Ruth Kempson, Gijs Wijnholds, and Julian Hough. 2021. Incremental composition in distributional semantics. *Journal of Logic, Language and Information*, 30(2):379–406.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718, Online. Association for Computational Linguistics.

Morteza Rohanian and Julian Hough. 2021. Reframing incremental deep language models for dialogue processing with multi-task learning.

Yo Sato. 2011. Local ambiguity, search strategies and parsing in Dynamic Syntax. In E. Gregoromichelaki, R. Kempson, and C. Howes, editors, *The Dynamics of Lexical Interfaces*. CSLI Publications.

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.

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.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 1–8, Tokyo, Japan. Association for Computational Linguistics.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 745–753, Athens, Greece. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. volume 2017-December.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of SIGDIAL 2016, 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 339–349, Los Angeles.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2017. Learning how to learn: an adaptive dialogue agent for incrementally learning visually grounded word meanings. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 1–10. ACL.

Evaluating Automatic Speech Recognition and Natural Language Understanding in an Incremental Setting

Ryan Whetten
ryanwhetten@u.boisestate.edu
Computer Science
Boise State University

Enoch Levandovsky
enochlevandovsky@u.boisestate.edu
Computer Science
Boise State University

Mir Tahsin Imtiaz
tahsinimtiaz@u.boisestate.edu
Computer Science
Boise State University

Casey Kennington
caseykennington boisestate.edu
Computer Science
Boise State University

Abstract

Spoken dialogue systems enable people to interact with machines using speech, many of which involve the use of automatic speech recognition and language understanding in order to react to and determine a decision about how to respond. Unlike humans, many systems operate on complete sentences, waiting for a length of silence before attempting to process the input. In contrast, incremental spoken dialogue systems enable faster and more natural interaction by operating at a more fine-grained level. In this work, we evaluate six speech recognizers and RASA for language understanding in an incremental spoken dialogue system. The results suggest that, for speech recognition, online/cloud models can be slower and less stable than local models and we show that incremental language understanding can enable a system to make decisions earlier than waiting for the end of the utterance.

1 Introduction

Interacting with technology using a spoken dialogue system (SDS) has become more prevalent with applications such as voice search, dictation, and virtual assistants (Yu and Deng, 2016). A fundamental step in how these systems process input, whether implemented in a chatbot, on a website, or on a robot, is to understand what is uttered by the user and produce some kind of action, often by responding using speech back to the user. This is usually performed by first transcribing what the user says using *Automatic Speech Recognition* (ASR), followed by using a model of *Natural Language Understanding* (NLU) to map from the ASR's transcript to a computable abstraction, often a semantic frame. Existing models for NLU, including large language models, are becoming more commonplace, but most have an important drawback: they operate on complete sentences.

Incremental systems, in contrast, operate at more fine-grained levels of information, usually at the word-level instead of the sentence-level, and begin to process the input as soon as it is received. Incremental systems have been shown to offer a more natural interaction (Aist et al., 2007; Edlund et al., 2008) likely due to the fact that humans also produce and understand language incrementally (Tanenhaus and Spivey-Knowlton, 1995). However, most existing ASR and NLU models are either non-incremental or have not been evaluated incrementally. With incremental systems offering more natural interactions, it is crucial to evaluate and understand how ASR and NLU perform in an incremental setting.

In the spirit of prior work, which evaluated several existing ASR models and their relationship to NLU to inform the research community (Morbini et al., 2013), in this work, we evaluate six ASR models (two online/cloud and four local). However, in this work, we experiment in an incremental SDS setting. We evaluate on two English datasets using incremental metrics proposed from Baumann et al. (2009, 2016), as well as propose a new metric *Revokes per Second* to observe how frequently the predictions of an ASR model change (section 3.1.1). Moreover, we incrementalize a recent version of RASA, a framework for NLU and building conversational agents, and evaluate its incremental performance on the SNIPS and SLURP datasets (Coucke et al., 2018; Bastianelli et al., 2020a) in conjunction with an ASR model.

Results show that cloud ASRs, although being some of the most accurate, can have a higher latency and change predictions more frequently than the local ASRs. For incremental NLU, results show that even without a perfect transcript (i.e. a transcript generated by an ASR instead of the ground-truth), a system could be ready to take an action up to six words on average before the end of an utterance. The results provide insights when considering which ASR to use and for designing SDSs that are more natural and responsive in their in-

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

teractions. All of the models are implemented as modules in the Retico framework (Michael, 2020) for ease of use in incremental systems.

2 Background & Related Work

Incremental ASR Many ASR models operate incrementally in that they produce word or sub-word outputs as the recognition unfolds (Morbini et al., 2013). This ability to function incrementally is an important requirement for spoken dialogue systems (SDS), especially ones that are multimodal or part of a robot platform because there is a high expectation of timely interaction from human dialogue partners (Kennington et al., 2020). Although ASR can function incrementally, most ASR models use the word-error-rate (WER) metric for evaluation, even in conversational settings (Morris et al., 2004; Morbini et al., 2013; Georgila et al., 2020). However, WER solely captures the end performance and does not take into account incremental performance and speed. Morbini et al. (2013) mentions the importance of incremental ASR stating, "incremental results allow the system to react while the user is still speaking", yet evaluates ASR performance using only WER. We build on this prior work by using WER as well as metrics to evaluate incremental performance.

Baumann et al. (2009, 2016) proposed metrics for evaluation of incremental performance such as Edit Overhead, Word First Correct Response, Disfluency Gain, and Word Survival Rate. All of the metrics, including WER, can be classified into one of the following three general areas of interest: overall accuracy, stability (which can be thought of as measuring the incremental performance), and speed. However, these metrics focus on discrete word-level output and not the relationship of between incremental performance and speed. To capture the relationship between incremental performance and speed, we propose to measure the number of *Revokes per Second* (introduced in Section 3.1.1).

Incremental NLU NLU maps words onto a meaning representation, such as a semantic frame (see example in Section 3.2). Among many methods for doing this mapping, in this paper, we focus on RASA (Bocklisch et al., 2017) which is open source and has been shown to work well for NLU (Liu et al., 2019). RASA was made to work incrementally in Rafla and Kennington (2019), but subsequent updates to RASA have left the incremental version obsolete and the original evaluation did not include ASR as the evaluation was performed using only text data (i.e. ground-truth transcriptions). In this work, we incrementalize a recent version of RASA that will be more maintainable in the future and we evaluate performance using incrementally produced transcriptions from an ASR as well as the ground-truth transcriptions.

3 Methods

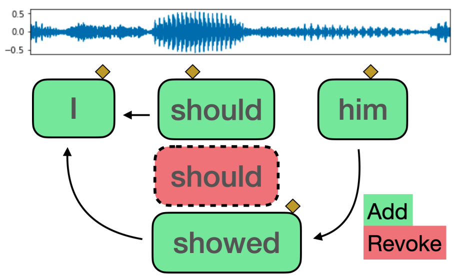


Figure 1: An example of adds and revokes. The word *should* is added, then revoked and replaced by *showed*. The diamonds represent the time when the predictions are made.

The Incremental Unit Framwework We adopt the *Incremental Unit* framework from Schlangen and Skantze (2009) for its flexible design. The framework is built around *incremental units* (IU), discrete pieces of information (e.g., a chunk of audio, a word, an image), that are produced by specific modules. These modules process IUs as input and can pass IUs that they produce to other modules. For example, a microphone module can output chunks of audio as IUs that are passed to an ASR module that outputs individual words as IUs which can in turn be passed on to an NLU module, and so on.

The IU framework has provisions for handling cases where a module's output was found to be in error, given new information. To handle these cases, there are three operations for IUs: `add` (to mark an IU to be added to the output), `revoke` (to mark an IU to be removed from the output), and `commit` (to mark that an IU will not longer change). A perfect ASR would only add new words to the growing list of previously recognized words. But as most ASRs have errors—particularly when they work incrementally—the `revoke` operation allows the ASR module to remove an erroneous IU and replace it (i.e., through another add operation) in the recognized output. Importantly, the `revoke` operation propagates to downstream modules that may have acted on prior input, signalling the error. An example of incremental `add` and `revoke` for ASR is shown in Figure 1. We use Retico, a Python implementation of the IU framework, to implement and evaluate ASR and NLU models (Michael, 2020).

3.1 ASR

We use six different, readily available ASR models: 2 cloud-based and 4 local, chosen due to their respective results and accessibility. The cloud-based models are Google Cloud's Speech-to-Text and Microsoft Azure's Speech. We use Wav2Vec2 (W2V), DeepSpeech (DS), PocketSpinx (PS), and Vosk (Baevski et al., 2020; Hannun et al., 2014; Huggins-Daines et al., 2006).

Due to the limited amount of information given about the online ASR models, we can not go into depth about the architecture and training behind these models. The local models are summarized in Table 1 and described below.

Wav2Vec (W2V): We use Meta's Wav2Vec model from a checkpoint provided by HuggingFace where the model has been pre-trained and fine-turned on 960 hours of Librispeech (Baevski et al., 2020). This architecture is unique in that it is pre-trained on hours of unlabeled raw audio data. While other models first convert the audio into a spectrogram, Wav2Vec operates directly on audio data.[1]

DeepSpeech (DS): Mozilla's DeepSpeech model, is based on work done by Hannun et al. (2014). This architecture uses Recurrent Neural Networks that operate on spectrograms of the audio to make predictions. We use the 0.9.3 model and scorer for predictions. This model was trained using a wider variety of data from Fisher, LibriSpeech, Switchboard, Common Voice English, and approximately 1,700 hours of transcribed WAMU (NPR) radio shows explicitly licensed to them to be used as training corpora.[2]

PocketSphinx (PS): One of the lighter ASRs we tested is CMU's PocketSphinx (Huggins-Daines et al., 2006). PS is a light-weight ASR that is a part of the open source speech recognition tool kit called the CMUSphinx Project. This model was trained on 1,600 utterances from the RM-1 speaker-independent training corpus. Unlike the previously mentioned models, PS does not use neural networks and is instead based on traditional methods of speech recognition by using hidden Markov models, language models, and phonetic dictionaries.[3]

Vosk: Alpha Cephei's Vosk (with the vosk-model-en-us-0.22 model) is built on top of Kaldi (Povey et al., 2011), and like PocketSphinx, uses an acoustic model, language model, and phonetic dictionary. Vosk uses a neural network for the acoustic part of the model.[4]

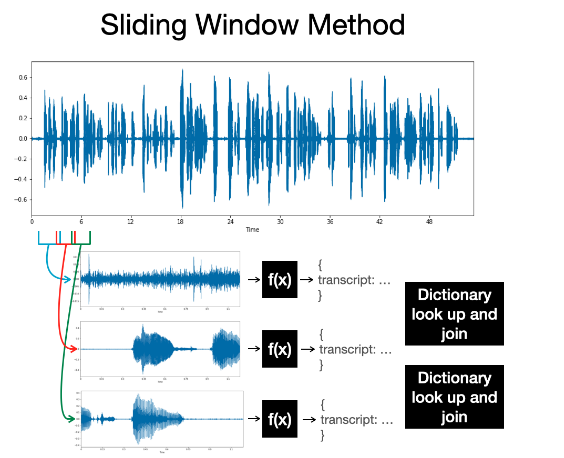


Figure 2: In the Sliding Window method, the ASR model makes predictions on partially overlapping portions of audio. Dictionaries are used to join the incoming predictions together.

3.1.1 ASR Metrics

As mentioned, all previously proposed metrics for evaluating incremental ASR can be divided into three broad categories: overall accuracy (using WER), speed, and stability. In this section, we describe the specific metrics used and introduce our new metric which combines these last two categories of speed and stability into a single metric.

Overall Accuracy: WER Although there are different metrics to measure overall accuracy as compared in (Morris et al., 2004), we only use the most common metric, Word Error Rate (WER), which is defined by the the number of edits, substitutions (S), insertions(I), and deletions (D), divided by the total number of words (N): $WER = \frac{S+I+D}{N}$.

Predictive Speed: Latency In order to measure the general speed of an ASR model, we measure the time it takes from the time the ASR model gets the

[1] https://huggingface.co/facebook/wav2vec2-base-960h
[2] https://deepspeech.readthedocs.io/en/r0.9/
[3] https://github.com/cmusphinx/pocketsphinx-python
[4] https://alphacephei.com/vosk/

Name (abbreviation)	Model	Training Data
Wav2Vec (W2V)	wav2vec2-base-960h	LibriSpeech
DeepSpeech (DS)	0.9.3	Fisher, LibriSpeech, Switchboard, Common Voice English
PocketSphinx (PS)	N/A	1600 utterances from the RM-1
Vosk	en-us-0.22	N/A

Table 1: Local ASR models along with their used models and training data if available.

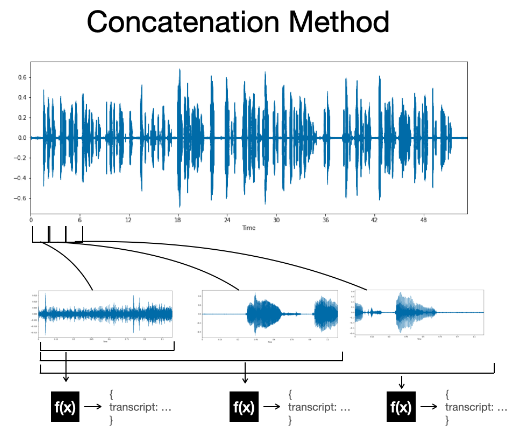


Figure 3: In this method, the incremental audio is concatenated together, and a prediction is made on the entire audio that has been given up to that point.

audio until the prediction is made. We then take this time and divide by the number of words in that particular prediction. With this, we define latency as the average amount of time per word it takes an ASR model to make a prediction: $LAT = \frac{Time}{N}$, where time is measured in seconds and N is the total number words in a given prediction.

Stability: Edit Overhead For measuring stability, we measure the edit overhead (EO). EO is the total number of revokes (R) divided by the total number of edits, or additions (A) and revokes (R), that the ASR model makes. $EO = \frac{R}{A+R}$.

Revokes per Second Our proposed and final metric is the number of *Revokes per Second* (RPS). We propose this metric as a way to capture the relationship between both speed and stability in an interpretable fashion by measuring how often an ASR changes its predictions. In an incremental SDS setting, this is the average number IUs that are labeled as type `revoke` per second. In an online meeting where real-time subtitles are available, this would represent the number of times you could expect a word to change per second in the transcript.

In such settings, a high RPS in a model's incremental predictions could result in confusion in downstream modules in an SDS setting (such the NLU module) or in humans trying to follow an online meeting using the real-time transcript.

We also look at the inverse *Seconds per Revoke* (SPR) as a simple adjustment to this metric to see how many seconds will pass by before one can expect to see a revoke. This SPR value is useful in interpretations when the RPS is low. Taken together, the formulas for these metric are as follows: $RPS = \frac{R}{Time(s)}$ and $SPR = \frac{Time(s)}{R} = \frac{1}{RPS}$

Combining Sub-word Output Both Google and Azure offer incremental ASR results. For these two ASRs, the audio files are sent to the cloud services in chunks, and the service returns a prediction with other meta-information. Google and Azure ASRs handle the concatenation, combining the predictions into a string that grows as the utterance unfolds. For local ASR models, we have control over how the predictions are combined and processed. We apply and compare two methods in this evaluation: Sliding Window and Concatenation.[5]

One limitation of many ASR models is the amount of audio they can process. For longer audio files (> 30 seconds), ASR models will start slow down and even crash. For this reason, we experiment with a sliding window of audio. For this Sliding Window method, we pass the audio from the file in chunks that are a bit longer than one second. These are then concatenated together as an audio buffer and given to an ASR model until it produces a prediction of at least 5 words or when it is indicated that it is the end of that particular audio file. Once a prediction of 5 words is made we remove the first 35% of the audio buffer. This results in a series of predictions on segments of audio. When a prediction is received, it is joined together with previous predictions. Due to overlap in incoming predictions, the way that the predictions are joined together is non-trivial. We used string filtering and matching functionalities to fil-

[5] We used the same PC with a GTX1080TI GPU for the local models.

ter out noise and join predictions appropriately by finding the overlapping string using dictionaries from WordNet and NLTK (Miller, 1995; Bird et al., 2009).

In the audio datasets we use, generally the files are short. Therefore, as a comparison we also implement a more simple Concatenation Method. For the Concatenation method, we present the audio in chunks into an audio buffer in the same manner as the Sliding Window method, except the audio buffer is a concatenation of all the audio (i.e., no audio ever gets removed from the buffer). Essentially, with this method, the ASR model makes a prediction from the very beginning of the file to the most recent audio given to the buffer. This is computationally more expensive and takes more memory because the ASR model has to make predictions on longer pieces of audio as time goes on, but this method eliminates the need for string matching between overlapping predictions. Diagrams showing these two methods can be seen in Figures 2 and 3. We compare these two methods as part of our evaluation.

3.2 NLU

RASA is a NLU framework that is made up of components that work in a sequential pipeline. In RASA, at least three components are usually required: a tokenizer which splits inputs into smaller tokens (usually words), a featurizer that maps words into a vector, and a classifier that maps from vectors to slots, but others can be included.

The output of this classifier becomes a *meaning representation*, which is a semantic *frame* made up of *slots*, with an overarching *intent*. The example below shows how the utterance *I would like a flight from Boston to Berlin* is represented as a semantic frame made up of 3 slots, one being the intent:

`intent`	flight
`source`	Boston
`target`	Berlin

The dialogue designer determines the slot names based on the domain, e.g., `source` for departure airport and `target` for destination airport.

Instead of making each of the individual components in RASA work incrementally, we follow Khouzaimi et al. (2014) by inserting an incremental manager component at the beginning of the RASA pipeline that allows word-level IUs (i.e. word and IU operation type) to be used as input.

Figure 4 shows a typical minimal pipeline for RASA with our incremental manager component added to enable the entire pipeline to process with word-level IUs. This incremental manager component maintains the unfolding utterance by adding each new word (i.e., from an incremental ASR) to a growing utterance prefix, or an incremental cache, that is re-processed at each word (revoked words are removed from the prefix, as needed).[6]

For example, the utterance *from Boston to Berlin* as part of an ongoing dialogue about booking flights is processed word-by-word, but RASA processes each prefix as a separate utterance:

from			
from	Boston		
from	Boston	to	
from	Boston	to	Berlin

Another challenge to incrementalizing RASA is only outputting new updates to the NLU frame. For example, at each word in the above utterance, RASA should only produce the `source:Boston` slot of the frame when the words *from Boston* are uttered, and not again even though the prefix is being reused at each increment. Likewise, the slot `target:Berlin` should only be produced once when the relevant words *to Berlin* are uttered.

For evaluating NLU, we use accuracy and F1 scores at each word increment.

4 Experiment 1: Incremental Evaluation of ASR Models

Data To evaluate, we use datasets from two different domains: LibriSpeech and a recently assembled dialogue dataset of simulated medical conversations (Fareez et al., 2022).[7] The LibriSpeech test-clean dataset contains 5.4 hours of speech from 40 different speakers, 20 male and 20 female. This audio is divided into over 2,600 files with an average of about 20 words per file containing a vocabulary of over 8,100 words. To ensure the audio would work on all of our models, we converted the audio files to WAV files.

[6]This version of incremental processing is called *restart* incremental because it resets the internal model at each word increment. More ideally, we would use a model that could maintain its internal state and work incrementally (known as *update* incremental), but recent language understanding models are not amenable to word-level processing.

[7]We were unable to obtain the Switchboard corpus due to prohibitive costs.

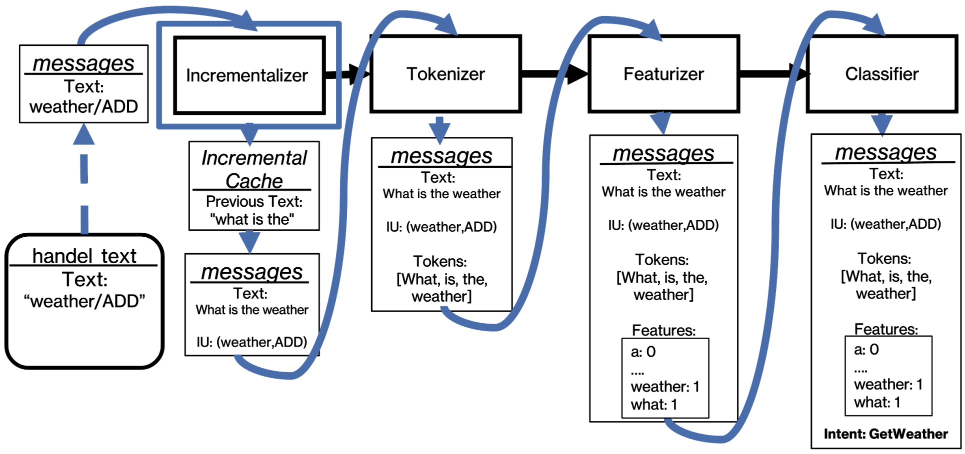


Figure 4: Adapted from rasa.com: our Incrementalizer component at the beginning of any pipeline allows the entire pipeline to process at the word level by managing and caching relevant incremental information.

The medical conversation dataset contains 272 audio files with corresponding transcripts. The purpose of using this dialogue data is 1) to test each model on domain data that presumably none of them have been trained on (since this dataset was just made public in 2022), and 2) to test how each model performs on a dialogue dataset that contains disfluencies such as fillers, corrections, and restarts.

The audio files range from around 7 to 20 minutes in length or about 800 to 2,200 words. Due to the size of these audio files, we split up the files into utterances based on silence and then randomly sample a set of 40 utterances, 17 of which were able to be processed by all 6 ASR models (max 40 seconds, min 0.8 seconds, 6.1 seconds in length on average) due to the length of some of the utterances and the constraints that each model can handle.

Results The results are shown Table 2. When using the Sliding Window method, local models had lower latency (i.e. faster) than both the online models. Some of the local ASR models using the Concatenation method were faster than both of the online ones. However, tests using the Concatenation method was slower and had a higher EO than the Sliding Window method given the same ASR.

Although slower and less stable (as measured by EO), the Concatenated versions performed better than their corresponding Sliding Window version in overall all accuracy or WER. This makes sense as the Concatenation method has access to the entire context to make predictions where as the Sliding Window has only a small portion of the context. Comparing the online models, Google is less accurate and more revoke dependant than Azure. However, Google is considerably quicker which could be crucial in an interactive dialogue setting. The cloud models had surprisingly low latency (though the latency is dependent on the internet speed), but the local ASRs tended to have the lowest latency.

The local ASR model which performed the best in terms of WER was the W2V model using the Concatenation method on the LibriSpeech data and Vosk on the Medical Dialogue data, while the model with the lowest Edit Overhead was the DS model using the Sliding Window method. Though a low WER is generally better, the number of revokes has implications for downstream modules in an SDS; keeping the EO low and Revokes per Second low with a low WER means the model was correct early, which is ideal.

Our results are consistent with previous evaluations on Incremental ASR (Baumann et al., 2016) that show that Google's ASR predictions, although fairly accurate overall, are not as stable as the others, with the highest Edit Overhead of 0.279/0.228 and an average of about 4.5/5.1 Revokes per Second on the LibriSpeech dataset and Medical Dialogue dataset respectively.

The DS and Vosk models' WERs were higher than some of the other models, but the low EO and infrequent number of revokes make them potentially good candidates for an SDS that requires high accuracy as well as low latency and EO, for example in a robotic platform. We suggest Concatenation for live microphones, with voice activity detection or with certain models chunking,[8] to pre-

[8]https://huggingface.co/blog/asr-chunking

Incremental ASR Results on LibriSpeech										
	Google	Azure	W2V	W2V (Con.)	DS	DS (Con.)	PS	PS (Con.)	Vosk	Vosk (Con.)
WER	13.2	9.1	10.6	**4.0**	18.3	8.4	*40.4*	31.8	33.4	6.4
LAT	0.197	0.539	**0.099**	0.127	0.181	*1.443*	0.105	0.220	0.104	0.167
EO	*0.279*	0.065	0.011	0.093	**0.001**	0.013	0.014	0.147	0.072	0.019
R/Sec	*4.564*	0.679	0.141	1.919	**0.008**	0.012	0.178	1.688	0.910	0.143
Sec/R	*0.219*	1.473	7.083	0.521	**123.135**	80.489	5.613	0.593	1.099	7.004

Incremental ASR Results on Medical Dialogue Dataset										
	Google	Azure	W2V	W2V (Con.)	DS	DS (Con.)	PS	PS (Con.)	Vosk	Vosk (Con.)
WER	41.1	**21.0**	47.8	42.3	42.5	38.7	*85.6*	80.0	38.4	23.2
LAT	0.287	0.623	**0.125**	0.217	0.245	*1.452*	0.131	0.394	0.307	1.296
EO	*0.243*	0.055	0.016	0.211	**0.000**	0.014	0.005	0.240	0.048	0.025
R/Sec	*5.944*	0.207	0.253	6.376	**0.000**	0.013	0.046	2.447	0.215	0.079
Sec/R	*0.168*	4.837	3.953	0.157	**inf**	75.616	21.734	0.409	4.649	12.733

Table 2: Summary of results. Local ASRs had lower latency than cloud-based ASRs. The Concatenation method, shown in the columns that contain a *(Con.)*, had higher latency and resulted in a higher EO and RPS, but not as many revokes as the online ASRs. *inf* means zero revokes per second.

vent running out of memory because it is more accurate and does not require string matching.

5 Experiment 2: Evaluation of Incremental RASA

In this section, we explain our experiment to evaluate our incremental version of RASA NLU.

Task & Procedure For this experiment, we were restricted to only use datasets that contain audio, text transcriptions, and annotated frames. Since SNIPS (Coucke et al., 2018) and SLURP (Bastianelli et al., 2020b) datasets have these requirements, they are used for evaluation in this experiment. Both datasets have speech (though in the case of SNIPS, the audio is synthesized). We compare incremental results using oracle (i.e., hand-transcribed) speech from the two datasets as well as ASR output from Google ASR (which had good WER in Experiment 1 and has low latency, but high edit overhead which is desired so RASA has a chance to handle revokes).

The SNIPS dataset contains 14,484 entries with seven categories of intents evenly distributed. There is an average of 9 words per utterance with 66,500 entity annotations with the largest entity representing 8.2% of the annotations. The SLURP dataset consists of 14,488 utterances with 18 intents unevenly distributed with the largest intent representing 14.4% of the data. There is an average of 7 words per utterance with 21,662 entity annotations with he largest entity representing 14.9% of the annotations.

Metrics & Baseline We calculate F1 score of the intent and slots (termed *entities* in RASA; a false positive is when a slot is filled erroneously, and a false negative is when a slot is unfilled). The F1 score at the end of an utterance is the highest possible because at that point it has received all of the audio information. To show how well RASA works incrementally, we show F1 score at the end of the utterance along with show how the F1 score is affected when incrementally removing up to 7 frames/words before the end of the utterance.

We report the F1 score for both hand-transcriptions and ASR output for both datasets. For intent detection, the majority classifier baseline for SNIPS is 14.3% and for SLURPS, 14.4% Similarly the majority classifier baseline for entity detection for SNIPS is 8.2% and for SLURPS entities, is 14.9%. Here we are not attempting to evaluate the RASA model to achieve state-of-the-art results, rather we are trying to evaluate the potential for RASA as an incremental NLU.

Results Figure 5 shows the results for this experiment. Naturally, the closer the utterance was to the end of the utterance, the higher the F1 score. Furthermore, the F1 scores for intent (a single classification of the entire utterance) was higher than the F1 scores for the entities. This too is expected as the single classification task of detecting intent from either seven (SNIPS) or 18 intents (SLURPS), is much simpler than detecting multiple entities from a wider range of categories.

Results moreover show that even when the transcripts are not 100% correct (i.e. come from an ASR, the solid blue line in figure 5), RASA can achieve a higher F1 score than the majority classifier as early as 6 words out for the more difficult task of entities detection. For intent detection, RASA performs significantly better than the

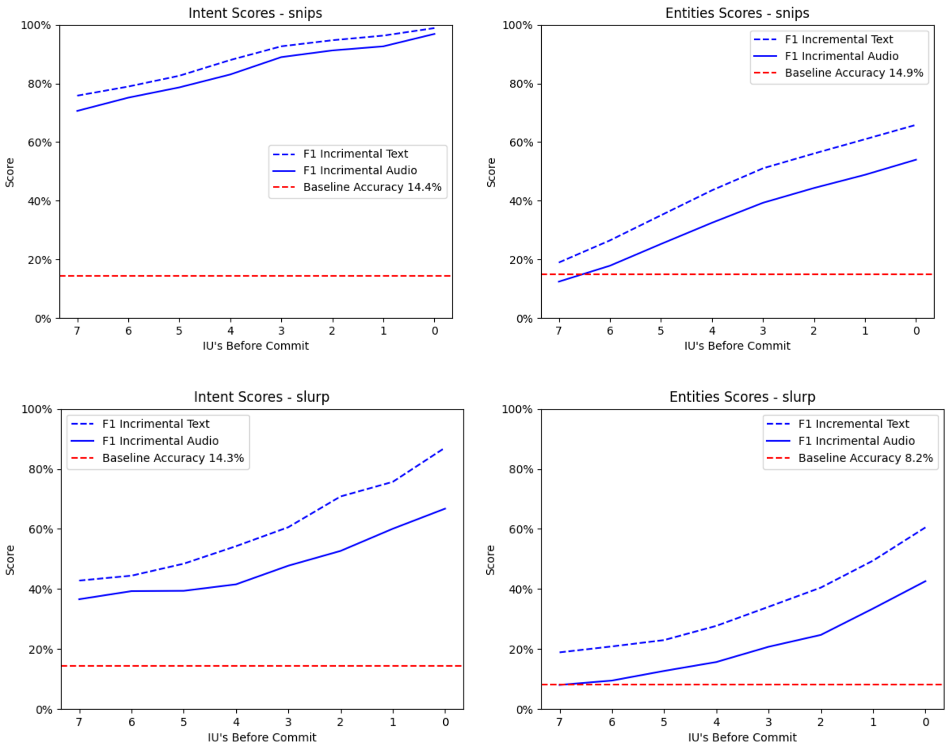


Figure 5: TOP: Incremental results on the SNIPS dataset: transcriptions and ASR. The y-axis is the F1 score, the x-axis is the number of words before the end of the utterance (i.e., before commit). BOTTOM: Incremental results on the SLURP dataset: transcriptions and ASR. The y-axis is the F1 score, the x-axis is the number of words before the end of the utterance.

baseline very quickly into an utterance and on the SNIPS dataset, with imperfect transcripts from the ASR, RASA achieves an F1 score over 80% as early as three words before the end of the utterance.

This evaluation shows the potential for RASA to be used effectively in an incremental setting, allowing a system that uses this incremental setup to be able to make decisions, start acting, or formulating queries before the end of an utterance. This is agreement with with Manuvinakurike et al. (2018) who showed that incremental NLU can be more efficient. For example, an utterance such as *go to the right to pick up...*, a robot could start moving in a predicted direction before the robot even 'knows' that it is to *pick up* and before it 'knows' what to pick up. In the setting of booking an airline or flight, the words *I would like to book...*, the SDS could already begin to start formulating the query to check availability before the end of the sentence.

6 Conclusion

In this work, we tested six different ASR models and RASA for NLU in an incremental setting and we proposed a new metric for incremental ASR, Revokes per Second as an informative addition to existing incremental metrics. We showed that, generally, as might be expected, online ASR (in our evaluation, Google Cloud and Azure cloud services) is not as fast as most of the local ASR models tested, and while the online ASRs are some of the most accurate ASRs we tested, they both have a relatively high number of Revokes per Second and Edit Overhead which, in combination with the latency, could potentially lead to more issues in an incremental setting because high edit rates could require unnecessary processing. Our results are informative as to the *out of the box* performance. Furthermore, we also believe that our proposed metric, Revokes per Second, is an interpretable useful metric that should be used as ASR becomes more prevalent in *live* settings such as in Spoken Dialogue Systems on robots or in live captioning in online meetings.

For NLU, we showed that RASA can work well incrementally, offering designers and users earlier than end-of-utterance predictions of user utterances. This will enable systems to have the option to make earlier decisions and actions, and our changes will be beneficial for long-term maintenance.

Each of the modules described in this paper are implemented in Retico and will be made public.

Acknowledgments

We deeply appreciate the useful feedback from the anonymous reviewers.

References

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. 2007. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, volume 1, pages 149–154, Trento, Italy.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020a. SLURP: A Spoken Language Understanding Resource Package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020b. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7252–7262. Association for Computational Linguistics.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and improving the performance of speech recognition for incremental systems. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 380–388.

Timo Baumann, Casey Redd Kennington, J. Hough, and David Schlangen. 2016. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *IWSDS*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *Proceedings of the 31st Conference on Neural Information Processing Systems*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, pages 12–16.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Commun.*, 50(8-9):630–645.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):1–7.

Kallirroi Georgila, Anton Leuski, Volodymyr Yanov, and David Traum. 2020. Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6469–6476, Marseille, France. European Language Resources Association.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alexander I Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. 2020. rrSDS: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 132–135, 1st virtual meeting. Association for Computational Linguistics.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. 2014. An easy method to make dialogue systems incremental. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 98–107, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv*.

Ramesh Manuvinakurike, Trung Bui, Walter Chang, and Kallirroi Georgila. 2018. Conversational image editing: Incremental intent identification in a new dialogue task. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–295, Melbourne, Australia. Association for Computational Linguistics.

Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, and Ron Artstein. 2013. Which ASR should I choose for my dialogue system? In *Proceedings of the SIGDIAL 2013 Conference*, pages 394–403, Metz, France. Association for Computational Linguistics.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Andrew Rafla and Casey Kennington. 2019. Incrementalizing RASA's Open-Source natural language understanding pipeline. *arXiv*.

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.

Michael K Tanenhaus and Michael J Spivey-Knowlton. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

First users' interactions with voice-controlled virtual assistants: A micro-longitudinal corpus study

Mathias Barthel
Institute for the German Language
Mannheim, Germany
barthel@ids-mannheim.de

Henrike Helmer
Institute for the German Language
Mannheim, Germany
helmer@ids-mannheim.de

Silke Reineke
Institute for the German Language
Mannheim, Germany
reineke@ids-mannheim.de

Abstract

We present a collection of (currently) about 5.500 commands directed to voice-controlled virtual assistants (VAs) by sixteen initial users of a VA system in their homes. The collection comprises recordings captured by the VA itself and with a conditional voice recorder (CVR) selectively capturing recordings including the VA-directed commands plus some surrounding context. Next to a description of the collection, we present initial findings on the patterns of use of the VA systems during the first weeks after installation, including usage timing, the development of usage frequency, distributions of sentence structures across commands, and (the development of) command success rates. We discuss the advantages and disadvantages of the applied collection-specific recording approach and describe potential research questions that can be investigated in the future, based on the collection, as well as the merit of combining quantitative corpus linguistic approaches with qualitative in-depth analyses of single cases.

1 Introduction

Human-computer interaction becomes increasingly more prevalent in industrialised societies. More recently, especially interactions with in-home intelligent virtual assistants (VAs) quickly grows in popularity and amount of use. While research on human interaction with technology moves more into the focus of the language sciences lately, it was established early with Suchman's (1987) seminal work on situated practices in the usage of "intelligent" machines (at that time a printer). Since then, ethnomethodological and conversation analytic (CA) research has addressed a variety of phenomena regarding the interaction between humans and AI-based technologies (for a comprehensive overview of studies see Mlynář et al. (in prep.)). CA-related studies (and studies interested in conversation analytic concepts), especially when investigating interaction with verbally controllable technology (voice-based virtual assistants, robots, chatbots etc.), have examined the organization of talk, like openings and closings (e.g., Pitsch et al. (2009)) and turn-taking in dyadic and multi-party interaction (Skantze, 2021), as well as on miscommunication and repair sequences (e.g., Krummheuer (2008); Pelikan and Broth (2016)).

Recent studies on interaction with VAs like *Amazon*'s *Echo Dot* or *Alphabet*'s *Google Home* have shown how VA systems are designed to help users diagnose and repair trouble (e.g. by rephrasing requests or asking clarifying questions (see Porcheron et al. (2018); Reeves et al. (2018)). Previous research also touched upon the question of how VAs are embedded in multiple ongoing activities in private settings (Porcheron et al., 2018), how reactions of VAs have effects on the progressivity in interaction (Fischer et al., 2019), how the integration of systems into everyday practices is connected to agency (Habscheid et al., 2023), and how a machine's 'participation' can be seen as situational and regulatory participation which becomes part of meaningful talk-in-interaction (Reeves and Porcheron, 2022).

While VAs are claimed to be designed to more and more resemble human interlocutors in their verbal behaviour, they still fall short of human-like interactional capacities in many tasks and on many occasions. Users however do not apply social rules 'mindlessly' onto VAs (Reeves and Porcheron, 2022). They adapt their talk in order to improve interaction with a VA (Pelikan and Broth, 2016), e.g. by altering prosody or rephrasing instructions (Porcheron et al., 2018), and they learn how to formulate probably successful commands (Reeves et al., 2018). Learning to efficiently deal with these weaknesses thus becomes a task of human users.

First users of a VA system hence need to learn the peculiarities of the system to be able to achieve successful goal-oriented interaction with the VA.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

Studying such adaptations to systems, CA-related research has hitherto mainly used single-case analyses only, typically focusing on specific moments of trouble. To systematically analyse and understand in what ways users adapt their use of VAs to the capabilities and limitations of the system, how they learn which strategies turn out to be successful, and which are the overarching longer-term patterns of use, we need to collect data of human-VA interaction over time and analyse them from a micro-longitudinal perspective. In our project, we aim at addressing this desideratum and adopt a mixed-methods approach that combines conversation analysis and interactional linguistics with quantitative analysis. Our overarching goal is a micro-longitudinal analysis of first users' adaption in interacting with the VA. The focus of this paper is on a quantitative overview of developments over time with regard to the timing dynamics of commands, their linguistic structures, and their success-rate.

The methods of data collection are described in section 2. Section 3 presents a description of the resulting collection of audio recordings, as well as a number of first exemplary findings. Finally, section 4 will offer a discussion of the achievements and downsides of the presented methods of data collection and processing, and give an outlook on future use cases for the collected data and the kinds of questions that can be investigated on their basis.

2 Methods

2.1 Participants

To be able to draw a picture of how humans use a VA in a natural setting, and in line with conversation analytic methodology, we recorded naturally occurring interactions of human users with a VA, focusing on recording human-VA interactions with ecological validity. We recorded first users' authentic interactions with VAs in their private living environments during their first weeks of using the VA. We searched for participants who had an a priori interest to get a VA system for their homes and asked them if they would be willing to participate in our study over a period of several weeks. We only included users who had no significant prior experience in using a VA system, so they are all novices in the field of VA communication. Participants got a small monetary compensation for their participation in the study and could keep the VA system after the end of the recording period. We obtained all participants' advanced informed written consent to use the recordings and VA log-files they provided for the purpose of the project. To date, we recorded six single participants or participating families with two to four members (mean age = 20 years, min = 3 years, max = 37 years) over a period of seven to ten weeks (mean = 66 days, min = 49 days, max = 72 days), starting from the first day of their usage of the VA.[1] This way, we were able to track potential changes over time in participants' usage behaviour and formulations of commands during the initial phase of interacting with their newly installed VA system.

2.2 Recording Methods

For data collection, a new VA system (*Amazon Alexa EchoDot*) was installed together with the participants in their home, either in the kitchen or in the living room. Additionally, a conditional voice recorder (CVR)[2] was placed in close proximity to the VA speaker for the recording period. The CVR is a device developed and previously used by Martin Porcheron (see Porcheron et al. (2018)) that captures audio snippets containing commands to the VA. The CVR-software uses a speech detection model[3] and is installed on a Raspberry Pi supplied with a conference microphone. We replicated the CVR and adapted it for our purposes.[4] Our version of the CVR continuously recorded 90-second stretches of audio, constantly overwriting these 90 seconds in a loop. Upon detecting the wake-word (*"Alexa"*), the CVR would save the last 90 seconds of recording and attach the following 90 seconds of recording to the file, creating 3-minute long audio snippets around each user command to the VA. This way, we were able to record the context in which users addressed the VA, the commands to the VA themselves, as well as the reactions by the VA plus potential follow-up context.[5] Whenever

[1] One additional household was excluded from data analysis in this study due do data scarcity, as the participants made use of the VA only in 11 days, producing only 81 commands.

[2] https://github.com/MixedRealityLab/conditional-voice-recorder

[3] The respective models were obtained by the Snowboy Hotword Detection Engine: https://github.com/Kitt-AI/snowboy

[4] Main changes were: We extended the recording time of the audio snippets from 120s to 180s, we wrote timestamps into the recording file names, we changed the LED-setup due to a mutable microphone and we added an RTC module that guarantees a power supply for the integrated system clock, so that we could disable wifi and bluetooth connections for privacy reasons.

[5] We only analyse the stretches of context that are relevant to the interaction with the VA.

they chose, participants could switch off the microphone attached to the CVR.

In addition to the CVR data, we also collected the audio recordings captured and stored by the VA system itself. The VA system saved audio recordings containing only the user commands, starting with the wake-word *"Alexa"*. Thus, these recordings are generally only a few seconds in length. On top of these VA audio recordings, the VA system kept a log of all user commands in a csv-file. These log files contain a transcription of each user command, generated by the VA's speech detection algorithm.[6]

Since both types of recordings have their advantages and drawbacks, both types of recordings were important for our purposes in order to achieve a collection of commands (and relevant context) that was as exhaustive as possible: CVR-recordings are based on a less well trained speech detection model than the one available to the VA. Hence, the CVR is prone to detection failures, occasionally missing to record actual user commands (i.e. false-negatives) (see section 3). Additionally, the CVR sometimes saves files based on false-positive detections of the wake-word. Due to the inferior speech detection model, false-positives and false-negatives are more common in the CVR-recordings than in the VA recordings. On the other hand, commands that did not trigger a verbal reaction by the VA are sometimes omitted from the list of VA recordings (and the respective csv-logfiles)[7]. Similarly, false-negatives also occur on the side of the VA, leading to no reaction in response to the wakeword. These false-negatives in turn can regularly be found within the CVR-recordings. Moreover, in comparison to the VA system, CVR-recordings contain context information leading up to the user command, the audio of the VA reaction to the command, and follow-up context including user reactions in third position following the VA reactions. Thus, CVR recordings are best suited for all studies that need to take into account the preceding context as well as the VAs response. As a complementary completion of the commands not recorded by the CVR, the VA recordings are however important for micro-longitudinal studies (e.g., on success- and failure-rates) that need to rely on a dataset as exhaustive as possible.

2.3 Data Pre-Processing

For the collection of human-VA interactions, the obtained recordings went through a number of pre-processing steps. After obtaining the CVR-recordings and the VA-recordings plus the VAs' log lists of commands that were issued by the participants during the recording period, we cleared the list of CVR-recordings from false positives by automatically matching the time stamps of the recordings with the time stamps of the logged commands in the VAs' log lists: Only CVR-recordings that contained at least one time stamp of a logged command were kept for further processing and inspection. As a next step, we manually checked and transcribed the remaining CVR-recordings that contained at least one logged command.[8] During this checking and transcription process, any additional commands that were contained in the CVR-recordings but not logged in the original list by the VA were also transcribed and added to the log list of issued commands. In a following step, all recorded commands were manually annotated for a number of factors, including whether the kind of command has been used before by the same user (form-based); a coarse category of what action was requested of the VA; what sentence type has been used for the command; what intonation contours have been used in the wake word and in the command proper; whether the command was successful in terms of the VA-output fitting to the command; and whether the output was followed by any additional comments on the side of the participant in third position. Moreover, we coded whether the original transcription by the speech detection algorithm of the VA was erroneous. In these cases, we corrected the transcription in question and kept a record of the original transcription of the command.[9]

[6] Copies of the VA log files and VA audio recordings were sent to us by the participants after the end of the respective recording periods. Before sending these data to us, participants had the chance to read the log file and listen to the recordings and decide to delete entries and recordings that they did not wish to share without any disadvantages or other consequences.

[7] Typically omitted commands include setting the volume or stopping a running playback of music. While these are omitted from VA recordings, they would still be present in the CVR-recordings.

[8] Iteratively developing and exhaustively implementing a coding scheme is a time-consuming process (Mundwiler et al., 2019; Stivers, 2015). At the point of submission, this checking and transcription has been completed for one participating family. See section 3.2 for more details.

[9] The number of VA speech detection errors varied between participants, see section 3.1 for details.

3 Results

As described in section 2, our collection consists of two kinds of recordings: short audio files captured by the VA, containing just the user commands, and 3-minute recordings captured by the CVR, containing the preceding context of a command, the command itself plus the VA's reaction, as well as follow-up context after the exchange. We will first present the results of our analyses of the obtained VA-recordings in section 3.1, followed by a presentation of the results of initial analyses of a subset of the obtained CVR-recordings in section 3.2.

3.1 Analyses of obtained VA-recordings

For our analysis, we included audio-recorded human-VA interactions in six households for the first 49 to 72 days after the VA had been installed by the users. In total, we obtained 5502 commands that were recorded and logged by the VAs. On average, commands were 4.23 words long, including the wake word (SD = 2.49).[10] The intensity of usage and thus the number of commands recorded and logged by the VA varied considerably between participants (see Table 1): While participating household 5, for instance, only issued 165 commands that were logged by the VA, making use of the VA in 44 out of the recorded 72 days (61%), household 6 produced 2186 logged commands, using the VA in 55 out of the 67 days in the recording period (82%). Listening to all VA-recorded commands and comparing them to the VA-logged transcriptions, we found that of all commands logged by the VA system, the transcript of the command was erroneous in 8%. Proportions of mis-detections of speech input were found to vary across participating households: 1: 10.1%, 2: 9.1%, 3: 9.5%, 4: 3.0%, 5: 6.8%, 6: 9.9%.

Human-VA interactions were found to commonly happen in clusters of commands[11]. This means that, across all logged commands, the probability of a command being issued is highest right after a previous command and drops considerably after a few seconds, with 25% of commands being issued within the first 10 seconds after a previous command, 50% of commands being issued within 22 seconds, and 75% being issued within 182 seconds.This general pattern holds across all recorded households (Figure 1).

Household (*N* members)	*N* days of recording	*N* days of use	*N* logged commands
1 (2)	49	29	313
2 (4)	70	44	1033
3 (1)	69	40	429
4 (3)	68	61	1377
5 (2)	72	44	162
6 (4)	67	55	2186
total:	**395**	**273**	**5502**

Table 1: Recording details by participating households. *N* members specifies the number of regular users of the VA, *N* days of recording specifies the length of the recording periods, *N* days of use specifies the number of days containing commands to the VA, *N* logged commands specifies the number of commands logged by the VA during the recording period.

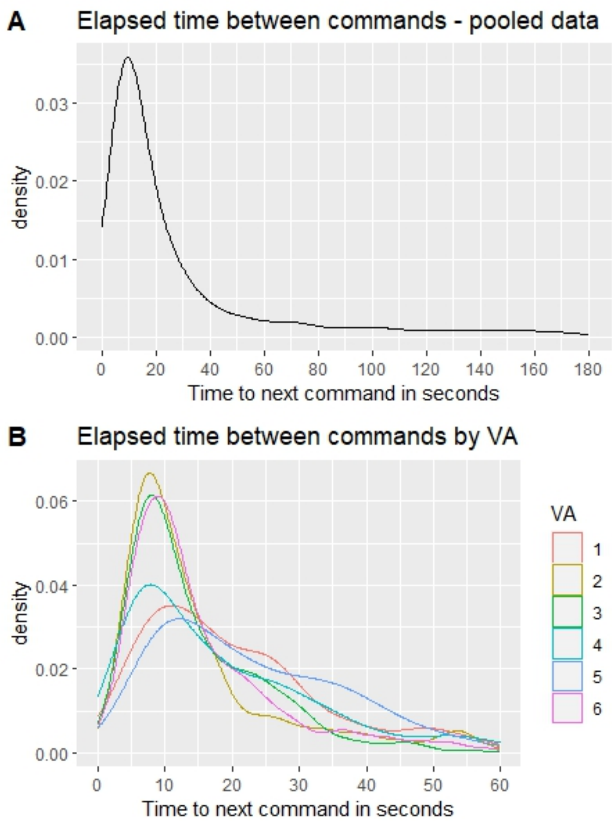


Figure 1: Density plots illustrating probabilities over time for a next command after a previous command. Top panel A shows data pooled by all users. Bottom panel B shows data by VA. In all VA users, the probability for a new command peaks between 7 and 13 seconds after the previous command. *N* = 5502.

[10]The average lengths of commands differed only slightly between households, with the smallest household mean being 3.95 words and the largest being 4.74. SDs for the different households were all between 2.39 and 2.65. VA responses varied much more in length, with a grand mean response length of 12.15 words (SD = 11.63).

[11]With 'commands', we mean the VA being addressed in an utterance by the user starting with the wakeword, mostly containing a request to the VA.

Another observation that holds across all recorded households is that the frequency of commands declines during the recording period. To quantify this observation, we built a general linear mixed effects regression model using the *R* package *lme4* (Bates et al., 2015; R Core Team, 2023), using a Poisson distribution to model the number of commands by the consecutive days of use during the recording period, with random intercepts and random slopes for day of use by household. The model output showed a significant effect of day of use ($\beta = -0.039, SE = 0.007, z = -5.453, p < .001$) and an intercept estimate of 3.438. Note that the link function is logarithmic, meaning that the modeled grand-average number of commands per day at the beginning of the recording period is 31 commands, with the number of commands decreasing by factor 1.04 on each consecutive day (Figure 2). While this factor (as well as the intercept) varied between households, it was found to be smaller than 1 for all households, meaning that the number of commands per day tended to decrease during the recording period for all households ($\beta_1 = -0.058, \beta_2 = -0.055, \beta_3 = -0.049, \beta_4 = -0.010, \beta_5 = -0.033, \beta_6 = -0.031$).

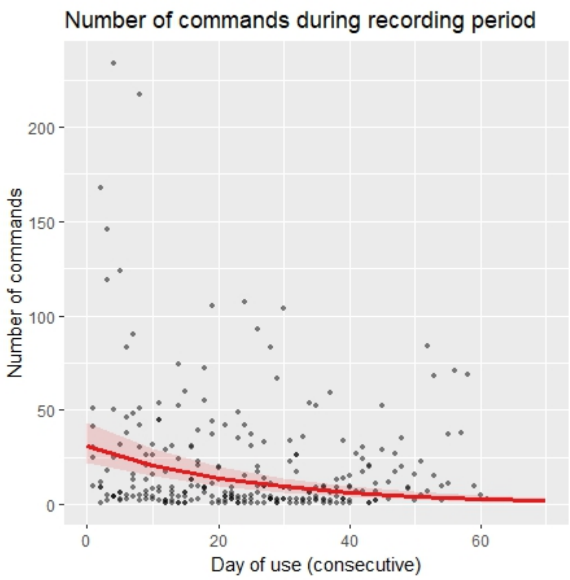


Figure 2: Development of frequencies of commands for consecutive days of use across all households. Dots represent the number of issued commands for each day of usage by any one household. Days without any commands are ignored. The red line represents the fit of a general linear mixed effects regression model (formula = $N_{commands} \sim dayOfUse + (1 + dayOfUse \mid household), family = poisson(link = log)$, see main text for details). The red ribbon represents 68% confidence intervals.

3.2 Initial analyses of a subset of CVR-recordings

As described in section 2.2, our collection consists of two kinds of audio recordings: short recordings of the commands made by the VA and 3-minute recordings containing the commands made by the additionally installed CVR. Analyses of the CVR-recordings are time consuming and still ongoing. Nevertheless, we exhaustively listened to all CVR-recordings of one of the participating households (household 1) that remained after excluding false positives as described in section 2.3. In this section, we present the analyses of the subset of the collection containing data of this example household, serving as a test case for the obtained recordings.

In addition to the 313 commands logged and recorded by this household's VA, we identified another 155 commands in the CVR-recordings that were not originally logged or recorded by the VA, and added these to the list of identified commands, leading to a total number of 468 identified commands. Note that while the VA did not log about a third of the issued commands, this does not mean that the VA was generally unresponsive to these commands. While the VA did indeed not respond to 50 of the total of the 468 identified commands (10.7%), the remaining commands triggered a response in the VA. Most of the originally unlogged commands were either commands to stop the ongoing output of the VA, or to adjust the output volume. These kinds of commands did not trigger a verbal response by the VA, but were generally complied to by directly stopping the current output or adjusting the output volume accordingly. On the flipside, 86 commands (18.4%) that were logged and recorded by the VA were not recorded by the CVR, in most cases probably because the wake word had not been detected, leading to false negatives.

We were interested in the distribution of success rates over different types of sentences (Figure 3). We thus coded all commands regarding their sentence type based on their syntactic structure. Of the 468 identified commands, 15 have a declarative sentence structure (3.2%, e.g., *"Alexa, ah das ist zu schwer"* ("Alexa, ah that's too difficult")), 135 have an imperative sentence structure (28.8%, e.g., *"Alexa, spiel mein Hörbuch"* ("Alexa, play my audiobook")), 105 have an interrogative sentence structure (22.4%, e.g., *"Alexa, wie wird das Wetter heute"* ("Alexa, how is the weather today")), 200 have an elliptical sentence structure (42.7%, e.g.

“Alexa, lauter” (“Alexa, louder”)), and a single case has a deontic infinitive structure (*“Alexa, Werbung überspringen”* (“Alexa, skip ads”)).[12] Another 12 commands have been aborted and not completely uttered (2.6%), mostly consisting of the wake word only.

We annotated all 468 commands regarding their outcome success. If the triggered VA response or output was relevant to (the surface structure of) the uttered command, the command was coded as ‘success’. If, on the other hand, the VA response or output did not fit the command, it was coded as ‘failure’. 298 commands successfully triggered relevant VA reactions, while 149 commands did not trigger the requested response or output and were thus coded as failures (Figure 3).[13] The proportion of failures was found to decline with increasing numbers of commands in a given sentence structure: In the most frequent category, imperative commands, only 26.5% of commands failed; in elliptical commands without a verb form, 29.6% failed; in commands with interrogative sentence structure, 40.7% failed; in the greatly rarer commands with declarative structure, 90.9% failed; and the single case with an infinitive verb form also failed. In a generalized linear model built with the *R* package *lme4*, the number of commands observed per sentence structure as a linear and a quadratic predictor for command success both turn out to be significant ($\beta_{linear} = 0.048, SE = 0.018, z = 2.587, p < .01$; $\beta_{quadratic} = -0.014, SE = 0.001, z = -1.956, p = .050$). This means that the higher the number of total commands used with a given sentence structure, the higher the proportion of successful commands in that structure (Figure 4).[14]

Interestingly, failing commands were found to be produced in clusters, with the probability of a command to fail being greatest right after a failing command, with no or not more than one successful command in between (Figure 5).

In order to test whether users successfully adapt their commands to the VA system over time, adjusting the input so as to increase the success rate, we analysed how the frequency of failing commands changes over time as the users get more experienced with the VA system. While failed com-

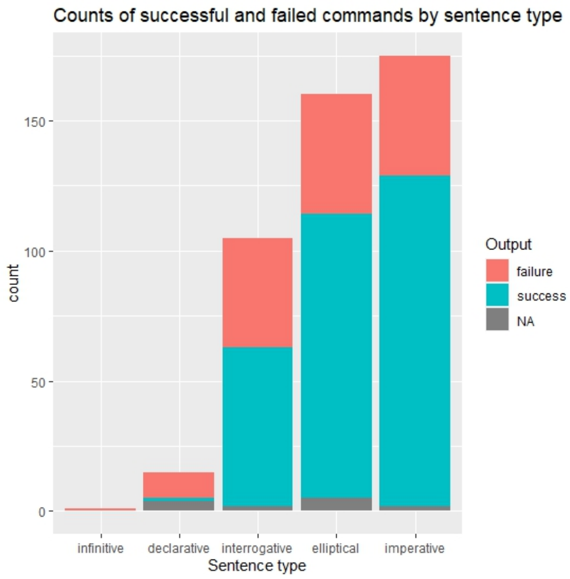


Figure 3: Number of successful and failed commands by sentence type in analysed subset. $N = 454$.

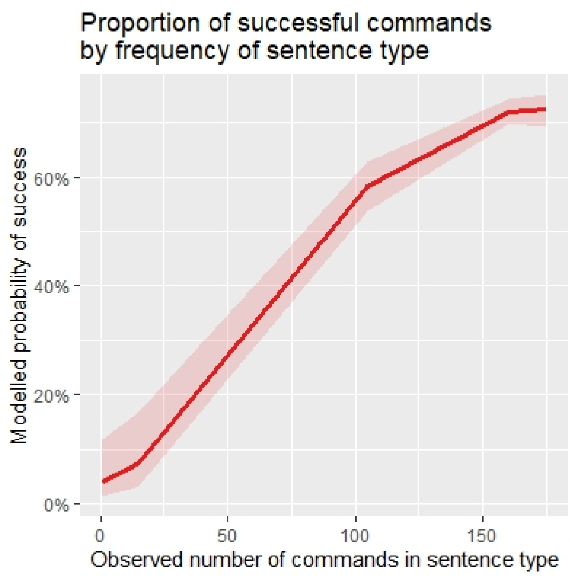


Figure 4: Modeled probability of success of a command as predicted by the observed frequency of the command’s sentence structure. More frequent sentence types show higher proportions of successful commands (see main text for details). Formula = $success \sim N_{structure} + I(N_{structure}{}^2), family = binomial(link = logit)$.

[12]”The vast majority of commands are single sentences. Rare instances of multi-sentence commands are generally not successful, apparently mainly because the VA does not log more than the first main sentence of the command. Example (translated): *User:* alexa play macklemore and like this <<singing> this is the moment>. *VA:* this is macklemore on spotify. (plays some other song by same artist instead of specific song). In this example the VA did not log the complete command and responded only to the first part of the command (logged command: *alexa play macklemore and like this*).

[13]7 aborted commands and 14 uncategorisable commands were not coded.

[14]Note that the relatively frequent stopping-commands (“*Alexa stop*”) were coded as ‘imperative’ here. If they were coded as ‘elliptical’, the general pattern of this result would not change.

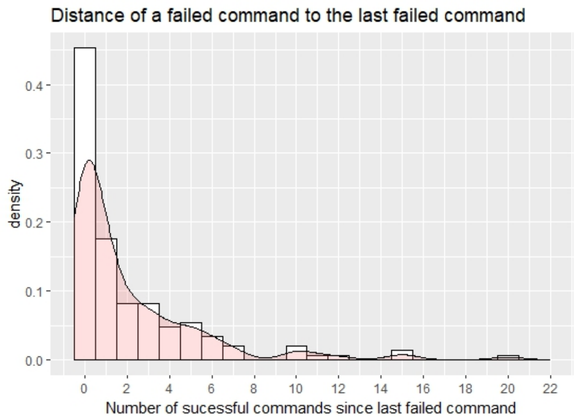


Figure 5: Frequency distribution of the number of successful commands between two failing commands. As the most frequent case, a failing command follows directly after a previous failing command ($N = 67$), with the second most frequent case being a single intermittent successful command ($N = 26$). $N_{total} = 149$.

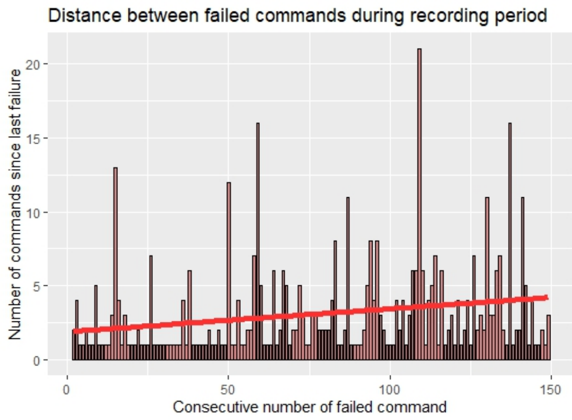


Figure 6: Development over time of the number of commands since the last failing command in analysed subset. For each failed command on the x-axis, bars show how many commands ago the last failure was located in the usage history. If the number of commands since last failure is shown to equal 1, this failed command followed directly upon a previous failed command; if the number is shown to be equal to 2, one successful command has been issued after the previous failed command, and so on. The regression line in red shows that the frequency of failing commands significantly decreased during the recording period, with a slope of 0.02 (formula = $distanceToLastFailure \sim position_{failedCommand}$; $p = .013$).

mands were very common initially, with about every second command failing to trigger an intended response or output, the success rate in the analysis subset of the collection approximately doubled during the recording period of 49 days (out of which the VA was used on 29 days). Hence, at the end of the recording period, only about one in four commands failed to elicit a desired response or output, which makes for an average increase in success rate of 0.66% per failed command (Figure 6).

Given the observation of an increasing success rate over time in combination with the distribution of failing commands across sentence types, we investigated the development of success rates by sentence type in more detail (Figure 7).

Elliptic commands, including standardized commands like setting the volume ('louder', 'softer'), are found to be constantly used over time, showing a high success rate already in the first week of use and even becoming more successful over time. Imperatives, including highly frequent standardized commands like 'stop', are continuously used over time as well. In contrast to ellipticals, however, they don't tend to become more successful over time. Declarative commands, which are mostly failing, are rather rare from the beginning and eventually fade out completely. In qualitative single-case analyses, we can see how declaratives that are not successful in the local sequential context are repaired and, more specifically, replaced by other formats (like interrogatives) for the same use case. E.g. *"Alexa, we'd like to play a game for five year olds"*

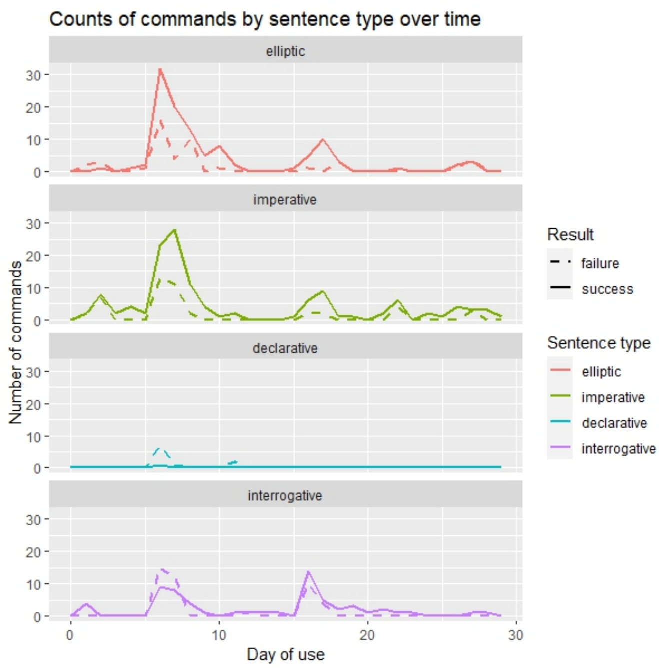


Figure 7: Numbers of failing and successful commands by sentence type during recording period. A single failing infinitive command on day 27 is not plotted here. 12 aborted commands, which are generally failing, are also not plotted here.

(see Transcript 1, lines 01-04 in the Appendix) is locally repaired by the interrogative *"Alexa, can we play a game with you"* (Transcript 1, line 10). The fact that alternative commands are successful (either directly or after several attempts, see below and Transcript 2 in the Appendix for the eventually successful request) and declarative commands fade out in the course of the recording period suggests that declarative sentence types are abandoned in favor of more successful command types. In interrogative commands, we find the success rate to increase over time. Investigating the data with more in-depth analyses shows that this is not the case because interrogatives are used with less trouble in general. Instead, unsuccessful variants of interrogatives are also locally repaired and replaced by types of interrogatives that turn out to be more successful. Transcripts 1 and 2 in the Appendix show that an unsuccessful *can we*-interrogative (*"Alexa can we play a game with you"*, Transcript 1, line 10) is replaced by a successful *wh*-interrogative (*"Alexa what games are there"*, Transcript 2, line 08). Overall, we find that (typically unsuccessful) *can you/we/I*-interrogatives fade out over time in favor of other, more successful, types of interrogatives. We take these first examples as evidence for experience-based, goal-oriented adaptations of users' behaviour in interaction with the VA that lead to a reduction in the proportion of failing commands over time.

4 Discussion and Prospect

In this first description of the new collection of first users' interactions with virtual assistants (VAs), we presented initial observations of patterns of use during the first weeks after installation of the VA. Comprising over 5.000 commands to the VA that were captured in six households with a total of sixteen members, the collection was found to be suitable for micro-longitudinal analyses of the development of patterns of interactions with the VA system. A CVR, selectively recording audio snippets only, has proven to be suitable for field recordings in private settings over a longer period of time. Continuous recording, as well as longitudinal video-recording, would be much more intrusive and less efficient in terms of capturing sequences of focal interest (i.e., sequences featuring interactions with the VA). Moreover, selective recording with a CVR proved to be a practical approach to meet relevant ethical questions, since recordings get limited to stretches of time that are directly relevant to the target research questions of the project. Three-minute stretches of recording have proven to be an apt compromise to grasp sufficient context without covering excess unrelated interaction. This approach also minimised the amount of recorded data, leading to computational efficiency during data curation, inspection, and annotation. Notwithstanding these advantages of this way of audio-only recording, they obviously come with the downside of some situational aspects remaining unanalysable to us: Without video recordings of the relevant sequences, most of the time we are unable to detect with certainty if or when and how users turn to another channel of input (like, e.g., their cellphone), or when they chose to control the VA by pressing a button (e.g. to adjust the output volume). Nonetheless, the applied recording methods strike a worthwhile balance between highly informative, goal-focused data and low invasiveness.[15]

Analysing the data captured by the VA, i.e., audio-recordings and lists of logged transcripts of just the commands (section 3.1), we found a clear and consistent pattern of users to interact with the VA in clusters of commands, meaning that the probability of a command being uttered is highest right after a command has been uttered and quickly declines within a few seconds. Additionally, we found failing commands to cluster as well, with frequently no or not more than one successful command between two failing commands that do not trigger the intended response or output in the VA. These two results seem likely to be related. If a user tries to achieve a certain goal and fails with an initial attempt, any follow-up pursuit of that goal might also fail, due to limitations of the VA. Similarly, a regularly observable pattern of commands that leads to one successful command between two failing commands is a successful stopping command after an initial failing command that triggered unintended output in the VA, followed by a second (possibly again failing) attempt to pursue the initial goal.[16]

While we found failing commands to be initially very frequent in the subset of our collection that

[15] Alongside this project, we collect non-longitudinal video data of VA users to be able to study the use of alternative types of input (like button presses, phone control, etc.) and embodied orientation and conduct.

[16] Example (translated): *VA*: (starts song). *User*: alexa that's the wrong one. *VA*: (beep). *User*: alexa stop. VA: (stops playback). *User*: alexa what other version is there? (no response by VA).

we analysed in more detail, the frequency of failing commands was found to decrease with increasing time of use (see section 3.2). While future analyses will still have to show whether this finding generalises across users, it might be related to the globally observed pattern that the number of uttered commands generally declines during the recorded first weeks of using the VA. The combination of these two findings offers at least two non-exclusive explanations: First, the number of commands might decline over time because users get to know and memorise the limitations in the VA's use cases and consequently try to use the system for less goals, hence uttering less commands. And second, with more experience, users learn to need less commands to achieve their interactional goals. This last scenario of increasing user efficiency might at least partly be caused by users reducing the variation of commands as they get used to the VA system, honing in on more standardised formats that become known to work. Speaking in favour of this possibility, we found that, at least in the subset of the collection that was analysed in more detail, the number of commands using sentence structures that lead to more failures (i.e., declaratives and interrogatives) is lower across the recording period than the number of sentence structures that lead to less failures (i.e., elliptical commands and interrogatives). Given the observation that the frequency of failures tends to decline over time, this structure-frequency effect might well be the result of a learning process that intensifies over time: As users repeatedly fail to achieve their intended goals with commands in a particular sentence structure, they might use the structure less frequently and learn to use other, more successful structures instead. First qualitative analyses of commands support this hypothesis: Local failures (like with declaratives) are found to lead to a local variation of formats in order to repair the trouble (e.g. substituting an unsuccessful declarative with an interrogative format). As a consequence, this can lead to a consistent usage of successful formats and strategies over time. We aim to validate this conceivable pattern, analysing a greater number of users in the course of the current project.

The learning effects contributing to the development and changes of usage patterns, including adaptions to characteristics of the VA system, are a central aspect of our intended future investigations that can be run on the presented collection of human-VA interactions. While we expect meaningful insights to be based on further quantitative analyses on a larger data basis of CVR-recordings, we also plan to adopt more in-depth qualitative analysis regarding the occasions and reasons for specific quantitative results (de Ruiter and Albert, 2017). For instance, we intend to identify possible 'crucial' moments, e.g. at the end of repair sequences, after which users learn how to successfully formulate a specific request, adapting their usage behaviour. Similarly, we plan to analyse in more detail which types of commands do more typically work and why, taking into account both the characteristics of the commands as well as the inherent limitations of the VA system that cannot be mitigated by adaptions in users' behaviour (Pelikan and Broth, 2016; Reeves et al., 2018). Moreover, as the CVR-recordings cover the context around commands, we will be able to investigate the sequential structure of user-VA interactions more thoroughly, analysing user comments in third-position after the VA's response to a command, as well as potential explicit ascriptions (of actions, intentions, etc.) to the VA, both addressed to co-present users and to the VA itself (see also Habscheid et al. (2023)). These investigations of longer sequences will also enable us to conduct a more in-depth analysis of repair sequences (see also Krummheuer (2010)) and their outcomes, as well as their development over time of usage. The results of these future analyses should shed light on the question on what levels users adapt to the VA, and how human-VA interactions change with an accumulating history of interacting with the respective VA.

On the basis of the presented new collection, we expect to generate fruitful insights into the dynamics of human-VA interaction. Due to its size and focus, the collection lends itself to mixed-methods approaches, with intended future investigations likely profiting from mutually informing insights from quantitative and qualitative analyses. While the former offer powerful tools to discover global usage patterns, the latter, especially conversation analytic qualitative single case analyses, offer apt methods to identify fine-grained aspects of sequential patterns and unveil additional information about occasions, reasons and routinization of users' behaviour and the practices they develop over time. A combination of both approaches will be necessary to draw an encompassing picture of change in practices of VA users over time.

References

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4.

de Ruiter, J. P. and Albert, S. (2017). An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology. *Research on Language and Social Interaction*, 50(1):90–107.

Fischer, J. E., Reeves, S., Porcheron, M., and Sikveland, R. O. (2019). Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–8, Dublin Ireland. ACM.

Habscheid, S., Hector, T., and Hrncal, C. (2023). Human and Non-Human Agency as Practical Accomplishment: Interactional Occasions for Ascription and Withdrawal of (Graduated) Agency in the Use of Smart Speaker Technology. *Social Interaction. Video-Based Studies of Human Sociality*, 6(1):1–31.

Krummheuer, A. L. (2008). Zwischen den Welten: Verstehenssicherung und Problembehandlung in künstlichen Interaktionen von menschlichen Akteuren und personifizierten virtuellen Agenten. In Willems, H., editor, *Weltweite Welten*, pages 269–294. VS Verlag für Sozialwissenschaften, Wiesbaden.

Krummheuer, A. L. (2010). *Interaktion mit virtuellen Agenten? Zur Aneignung eines ungewohnten Artefakts.* Lucius, Stuttgart.

Mlynář, J., de Rijk, L., Liesenfeld, A., Albert, S., Stommel, W., and Topinková, R. (in prep.). Interacting with AI: A scoping review of empirical ethnomethodological and conversation analytic studies.

Mundwiler, V., Kreuz, J., Müller-Feldmeth, D., Luginbühl, M., and Hauser, S. (2019). Quantitative und qualitative Zugänge in der Gesprächsforschung: Methodologische Betrachtungen am Beispiel einer Studie zu argumentativen Gruppendiskussionen. *Gesprächsforschung*, 20:323–383.

Pelikan, H. R. and Broth, M. (2016). Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932, San Jose California USA. ACM.

Pitsch, K., Kuzuoka, H., Suzuki, Y., Sussenbach, L., Luff, P., and Heath, C. (2009). "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 985–991, Toyama, Japan. IEEE.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada. ACM.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reeves, S. and Porcheron, M. (2022). Conversational AI: Respecifying participation as regulation. In *Handbook of Digital Society*, pages 573–592.

Reeves, S., Porcheron, M., and Fischer, J. (2018). 'This is not what we wanted': designing for conversation with voice interfaces. *Interactions*, 26(1):46–51.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., and Uhmann, S. (2011). A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 12:1–51.

Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Stivers, T. (2015). Coding Social Interaction: A Heretical Approach in Conversation Analysis? *Research on Language and Social Interaction*, 48(1):1–19.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, Cambridge.

Acknowledgments

We thank Svenja Filthaut for automating the detection of false-positive CVR-recordings.

Author contributions

MB: field recording, data annotation, formal analysis, data curation, writing - original draft, writing - review and editing, visualization, supervision
HH: conceptualization, methodology, writing - original draft, software adaptation CVR, hardware conceptualization and building CVRs (final version), data annotation, supervision, project administration
SR: conceptualization, methodology, writing - review and editing, supervision, software and hardware prototype CVR (first replication)

A Appendix

List of coding categories

- commandID
  - unique identifier of each command
- date
  - calender date of command
- time
  - clock time of command
- commandTranscription
  - text transcript of user command
- sentenceType
  - grammatical sentence type of command
    * declarative
    * elliptical
    * imperative
    * infinitive
    * interrogative
    * abortion
- responseTranscription
  - text transcript of VA reponse to user command
- dayOfRecording
  - day of recording, also counting days when no command was produced
- dayOfUse
  - day of use of the VA, not counting days when no command was produced
- VAtranscriptionCorrect
  - coding if the automatic speech recognition process transcribed the user command correctly
    * yes
    * no
- commandSuccess
  - coding whether the command triggered a fitting output to from the VA
    * yes
    * no
    * unclear

Transcript 1:[17]
CVR03-recording-220724162605

MO = Mother; CH = Child; AL = Alexa

01	MO	aLExa,
02		wir möchten ein SPIEL spielen?
		we'd like to play a game
03		(0.3)
04	MO	fü:r FÜNFjährige.
		for five year olds
05		(1.6)
06	AL	entschuldigung das weiß ich leider nicht
		sorry I do not know that unfortunately
07		(2.0)
08	MO	hö? ((lacht))
		huh? ((laughs))
09	CH	((kichert))
		((chuckles))
10	MO	°hh alExa können wir ein SPIEL mit dir spielen.
		Alexa can we play a game with you
11		(2.5)
12	MO	aLExa,
13		(1.0)
14	MO	°h können wir ein SPIEL mit dir spielen.
		can we play a game with you
15		(1.2)
16	AL	um musik aus deiner amazon musik bibliothek abzuspielen frage einfach nach dem song interpreten oder dem album das du gerne hören möchtest
		to play music from your amazon music library *just ask for the song, artist or album* *you would like to listen to*

About ten minutes after Transcript 1, mother and child try again to play a game with Alexa. In between, there was one successful request done by the child: After saying ”Alexa what can we play”, Alexa responds ”Okay then let's choose a great game”, offering a list of possible games. The mother tries to replicate this, first using an unsuccessful interrogative, and then eventually formulating a successful request (Transcript 2):

[17]Transcripts were created based on GAT2 transcription conventions (Selting et al., 2011).

Transcript 2:
CVR03-recording-220724163721

MO = Mother; CH = Child; AL = Alexa

01	MO	((lacht)) °h mach DU doch nochmal;=
		((laughs)) you do that again
02		=das hast du eben SUper gemacht.
		you have just done great
03		°h WAS hast du sie gefragt-=
		what did you ask her
04		=was für SPIEle gibt es. gell,
		what games are there, right
05		(0.9)
06	CH	JAha.
		yes
07	MO	ja FRAG se nochma;=
		well ask her again
08		=aLExa was für SPIEle gibt es.
		Alexa what games are there
09		(1.7)
10	AL	okay spiele, lass uns eins zum spielen finden
		okay games, let us find one to play

Retrieval-Augmented Neural Response Generation Using Logical Reasoning and Relevance Scoring

Nicholas Thomas Walker
Norwegian Computing Center
Oslo, Norway
walker@nr.no

Stefan Ultes
University of Bamberg
Bamberg, Germany
stefan.ultes@uni-bamberg.de

Pierre Lison
Norwegian Computing Center
Oslo, Norway
plison@nr.no

Abstract

Constructing responses in task-oriented dialogue systems typically relies on information sources such the current dialogue state or external databases. This paper presents a novel approach to knowledge-grounded response generation that combines retrieval-augmented language models with logical reasoning. The approach revolves around a knowledge graph representing the current dialogue state and background information, and proceeds in three steps. The knowledge graph is first enriched with logically derived facts inferred using probabilistic logical programming. A neural model is then employed at each turn to score the conversational relevance of each node and edge of this extended graph. Finally, the elements with highest relevance scores are converted to a natural language form, and are integrated into the prompt for the neural conversational model employed to generate the system response.

We investigate the benefits of the proposed approach on two datasets (KVRET and GraphWOZ) along with a human evaluation. Experimental results show that the combination of (probabilistic) logical reasoning with conversational relevance scoring does increase both the factuality and fluency of the responses.

1 Introduction

Although Large Language Models (LLMs) are widely used for conversational response generation, they still suffer from a number of shortcomings, including their propensity to produce hallucinated content (Ji et al., 2023). Recent work has demonstrated how to exploit external information sources such as knowledge bases (KBs) to improve the output of LLMs in various downstream tasks (Yu et al., 2022a), including dialogue systems (Wang et al., 2021). A promising approach is Retrieval-Augmented Generation (RAG), which operates by first retrieving relevant information from external sources and then augmenting the input provided to the LLM with this retrieved content (Lewis et al., 2020). While RAG has been demonstrated to reduce hallucinations (Shuster et al., 2021), LLMs are nonetheless easily distracted by irrelevant information (Shi et al., 2023). For this reason, one should strike a balance between providing the model with potentially useful information and avoiding overloading it with too many spurious or irrelevant facts.

Moreover, while LLMs have recently shown some success at reasoning benchmarks (Bubeck et al., 2023), their ability to engage in multi-step reasoning remains poor. In particular, Dziri et al. (2023) provide a systematic investigation of the performance of LLMs on several compositional reasoning tasks, and find that those models largely rely on pattern matching shortcuts and fall short of exhibiting generic problem-solving skills.

This paper presents a novel approach to retrieval augmented generation in task-oriented dialogue systems that seeks to address those challenges. Following (Walker et al., 2022), we represent the background knowledge of the system as a *graph* of dynamically updated facts representing the dialogue state. This initial graph is first enriched at each turn with derived facts inferred through probabilistic logical programming using a limited number of rules, using ProbLog as a framework (Fierens et al., 2015). The conversational relevance of each fact is then scored using a neural model, based on various features expressing both the conversational saliency of each entity and semantic similarity between the fact and the recent dialogue history. The most relevant facts are then converted into sentences and incorporated into the input of the response generation model. Crucially, the relevance scoring model and the response generation model are optimised jointly based on dialogue examples. Figure 1 provides a general sketch of the approach.

The paper makes the following contributions:

1. The use of probabilistic logical programming

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

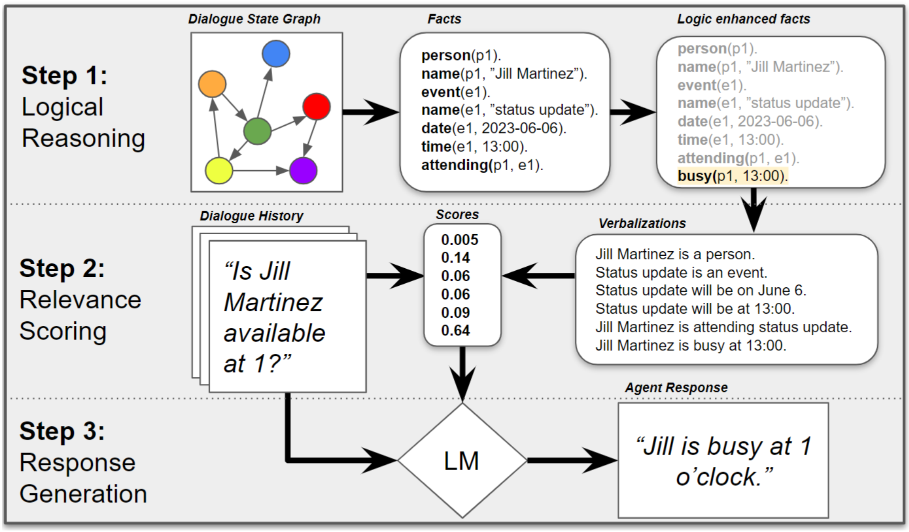


Figure 1: General sketch of the proposed approach. The starting point is a dialogue state represented as a knowledge graph that combines both background information and various features extracted from the dialogue turns (transcriptions, speakers, entity mentions). In Step 1, the facts representing the nodes and edges of the graph are first extended with derived facts using probabilistic logical programming. Those facts are then verbalized (converted into natural language sentences) in Step 2, and a neural model is employed to score their conversational relevance with regard to the current dialogue context. Finally, the k most relevant facts are included in Step 3 as part of the prompt for the neural language model responsible for producing the actual system response.

to extend the knowledge graph representing the current dialogue state with derived facts.

2. A neural scoring model that relies on both dialogue-level features (such as recency) and semantic similarity to determine the most relevant nodes and edges of this augmented graph.

3. An empirical evaluation of the above approach using two dialogue datasets (KVRET and GraphWOZ) and a human evaluation.

2 Related Work

Several papers have investigated the use of neural models to retrieve relevant information from knowledge sources and integrate their results in response generation. Dinan et al. (2018) distinguish between the tasks of knowledge selection and response generation for knowledge-grounded dialogue agents. This information may be structured or unstructured (Young et al., 2018; Zhao et al., 2020), and generally consists of documents describing entities which may be relevant to the dialogue. For an open-domain dialogue model, this background information can be drawn from sources such as Wikipedia. In task-oriented dialogue, relevant information will depend more heavily on the domain of the dialogues. Previous work has also demonstrated the effectiveness of jointly learning a language model with a knowledge retrieval model (Zhang et al., 2021), simplifying the task of identifying relevant items without labelled data.

External information, often in the form of KBs, is crucial to many dialogue models (Ghazvininejad et al., 2018; Parthasarathi and Pineau, 2018; Zhang et al., 2018; Madotto et al., 2018). Multiple approaches exist for combining retrieval and generation to yield higher quality responses, such as by first generating a response and subsequently refining it (Weston et al., 2018). The model used by Peng et al. (2023) queries an LLM and evaluates the output for factuality, and re-queries the system with feedback to elicit a more factual response.

Thulke et al. (2021) propose an approach which samples a subset of the background knowledge rather than optimizing over the entirety of it, a process that we also integrate in our model training process. The Global-to-Local Knowledge Selection model is an alternative which pre-selects information across the whole of the background knowledge using topic transition vectors (Ren et al., 2020). Meanwhile, He et al. (2021) proposed a model which integrated information about system API calls to the retrieval model.

Numerous models make use of large, static, knowledge bases to augment language models. The KETOD model (Chen et al., 2022) used Wikipedia data to enhance a task-oriented dialogue system's responses with information about entities in the dialogue. Likewise, Kim et al. (2020) and Zhan et al. (2021) modelled knowledge selection with a latent variable model, which have also shown strong results for RAG in a zero-shot setting (Li et al., 2020). Moon et al. (2019) investigated a graph decoder model using random walks over a knowledge graph containing dialogue relevant information. Paranjape et al. (2021) made use of a "guide" retriever model to use posterior information from responses to help the retriever model learn from relevance in both the input and reference output responses. Cai et al. (2019) made of of a "skeleton-guided" response generator in a dialogue system.

RAG has also been used to improve common sense reasoning (Yu et al., 2022b), or incorporate graphs of commonsense knowledge to the model (Zhang et al., 2019a). Common sense knowledge in the SenticNet KB has also been used as a source of knowledge for a dialogue model (Young et al., 2018), albeit without logical reasoning over the graph. Liu et al. (2019b) explored a model which used multi-hop reasoning to identify a relevant vertex in a graph of "factoids" which are each associated with unstructured sentences. Other models have made use of linguistic rule-based components to combine semantic representations of the dialogue state with background knowledge to improve empathetic responses and dialogue flow in task-oriented dialogue (Smith et al., 2011).

A hierarchical approach to knowledge grounded task-oriented dialogue was presented by Lee and Jeong (2023), where the pipeline is composed of domain identification, entity extraction, and a pre-trained language model to rank relevant documents. Other work proposed a novel factuality-specific sampling algorithm to improve LLM output (Lee et al., 2022), while Bonetta et al. (2021) used k-nearest neighbors to find relevant information.

3 Approach

As illustrated in Figure 1, the approach proceeds in three steps. Probabilistic logical programming is first employed to extend the initial knowledge graph with new facts based on a small set of rules. A neural scoring model then determines the relevance of those facts in the current dialogue context. The most relevant facts are then included as part of the input to the second neural model, which is responsible for the actual response generation. The next sections describe those steps.

3.1 Dialogue state representation

Following (Walker et al., 2022), we represent the current dialogue state (along with other background information that might be relevant for response generation) as a *knowledge graph* consisting of multiple entities connected by relations. The graph is always grounded in a specific dialogue and continuously evolves during the interaction, with new nodes and edges representing dialogue turns, speakers, or entity mentions. The dynamic and dialogue-specific nature of this knowledge graph stands in contrast with the static KBs (based on e.g. Wikipedia or similar sources) typically used in knowledge-grounded generation.

To account for uncertainties associated with noisy or partial observations (such as ASR transcriptions of user utterances or ambiguous referential links), both node attributes and labelled edges may be associated with probabilities.

3.2 Probabilistic logical programming

To explicitly reason over this graph, we rely on the probabilistic logical programming language ProbLog (Kimmig et al., 2011; Fierens et al., 2015). We assign each node to a unique identifier and represent the node attributes and edges as (ground) logical predicates, as illustrated in Figure 1. Node attributes and edges associated with a probability < 1 are expressed as probabilistic facts.

3.2.1 ProbLog

A ProbLog program consists of two parts: a set of ground probabilistic facts, and a logic program, expressed as a set of logical clauses. The clauses may be themselves associated with probabilities. ProbLog also allows for the definition of "annotated disjunctions" where mutually exclusive facts are coupled with a discrete probability distribution. Syntax-wise, ProbLog is a probabilistic extension of Prolog and supports both probabilistic and inductive reasoning. Given a set of logical rules and ground facts, ProbLog provides inference algorithms to efficiently query the probability of one or more predicates. This inference is done by converting the facts and logical program to a compact encoding such as Sentential Decision Diagrams (Vlasselaer et al., 2014) and then running weighted model counting (Chavira and Darwiche, 2008) on this compiled representation.

3.2.2 Entity linking rules

An important task in goal-oriented dialogues is to connect entity mentions to the actual entities present in the KB. For instance, if "Jill Martinez" is mentioned by the user, this mention must be linked to the actual node (p_1) for that person in the KB. Entity mentions may correspond to named entities, but may also take the form of pronouns ("she") or generic noun phrases ("the meeting").

We first detect entity mentions in user utterances using a neural sequence labelling model fine-tuned on labelled, in-domain data from a pretrained ROBERTA model (Liu et al., 2019a). A small set of probabilistic ProbLog rules is then employed to determine the most likely reference among the entities in the knowledge graph. Those rules take advantage of both edit distance metrics and recency measures (Walker et al., 2022). Each rule is attached to a probability reflecting its strength. Those probabilities are estimated empirically from partial interpretations on the training data, following the approach described in Gutmann et al. (2011). After applying those entity linking rules, the outcome is then written back to the knowledge graph as probabilistic `refers_to` edges linking each observed mention to the entity it refers to.

3.2.3 Commonsense rules

Consider a scenario where a task-oriented dialogue system must answer a user question:

"What events do I have today?"

Assuming the knowledge graph contains basic information about calendar events such as their date, time and attendees, answering this question rests upon multiple reasoning steps. As multi-step reasoning remains a challenging task for language models (Liu et al., 2023), we specify a small number of commonsense reasoning rules to automatically derive new facts from the current dialogue state. For the above example, the connection between dates and events in context can be made explicit with the following rule:

```
person(P), event(E),
attendee(E,P), date(E,D),
date(today,D)
⟹ attending_today(E,P)
```

A second example of a logical rule is as follows:

```
room(R), ¬(event(E),
location(E,R), date(E,D),
date(today,D), start_time(E,ST),
end_time(E,ET),
time_between(T,ST,ET,1))
⟹  room_available_today(R,T)
```

The above rule simply states that a room R is available today at a given time T if no event is scheduled at that time in that room.

The goal of those commonsense rules is to deduce facts that may provide useful information to the response generation model. Those logically derived facts will typically correspond to information that may be queried by the users, such as a person's agenda for today or the availability of a room at a given time. To avoid deriving too many spurious or irrelevant facts, we only query ProbLog for facts pertaining to entities recently mentioned in the dialogue history. For our experiments, we query entities mentioned in the current turn.

After applying both entity linking and commonsense rules, the facts are converted to a natural language *verbalization*. Each predicate is associated with a handcrafted template which creates a natural language form of the fact. For example, a person defined by the fact `person(p_123)` with a name `name(p_123, "Lisa Wilson")` can be verbalized as *Lisa Wilson is a person.*

3.3 Relevance Scoring

The second component of the proposed approach is a neural model that scores the relevance of the verbalized facts (including both the intial ones as well as the ones derived through logical reasoning). Given a dialogue history $x = [u_1, ...u_n]$ corresponding to a list of utterances and a set of verbalized facts Z, the relevance scoring model expresses the probability $P(z|x)$ that the fact $z \in Z$ is relevant for responding to x.

The model is expressed as a simple feedforward neural network based on the following inputs:

1. Semantic similarity measures between z and x, using the cosine similarity between the embedding of the verbalized fact z and the embedding of the most recent k utterances in the dialogue history x (concatenated if $k > 1$):

$$sim(z, x) = \frac{Enc(z) \cdot Enc(x_{[n-k:n]})}{\|Enc(z)\| \; \|Enc(x_{[n-k:n]}\|}$$

The *Enc* embeddings are obtained with a sentence-BERT model (Reimers and Gurevych, 2019) optimized for semantic search and question answering[1]. The cosine similarity is computed for $k = 1$ and 2.

[1] https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

2. BM25 information retrieval scores (Robertson et al., 2009) using the verbalized facts as database and the user utterance as a query.

3. Recency score expressing whether the fact z pertains to a recently mentioned entity. This score relies on the `refers_to` predicates derived from entity linking and captures the conversational saliency of entities and facts related to them. For instance, facts related to p_1 in Fig. 1 are salient since the person is mentioned in the last utterance.

The relevance model $P(z|x)$ is trained jointly with the response generation. Concretely, we define the probability of a response y given a dialogue history x as:

$$P(y|x) = \sum_{z \in Z} P(y|x, z) P(z|x) \tag{1}$$

where $P(y|x, z)$ is provided by the response generation model (see below), and express the probability of a response y given a prompt concatenating the dialogue history x and fact z, and $P(z|x)$ express the relevance of z for x. The relevance model $P(z|x)$ is then optimized by back-propagating the cross-entropy loss of Eq. (1) using a training set of dialogue examples. Intuitively, a fact will therefore be deemed as relevant if its inclusion in the prompt makes it relatively easier for the generation model to produce the correct response. To ensure the inference remains efficient, Eq. (1) is simplified by sampling the K most relevant facts instead of marginalizing over all possible facts.

3.4 Response Generation

The final step of our approach is to generate a response y based on both the current dialogue history x and a set of relevant facts $z_1 : z_K$, where K denotes the number of facts (sorted by relevance) to include in the input prompt. Any pretrained language model can be employed for this task. We rely for our experiments on both the GODEL model (Peng et al., 2022) which is specifically designed for goal-oriented dialog as well the generic GPT 3.5 model (Brown et al., 2020).

4 Evaluation

We evaluate the performance of the proposed approach on two existing dialogue datasets along with a human evaluation. We present below the experimental design, and discuss the results.

4.1 Datasets

GraphWOZ (Walker et al., 2022)

GraphWOZ contains task-oriented dialogue with dialogues discussing people and places in a fictional organization to schedule meetings and discover information. Each dialogue is paired with synthetically generated calendar events. The graphs contain fictive people, rooms, and events along with dialogue information such as utterances and mentions of entities in utterances.

Although the calendar information is synthetically generated, generation of new dialogue utterances with calendar information may not accurately reflect real system-human interaction. In consideration of this factor and to compensate for the small amount of training data, we augment the GraphWOZ training set with modified versions of the original dialogues where entities are replaced in both the knowledge base and dialogue history. We replace entity names with randomly sampled replacements, and the dates and times of the dialogues and events in the KBs are replaced such that relative terms such as "today", "tomorrow", "morning", and "afternoon" remain consistent in the modified dialogue.

The entity linking and commonsense rules for this dataset are provided in the Appendix.

KVRET (Eric et al., 2017)

This dataset contains task-oriented dialogue in three domains: weather, navigation, and calendar scheduling. Each type of dialogue contains associated KB information representing objects of interest which are discussed in the dialogue. The knowledge bases in KVRET were created by randomly sampling attribute values for defined slots according to the domain. We convert these KBs into a ProbLog program along with the user utterances and mentions of objects. For simplicity, we take a string equality match of a substring in the utterance to an object in the KB as a `refers_to` relation.

We rely on three simple ProbLog rules for this dataset. The weather domain has a rule which determines "today" along the weather for a particular day and location. In the calendar scheduling domain, we provide a rule to handle location names with multiple potential referents. Lastly, for the navigation domain we define a rule comparing the distance from the user to two points of interests and determining which one is closest. Each of

Model	Dev			Test		
	BLEU	METEOR	BERTScore	BLEU	METEOR	BERTScore
$\text{GODEL}_{\text{NoFacts}}$	0.17	0.37	0.89	0.11	0.36	0.88
$\text{GODEL}_{\text{AllFacts}}$	0.14	0.38	0.88	0.13	0.33	0.88
$\text{GODEL}_{\text{Relevance}}$	0.18	0.38	0.89	0.14	0.33	0.88
$\text{GODEL}_{\text{Relevance+Logic}}$	0.17	0.37	0.89	0.16	0.35	0.88
$\text{GPT}_{\text{NoFacts}}$	0.08	0.35	0.88	0.06	0.32	0.87
$\text{GPT}_{\text{AllFacts}}$	0.07	0.36	0.88	0.06	0.32	0.87
$\text{GPT}_{\text{Relevance}}$	0.07	0.35	0.88	0.06	0.35	0.87
$\text{GPT}_{\text{Relevance+Logic}}$	0.07	0.37	0.88	0.07	0.36	0.87

Table 1: Results with reference-based metrics on the development and test set of GraphWOZ.

Model	Dev			Test		
	BLEU	METEOR	BERTScore	BLEU	METEOR	BERTScore
$\text{GODEL}_{\text{NoFacts}}$	0.18	0.45	0.91	0.11	0.36	0.91
$\text{GODEL}_{\text{Relevance}}$	0.18	0.42	0.91	0.16	0.41	0.91
$\text{GODEL}_{\text{Relevance+Logic}}$	0.20	0.43	0.91	0.17	0.42	0.91

Table 2: Results with reference-based metrics on the development and test set of KVRET.

these rules therefore makes information explicitly available to the system which would be unavailable from context or otherwise require logical inference that an LLM is not optimized to perform.

4.2 Models

We experiment with the four following types of response generation models:

NoFacts Generation model that does not use the knowledge graph at all and produce a response based on the current dialogue history.

AllFacts+Logic Generation model using all verbalized facts (including logically derived ones), without relevance scoring. These facts are shuffled and truncated to fit into the context window of the generation model.

Relevance Generation model using the initial facts from the knowledge graph (but without logically derived ones) ranked using the relevance scoring model. The 10 most relevant facts are then prepended to the prompt.

Relevance+Logic Generation model using both the initial facts and the logically derived ones, along with the relevance scoring model to select the 10 most relevant facts.

We experiment with two generative models: the encoder-decoder GODEL (Peng et al., 2022), which is pre-trained on large volumes of multi-turn dialogues, and the recent GPT-3.5 model (Brown et al., 2020). We first test the response generation capabilities of GPT-3.5 with the three different approaches on GraphWOZ. For each turn, we provide the system with the dialogue history up to the current turn. When using all facts, background knowledge is added as a single document in the initial prompt, as repeating the entire document of the facts at each turn would result in truncation of the dialogue history without adding additional information.

4.3 Metrics

For both GraphWOZ and KVRET, we use standard evaluation metrics such BLEU, METEOR and the averaged BERTScore F1 (Zhang et al., 2019b). We also use the recently introduced UniEval (Zhong et al., 2022), a reference-free metric which has been shown to correlate well with human judgments.

We also evaluate the *factuality* of the responses by manually annotating them with two types of error. The first error type are *hallucinations*, which we define as either (a) a statement that contradicts the KB, including contradictions implied by the dialogue context ; (b) a statement referring to a nonexistent entity in the KB ; or (c) a statement describing a calendar action that would create a calendar conflict if enacted.

Model	Coherence	Groundedness	Naturalness	Understandability
$\text{GODEL}_{\text{NoFacts}}$	0.946	0.908	0.871	0.864
$\text{GODEL}_{\text{AllFacts}}$	0.975	0.943	0.903	0.896
$\text{GODEL}_{\text{Relevance}}$	0.916	0.878	0.862	0.855
$\text{GODEL}_{\text{Relevance+Logic}}$	0.979	0.951	0.868	0.861
$\text{GPT}_{\text{NoFacts}}$	0.951	0.880	0.943	0.938
$\text{GPT}_{\text{AllFacts}}$	0.952	0.878	0.931	0.925
$\text{GPT}_{\text{Relevance}}$	0.969	0.912	0.935	0.931
$\text{GPT}_{\text{Relevance+Logic}}$	0.949	0.883	0.928	0.922

Table 3: UniEval Score (Reference-free) on the test set of GraphWOZ.

Model	Dev		Test	
	Hallucinations	Retrieval Errors	Hallucinations	Retrieval Errors
$\text{GPT}_{\text{NoFacts}}$	34 (18%)	17 (9%)	32 (17%)	16 (8%)
$\text{GPT}_{\text{AllFacts}}$	23 (13%)	11 (6%)	23 (13%)	20 (11%)
$\text{GPT}_{\text{Relevance}}$	21 (12%)	13 (7%)	24 (13%)	16 (8%)
$\text{GPT}_{\text{Relevance+Logic}}$	15 (8%)	14 (8%)	25 (14%)	9 (5%)

Table 4: Turns containing hallucinations and retrieval errors (GraphWOZ, 181 turns in Dev, 180 in Test)

The second type of error occurs when the system fails to retrieve information it has access to or should have access to. We denote this type of error as a *retrieval error*, and we annotate a turn as containing a retrieval error if it contains:

- A statement denying having access to information which exists in the calendar, or does not answer a question while answering another ;
- A confirmation of inviting a person to a meeting without specifying whom ;
- A request to confirm a piece of information that was already confirmed.

4.4 Results

Tables 1 and 2 respectively provide the results of the reference-based metrics on GraphWOZ and KVRET. We observe modest differences between the models, particularly on the development set. This is expected as those metrics are not really suited to measure the extent to which a language model produces factual answers. We do nevertheless observe a slight increase in BLEU score when logic-derived facts are added. This result is consistent in both GraphWOZ and KVRET. The differences in UniEval, shown in Table 3 were also small, with the model using all facts performing comparably to the Relevance+Logic model.

However, for the manually annotated hallucinated and retrieval errors in Table 4 (only annotated for the GPT models), we do observe a substantial reduction in overall errors for the Relevance+Logic model compared to the baselines. Particularly in the development set, the model using all facts (without relevance scoring) frequently produce hallucinations, such as incorrect attributes of various people and events.

4.5 Error Analysis

Qualitatively, the errors produced by the models also differed. The GPT models using selected facts were more likely to provide all available information, whether requested or not. In some cases, the model produced statements connecting unrelated facts and implying an error, for example:

> *User*: Hi robot, my name is Curtis Williams, please let me know when the meeting called, deliverables team meeting. [*sic*]
> *System*: Hello Curtis! The deliverables team meeting will be held on Friday, April 08 at 11:30 AM in the Beta conference room. Annette Harding is organizing the meeting on behalf of the Mathematics group. You and Angela Jimenez are scheduled to attend the meeting. The meeting is scheduled to last for 30 minutes until the Niches team meeting starts at 12:00. Michael Glover is attending the Niches team meeting and Annette Harding is also attending the Deliverables team meeting with you. Is there anything else you need assistance with?

Although the general content is factual, the niches team meeting was on a different day despite the implication of the phrasing. The GPT model with all facts responded erroneously at the same turn as:

> Alright Curtis, I have checked your calendar and it looks like there is a "Deliverables Team Meeting" scheduled for tomorrow at 2:00 pm in room 619. Would you like me to add any notes or reminders for the meeting?"

While more concise, the system incorrectly states the time of the event (11:30) and date (today), as well as the location (Beta). This apparently arises from misattribution of background facts relating to other events scheduled across times and locations.

4.6 Human Evaluation

Experimental setup

To confirm the performance of the approach in actual interactions, we also conduct a human evaluation in the context of a receptionist scenario similar to GraphWOZ, where the participant interacts with the system to find information about entities and schedule events. We recruited 16 participants including students from the university and employees to interact with the dialogue systems through text. Users were instructed to interact with the system to accomplish a task, mark the conversation as finished when either the task appeared complete or the dialogue system unrecoverably failed. After each dialogue, the users were prompted to rate the dialogue on a scale of 1 to 5 for two statements, where 1 is "Never", 2 is "Mostly Not", 3 "Sometimes", 4 "Mostly", and 5 "Always":

- *The system responded to me in a conversationally relevant way.*
- *The system successfully completed my task and gave me the information I asked for.*

Users were instructed to repeat this process for 30 minutes, with priority given to conversation quality. For each dialogue, a model was randomly selected and a random dialogue state similar to the GraphWOZ dataset was generated for the dialogue. A task was then randomly generated from a set of task templates involving fictive entities. The collected dialogues were then manually annotated for both hallucinations and retrieval errors. Because the total number of turns varied from model to model, we evaluate the proportion of turns which contain hallucinations and retrieval errors.

Model	**Task**	**Appropriateness**
$\text{GODEL}_{\text{None}}$	3.35	3.07
$\text{GODEL}_{\text{All}}$	3.75	3.63
$\text{GODEL}_{\text{Logic}}$	4.08	3.75
GPT_{None}	4.18	4.59
GPT_{All}	4.09	4.32
$\text{GPT}_{\text{Logic}}$	4.37	4.11

Table 5: Average participant scores for the model task completion and appropriateness criteria.

Model	**Hallucinations**	**Retrieval errors**	**#**
$\text{GODEL}_{\text{None}}$	0.17	0.41	105
$\text{GODEL}_{\text{All}}$	0.24	0.32	84
$\text{GODEL}_{\text{Logic}}$	0.22	0.21	67
GPT_{None}	0.32	0.14	88
GPT_{All}	0.22	0.39	117
$\text{GPT}_{\text{Logic}}$	0.20	0.12	132

Table 6: Proportion of system responses containing either hallucinations or retrieval errors in the human evaluation experiments. The last column indicates the total number of system utterances from all dialogues with that model.

Results

As for the GraphWOZ results, the human interaction experiments indicate a reduction in the proportion of turns with hallucinations or retrieval errors, as shown in Table 6. This reduction is observed for both model types when the logic-enhanced relevance scoring model was used.

The participant scores in Table 5 ranked the models which used the relevance scored facts highest. While the GODEL model using the relevant facts scored higher in appropriateness, the opposite pattern is observable in the GPT models, although the scores remain relatively high. As not every participant interacted with every model, differences in scoring between individual participants cannot be discounted as a factor impacting these results, thus a larger study would be beneficial.

5 Conclusion

This paper presented a novel approach to retrieval-augmented response generation in task-oriented dialogue systems. The approach relies a dynamic knowledge graph representing the dialogue state, which is enriched at each turn with facts derived from a small set of rules specified in the ProbLog language. Those facts are then ranked by relevance using a dedicated scoring model which accounts for both the semantic similarity and conversational saliency of each fact. The most relevant facts are then incorporated to the background knowledge provided as input to the response generation model.

We provide experimental results showing that the combination of logical reasoning with a relevance scoring model leads to more factual responses. In particular, the logical rules seem to assist the generation model's ability to provide responses grounded in multi-step reasoning based on the available background knowledge.

The proportion of errors remains, however, relatively high, likely due to the very limited number of dialogues available for training in GraphWOZ and KVRET. Future work will focus on evaluating the potential of this approach in other (and potentially broader) dialogue domains.

References

Giovanni Bonetta, Rossella Cancelliere, Ding Liu, Paul Vozila, et al. 2021. Retrieval-augmented transformer-xl for close-domain dialog generation. In *The International FLAIRS Conference Proceedings*, volume 34, pages 1–4. Florida Online Journals.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, Hong Kong, China. Association for Computational Linguistics.

Mark Chavira and Adnan Darwiche. 2008. On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6-7):772–799.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Daan Fierens, Guy Van den Broeck, Joris Renkens, Dimitar Shterionov, Bernd Gutmann, Ingo Thon, Gerda Janssens, and Luc De Raedt. 2015. Inference and learning in probabilistic logic programs using weighted boolean formulas. *Theory and Practice of Logic Programming*, 15(3):358–401.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Bernd Gutmann, Ingo Thon, and Luc De Raedt. 2011. Learning the parameters of probabilistic logic programs from interpretations. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, pages 581–596. Springer.

Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zhengyu Niu, and Haifeng Wang. 2021. Learning to select external knowledge with multi-scale negative sampling. *arXiv preprint arXiv:2102.02096*.

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.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Angelika Kimmig, Bart Demoen, Luc De Raedt, Vitor Santos Costa, and Ricardo Rocha. 2011. On the implementation of the probabilistic logic programming language problog. *Theory and Practice of Logic Programming*, 11(2-3):235–262.

Hayoung Lee and Okran Jeong. 2023. A knowledge-grounded task-oriented dialogue system with hierarchical structure for enhancing knowledge selection. *Sensors*, 23(2):685.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

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

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019b. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792, Hong Kong, China. Association for Computational Linguistics.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. In *International Conference on Learning Representations*.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8697–8704.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cameron Smith, Nigel Crook, Daniel Charlton, Johan Boye, Raul Santos De La Camara, Markku Turunen, David Benyon, Björn Gambäck, Oli Mival, Nick Webb, et al. 2011. Interaction strategies for an affective conversational agent. *Presence*, 20(5):395–411.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*.

Jonas Vlasselaer, Joris Renkens, Guy Van den Broeck, and Luc De Raedt. 2014. Compiling probabilistic logic programs into sentential decision diagrams. In *Proceedings Workshop on Probabilistic Logic Programming (PLP)*, pages 1–10.

Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. 2022. Graphwoz: Dialogue management with conversational knowledge graphs. *arXiv preprint arXiv:2211.12852*.

Hao Wang, Bin Guo, Wei Wu, Sicong Liu, and Zhiwen Yu. 2021. Towards information-rich, logical dialogue systems with knowledge-enhanced neural models. *Neurocomputing*, 465:248–264.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022a. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022b. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2019a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. *arXiv preprint arXiv:1911.02707*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2021. Joint retrieval and generation training for grounded text generation. *arXiv preprint arXiv:2105.06597*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Entity linking rules

The ProbLog rules employed for entity linking are given below. The probabilities attached to the rules are estimated empirically using the parameter estimation approach implement in ProbLog library, based on Learning from Interpretations (Gutmann et al., 2011).

```
0.60838635::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), jw_similarity(N,S,O), O>0.9.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), jw_similarity(N,S,O), O>0.8.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), jw_similarity(N,S,O), O>0.7.

0.72255423::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), lev_distance(N,S,O), O < 2.

0.30394455::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), lev_distance(N,S,O), O < 3.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), lev_distance(N,S,O), O < 6.

0.0019686::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), lcs(N,S,O), O > 3.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), string(M,S), lcs(N,S,O), O > 6.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), nb_common_words(N,S,O), O > 0.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), nb_common_words(N,S,O), O > 1.

0::refers_to(M,E) :- new(U), mention(U,M), string(M,S),
    is_processable_time(S,0), name(E,N), nb_common_words(N,S,O), O > 2.

0.27142172::refers_to(M,E) :- new(U), mention(U,M), respond_to(U,AR1),
    mention(AR1,PM1), refers_to(PM1, E).

0.12752306::refers_to(M,E) :- new(U), mention(U,M), respond_to(U,AR1),
    respond_to(AR1,PU1), mention(PU1,PM1), refers_to(PM1, E).

0.07429096::refers_to(M,E) :- new(U), mention(U,M), respond_to(U,AR1),
    respond_to(AR1,PU1), respond_to(PU1,AR2), mention(AR2,PM1), refers_to(PM1, E).

0.01403269::refers_to(M,E) :- new(U), mention(U,M), respond_to(U,AR1),
    respond_to(AR1,PU1), respond_to(PU1,AR2), respond_to(AR2,PU2),
mention(PU2,PM2), refers_to(PM1, E).
```

A.2 Commonsense rules

The rules employed for commonsense reasoning on the GraphWOZ dialogues are provided below.

```
event_today(E,T) :- event(E), start_time(E,T), date(at_today,D), date(E,D).

event_tomorrow(E,T) :- event(E), start_time(E,T), date(at_tomorrow,D), date(E,D).

person_group(P,G) :- people(P), group(P,G).

group_members(G,L) :- group(G), findall(P, person_group(P,G), L).

count_members(G,N) :- group(G), refers_to(M,G), group_members(G,L), list_length(L,N).
```

```
room_available_today(R,T) :- room(R), \+(location(E,R), date(E,D), date(at_today,D),
    start_time(E,ST), end_time(E,ET), time_between(T,ST,ET,1)).

room_available_tomorrow(R,T) :- room(R), \+(location(E,R), date(E,D),
    date(at_tomorrow,D), start_time(E,ST), end_time(E,ET), time_between(T,ST,ET,1)).

room_available_now(P) :- room(P), \+(room_busy_now(P)).

room_busy_now(P) :- room(P), time(at_now,T), attendee(E,P), date(E,D),
    date(at_today,D), start_time(E,ST), end_time(E,ET), time_between(T,ST,ET,1).

person_available_today(P,T) :- refers_to(M,P), people(P),
    string(_,T), is_time_expression(T,1), \+(person_busy_today(P,T)).

person_busy_today(P,T) :- refers_to(M,P), people(P), string(_,T),
    is_time_expression(T,1), attendee(E,P), date(E,D), date(at_today,D),
    start_time(E,ST), end_time(E,ET), time_between(T,ST,ET,1).

person_available_tomorrow(P,T) :- refers_to(M,P), people(P), string(_,T),
    is_time_expression(T,1), \+(person_busy_tomorrow(P,T).

person_busy_tomorrow(P,T) :- refers_to(M,P), people(P), string(_,T),
    is_time_expression(T,1), attendee(E,P), date(E,D), date(at_tomorrow,D),
    start_time(E,ST), end_time(E,ET), time_between(T,ST,ET,1).

person_available_now(P) :- refers_to(M,P), people(P), time(at_now,T),
    \+(person_busy_now(P)).

person_busy_now(P) :- refers_to(M,P), people(P), time(at_now,T), attendee(E,P),
    date(E,D), date(at_today,D), start_time(E,ST), end_time(E,ET),
    time_between(T,ST,ET,1).

attending_today(E,P) :- attendee(E,P), date(E,D), date(at_today,D).

person_events_today(P,L) :-  refers_to(M,P), people(P),
    findall(X, attending_today(X,P), L).

attending_tomorrow(E,P) :- attendee(E,P), date(E,D), date(at_tomorrow,D).

person_events_tomorrow(P,L) :-  refers_to(M,P), people(P),
    findall(X, attending_tomorrow(X,P), L).

available_rooms_now(L) :- findall(R, room_available_now(R), L).

available_rooms_today(L,T) :- string(_,M), morning_time(M,1), between(8,11,T),
    findall(R, room_available_today(R,T), L).

available_rooms_tomorrow(L,T) :- string(_,M), morning_time(M,1), between(8,11,T),
    findall(R, room_available_tomorrow(R,T), L).

available_rooms_today(L,T) :- string(_,M), afternoon_time(M,1), between(12,17,T),
    findall(R, room_available_today(R,T), L).

available_rooms_tomorrow(L,T) :- string(_,M), afternoon_time(M,1), between(12,17,T),
    findall(R, room_available_tomorrow(R,T), L).

time_place(E,D,T) :- refers_to(M,E), event(E), date(E,D), start_time(E,T).
```

Toward Open-World Human-Robot Interaction: What Types of Gestures Are Used in Task-Based Open-World Referential Communication?

Mark Higger, Polina Rygina, Logan Daigler,
Lara Ferreira Bezerra, Zhao Han, Tom Williams
MIRRORLab, Department of Computer Science
Colorado School of Mines, Golden CO 80401

Abstract

Gestures play a critical role in human-human and human-robot interaction. In task-based contexts, deictic gestures like pointing are particularly important for directing attention to task-relevant entities. While most work on task-based human-human and human-robot dialogue focuses on closed-world domains, recent research has begun to consider open-world tasks, where task-relevant objects may not be known to interactants a priori. In open-world tasks, we argue that a more nuanced consideration of gesture is necessary, as interactants may use gestures that bridge traditional gesture categories, in order to navigate the open-world dimensions of their task environment. In this work, we explore the types of gestures used in open-world task contexts, and their frequencies of use. Our results suggest a need to rethink the way that gesture analysis is approached in the study of human-human and human-robot interaction.

1 Introduction

For task-based human-robot interaction (HRI), effective communication can greatly increase task effectiveness (Cantrell et al., 2011; Tellex et al., 2020). Critically, this includes both verbal and non-verbal communication (Mavridis, 2015). Effective non-verbal communication can include both gaze and gesture (De Angeli et al., 1998), and are fundamental in human-human communication and substantially contribute to fluent communication (Kendon, 1997; Kita, 2003; Goldin-Meadow, 1999; Ping and Goldin-Meadow, 2010). Accordingly, understanding and generating these non-verbal communication modalities is critical for effective human-robot communication.

Recently, work on natural language understanding in robotics (Han et al., 2022; Han and Williams, 2022; Culpepper et al., 2022; Williams and Scheutz, 2015b) has been expanding beyond traditional tabletop domains to consider open-world contexts. In contrast to a closed-world context where there is an assumption that all possible referents have an a-priori representation in the robot's system, open-world contexts have referents which may be new or unknown to robot ahead of time. For example, if a nurse asks an assistive robot "Go to the kitchen to find a glass of water, then bring it to the patient". When trying to figure out what "it" means in this sentence the robot needs to be aware of the glass of water. However if the robot is using a vision-based system to ground referring language it may only contain a representation of the objects currently visible in the room. In order to ground "it" to a particular object, the robot may need to create a new representation for the glass of water outside of the vision system. By allowing for new representations to be created outside of what the robot is already aware of (in this case through vision), it can then allow for understanding reference in an open-world context. Work in this area has led to a number of algorithms for *open-world reference resolution* (Williams and Scheutz, 2015b; Williams et al., 2016; Culpepper et al., 2022) to allow for this type of behavior.

While there is work on linguistic grounding in open-world contexts, work on robot gesture still largely assumes a closed world where gestural targets are visible, known, and close-by (Lücking et al., 2015; Sauppé and Mutlu, 2014). We argue that this has led to an overly narrow focus on understanding and generating specific, narrow classes of gestures in human-robot interaction; i.e., deictic gestures that focus an interactant's attention on a visible, nearby area of the task environment. Based on this argument, we analyze the gestures used in a recent corpus of human-human interactions, collected in a novel task environment designed by Han et al. (2022) to elicit a more ecologically valid range of referring forms. As we will show, our analysis of the gestures used in this task context backs up our argument, yielding a novel taxonomy

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

of gestures used in open-world dialogue contexts, and suggesting a need for human-robot interaction researchers to fundamentally rethink the types of gestures they are attempting to model in task-based human-robot interaction.

2 Related Work

2.1 Open World Communication

To understand the ways in which gesture in current task-based HRI are overly limited, let us first consider the linguistic work that specifically targets open-world interactions. While task-based natural language understanding and generation in human-robot interaction has traditionally considered only closed-world environments. There has recently been an increase in research relaxing this closed-world assumption to consider *open worlds*, especially in the context of reference resolution (Williams and Scheutz, 2015b; Williams et al., 2016; Culpepper et al., 2022).

Reference resolution is the process of identifying what knowledge associated with particular entities in a robot's memory is being referred to by a speaker's referring language. While traditional approaches to reference resolution (and the related process of language grounding) have only attempted to associate incoming referring expressions with pre-existing knowledge representations, work on open-world reference resolution has additionally modeled how listeners might assess when an incoming referring expression is likely to refer to a previously unknown entity, and how a new knowledge representation might be created in such cases through the process of hypothesization (Williams and Scheutz, 2015a,b). More recently, Culpepper et al. (2022) presented a novel algorithm that allows for *incremental* reference resolution, which allows for a real-time word-by-word processing within this type of open-world framework.

While there has been less work on open-world language generation, this too is starting to change. Han et al. (2022), for example, recently presented a novel experimental setup designed to capture natural references to both visible, previously visible, and not yet visible referents, in order to develop computational referring form selection models that can handle these sorts of open world references. Like prior work on open-world reference resolution, however, this work has been solely focused on natural language generation, and has not attempted to account for the role that nonverbal communication, especially gesture, plays in open-world dialogue.

2.2 Gesture in Human-Robot Interaction

Because gestures are known to be a fundamental part of human communication (Kita, 2003; Goldin-Meadow, 1999), the use of gesture to enhance human-robot interaction has attracted significant attention across the history of human-robot interaction (Waldherr et al., 2000). Effective gesture has been shown both to promote sociability and interactions, making robots more natural and enjoyable to work with (Kim et al., 2013; Salem et al., 2012); and to enhance the effectiveness and productivity of task-based interactions (Gleeson et al., 2013; Gross et al., 2017). Because of the important role gestures play in human-robot interaction, HRI researchers have devised a number of taxonomies for categorizing the different types of gesture that can be used – and understood – by robots. Many of the existing taxonomies used for gesture generation and understanding for HRI research primarily focuses on physical arm motions (Allwood et al., 2007; Dael et al., 2012). These taxonomies are often very complex: the BAP taxonomy (Dael et al., 2012), for example, has nearly 40 non-mutually exclusive codes for gesture, with complex and specific codes such as "Left arm action curved repetition" or "Asymmetrical arms action". For understanding what physical movements are needed to create with robots these taxonomies can be very helpful, however these taxonomies are often not grounded with the intent of the gesture in mind. Additionally, research in the semiotics community (Goodwin, 2003) shows that physical manifestation is not always a good indicator of the intent of the gesture, as context and subtle indicators also play a large role in gestural meaning. This means that despite their comprehensive nature in terms of physical motion, these large-scale categorization systems fail to account for the ways that different types of gestures are typically used to achieve different types of communicative purposes in human-robot interactions.

In contrast, the taxonomy devised by McNeill and Levy (1982) for use in human-human interaction categorizes gestures according to a small number of conversational roles: deictic, iconic, metaphoric, beat, and emblematic. Deictic gestures, like pointing, direct an interactants's visual attention to a particular object or location. Iconic gestures mime the physicality of an object or action to direct an interactant's internal attention to a particular concept. Metaphoric gestures are used to convey more abstract concepts such as time. Beat gestures are used for pacing or timing of linguistic structure. Emblematic gestures (e.g. a wave or thumbs-up) have distinct meanings derived through social and cultural context. While, this taxonomy for categorizing gestures (McNeill and Levy, 1982) and their incorporation into language-based communication (McNeill, 1985) has been widely adopted across the fields of psycholinguistics and human-robot interaction (de Wit et al., 2022) it lacks the specificity, and grounding to physical movement for interpretability that is often required, especially when it comes to the wide variety of gestures that fit under the broader umbrella of "deictic gestures".

Within the domain of human-robot interaction, there is wide recognition that understanding and generation of non-verbal communication is critical for situated interaction (Cantrell et al., 2011; Breazeal et al., 2005; Mavridis, 2015). While there has been work on all of the gestural categories described by McNeill and Levy (1982), deictic gesture in particular has attracted significant attention, due to its highly task-oriented and more easily interpretable nature. Indeed, in task-based human-robot interaction, there is often an exclusive focus on understanding and generating *deictic* gestures. In the foundational work of Sauppé and Mutlu (2014), a wide variety of subtypes of deictic gestures are studied, including pointing, presenting, touching, exhibiting, grouping, and sweeping, with each category defined according to the physical motion of the gesture. For example, a 'touching' gesture requires direct physical contact with the referent, while a 'sweeping' gesture used wide arm movements to direct attention to larger regions. Work on computational understanding and generation of deixis typically models deictic gestures by projecting a saliency cone from the origin of a deictic pointing gesture (Kranstedt et al., 2005) outwards in a particular direction. These cones can then be used in a multi-modal estimator that combines grounded language and non-verbal communication to identify which potential referent is most likely given the particular language and gestures used (Schauerte et al., 2010; Schauerte and Fink, 2010; Lücking et al., 2015).

Yet critically, these methods assume that the target of a deictic gesture is visible and known to the robot so it can find the most likely target that falls within the deictic cone. While this seems reasonable at first glance, it is clear that humans frequently use gestures to refer to objects that cannot be seen or may not be known to the other person they are gesturing to. A simple example of this is the way that iconic gestures help to draw users' *internal* attention to a target referent representation, rather than drawing users' *visual* attention to a target referent stimulus. Moreover, some researchers (Stogsdill et al., 2021; Enfield et al., 2007) have recently begun to explore types of gestures that arise in large-scale and open-world environments, which seem to further trouble the boundaries between these categories of gesture.

2.3 Gesture in Large-Scale and Open Worlds

One example of how the boundaries between traditional gesture categories are being troubled and contested within psycholinguistics is the work by Enfield et al. (2007). They examined how general "pointing" gestures could have a more complex meaning than what is traditionally associated with deictic gestures. Specifically, Enfield et al. looked at how pointing was used during conversation in small Laotian villages and classified pointing gestures into two types; primary Big (B) and secondary Small (S) points. The B-points were composed of large arm movements, while S-points were smaller, single-armed gestures that had more complex hand movements. While B-points seemed to be used in the way typically expected of deictic gestures (i.e., to point to a physical location in space), S-points were instead found to have more complicated usage. Sometimes the S-points were used similarly to deictic gestures specifying a physical object or person which would be ambiguous through language alone. But other times, the S-Points were also used to refer to locations that were not in the current view while not necessarily pointing in the exact location of that object.

These S-points thus trouble McNeill and Levy (1982)'s conception of deictic gestures, in which the objective of a pointing gesture is to direct attention to a physical location via spatial information. The notion of "abstract deictic" gestures are explored in McNeill et al. (1993), which looks at how deictic gestures can be used to point to objects or people which are not physically there. While McNeill et al. (1993)'s "abstract deictic" gestures are used to point to entities without a physical presence, they still retain spatial information within a narrative context, being used to denote relative position in a non-present space. However, some of the behaviors present in Enfield et al. (2007)'s S-point's includes pointing gestures which do not necessarily contain direct spatial information, but instead are used to create a gestural representation of an entity, present or not. This suggests there is non-spatial information that can also be conveyed by pointing. While this use of non-spatial pointing is explored within the semiotics community (Enfield et al., 2007; McNeill, 2003), these gestures are still often classified as "Deictic gestures", despite their lack of spatial information making them inherently non-deictic. Perhaps this is due in part to the grounding of McNeill's categories in studies of conversational rather than task-based dialog. We argue that because of the lack of spatial information, a non-spatial pointing gesture does not fit cleanly into McNeill's "Deictic gesture" category. As such, this suggests a need to our work that studies task-based gestures in order to better understand the use of gestures in complex, large-scale, open-world environments.

These troubled category boundaries have also been recently noted in work in the HRI community. Specifically, Stogsdill et al. (2021) explored the use of vague, non-deictic pointing gestures that are very similar to the examples of S-Points found in Enfield et al. (2007). For example, if someone is trying to refer to the room next door, they may vaguely wave toward the room in question without pointing precisely or directing their gaze. Or, if someone is referring to another faraway city, they may point in a completely random direction. This use of an abstract pointing gesture removes the spatial information that is central to McNeill et al. (1993)'s deictic gestures. Instead, Stogsdill et al. (2021) argues that this gesture attempts to merely convey the concept of "away", which may be closer to a metaphoric gesture than a deictic gesture.

This spectrum of category-spanning gestures remains understudied, yet critically points to a need to reconsider the categories of gesture used in task-based interaction. More specifically, in order to enable open-world Human-Robot Interaction, we argue that there is a fundamental need to question how and where existing approaches to gesture understanding and generation might fall short, and how these shortcomings might be grounded in the general taxonomies used by roboticists and psycholinguists to make sense of gesture.

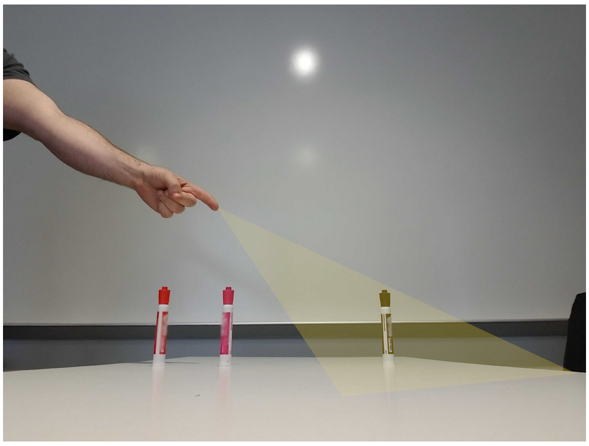


Figure 1: Example of the conical deictic projection often found in computer-vision-based deictic gesture representation. As can be seen, the gesture becomes less precise as the target becomes farther away from the source of the gesture.

In this work, we thus propose to investigate the following research questions:

R1: How might we better categorize the types of referring gestures used in open-world task-based environments?

R2: How prevalent are these different categories of gestures?

R3: Is the overwhelming focus of the HRI research community on precise, deictic, *B-point* gestures justified by this distribution of observed gestures?

3 Methodology

3.1 Dataset

To evaluate our research questions, we analyzed the experimental data from the experiments conducted by Han et al. (2022). We will briefly describe the context in which Han et al. collected that data, as reported in Han et al. (2022), to explain why this was an ideal dataset for answering our research questions. The task environment used in Han et al. (2022)'s work (shown in Figure 2) was partitioned into four quadrants, each containing a variety of colored blocks. In their experiment, pairs of participants (an *instructor* and a *learner*) participated in a sequence of four building tasks, one in each quadrant, in which the *instructor* taught the *learner* how to construct a different building from those blocks. Specifically, Han et al. (2022) designed their experiment so that each of these four building tasks required blocks not available within the task's quadrant, including blocks seen in previous quadrants or located in quadrants that were to be visited in the future. This task structure was thus inherently open-world in nature; the instructor was required not only to refer to blocks that were immediately visible, but also to refer to blocks that had been seen in previous quadrants, as well as blocks whose locations were as-yet-unknown. By having discreet sections with items that participants were aware of and other sections they not aware of, they established quadrants that belonged to a current, closed working context and quadrants that belonged to an open-world. While Han et al. (2022) report designing this experiment to analyze open-world *language* production, we realized that their dataset could also be a rich source of open-world *gesture* production.

We analyzed twelve videos from Han et al. (2022)'s video dataset, a total of 337 minutes with an average of 28 minutes per video. From these videos, we identified 1067 gestures in total and an average of 89 gestures per video. In the next section, we will thus describe how we analyzed the gestures found in these videos.

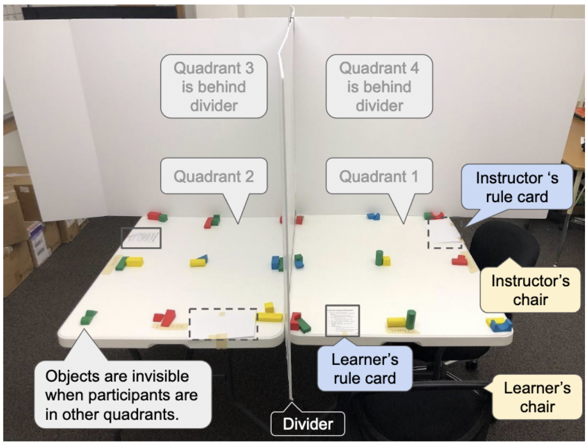


Figure 2: Setup for testing referring form selection from "Evaluating Referring Form Selection Models in Partially-Known Environments" (Han et al., 2022)

3.2 Qualitative analysis and gesture coding

To analyze those videos, we performed an iterative qualitative coding procedure. First, we began by breaking each video into a set of *communicative action segments* in which discrete gestures were used. Next, we performed open coding to identify, for each gesture, the (1) physical manifestation of the gesture and (2) information conveyed by the gesture. Then, we removed from consideration any gestures that did not appear to be related to the task itself, or which did not appear to provide the interactant with meaningful referential information to help them accomplish the task, such as beat gestures or emblematic gestures. Finally, we analyzed the literature on gesture classification to cluster gestures into discrete categories that were informed by prior literature whenever possible, while also striving to ensure (1) clear boundaries between gesture clusters, and (2) clear criteria for assignment of gestures to clusters.

Once a gesture taxonomy was identified, coders were used to identify when in the videos particular gestures were used. Then 17% of communicative action segments were used to establish an intercoder reliability rating (IRR) using Cohen's Kappa Coefficient (Cohen, 1960) of $\kappa = 0.35$, denoting fair agreement. After an IRR was established, a single coder was used for the remaining segments. In the next section, we will describe the taxonomy of gestures that resulted from this qualitative coding procedure.

4 Gesture Taxonomy

Based on our qualitative analysis, we formulated the following taxonomy of open-world task-based gestures which strike a balance between the intent of the gesture while allowing for the visually separable physical interpretability needed for robotics. This taxonomy is comprised of five key categories: (1) Precise Deictic Gestures, (2) Small Region Deictic Gestures, (3) Large Region Deictic Gestures, (4) Abstract Pointing Gestures, and (5) Iconic Reference Gestures. Summary statistics for the use of these gestures are shown in Table 1.

Gesture Type	Count	pct of total
Precise Deictic	391	36.6%
Small Region Deictic	434	40.6%
Large Region Deictic	13	1.2%
Abstract Point	103	9.7%
Iconic Reference	126	11.8%

Table 1: Number of gestures observed in the analyzed dataset

4.1 Precise Deictic Gesture

We identified three distinct types of deictic gestures that manifest in open-world task-based interactions.

The first category was what we term "Precise Deictic Gestures". These gestures were those most closely related to the traditional conceptualization of Deictic Gesture, and were relatively common (36.6% of all analyzed gestures) due to their important role in specifying which block will be needed next for the experiment. The apparent purpose of this gesture was to physically direct attention to a single target object with a high level of specificity. This type of gesture physically manifested either as touching that object, or as pointing to that object in a way where it was the only task-relevant object entirely within the saliency cone extending from the gesturer's hand (cf. Lücking et al., 2015; Schauerte et al., 2010). Thus, while the physical motions of precise deictic gestures are dependent on the environement, for example the gesturer may need to put their finger closer to an object if it it is near another object, the purpose of precise decictic gestures is maintained.

This category of gesture captures gestures that Sauppé and Mutlu (2014) would have categorized as *exhibiting, touching*, and *presenting*, as well as the single-target subset of those gestures Sauppé and Mutlu would have categorized as *pointing*. The difference in our categorization schemes is thus grounded in a difference in focus on *motion* vs *purpose*. That is, while Sauppé and Mutlu primarily focuses on observable differences in how gestures are physically executed, we instead primarily focus on differences in what gestures are intended to achieve, while still taking into account the contextual information that the physical motion brings.

By analyzing these gestures in terms of purpose rather than motion alone, we can understand both when and why these gestures are used. In cases where the gesturer's purpose is to direct attention to a single object, they generate a precise gesture in which only the target appears in the cone, because (cf. Schauerte and Fink (2010)) if the gesture were less precise, and multiple objects fell into the cone, then additional effort would be needed to further pick out the object through other channels such as language.

4.2 Small Region Deictic Gesture

The second category of Deictic gesture we observed was what we term "Small Region Deictic Gestures" which accounted for a plurality (40.6%) of all gestures observed. The apparent purpose of this gesture was to direct attention toward a small group of objects, either due to an intention to pick out the blocks as a group, or due to an intention to pick out a single object within the group, without certainty as to which object should be attended to. This type of gesture physically manifested as pointing toward the general area containing those objects, so that all objects in the group were entirely within the saliency cone extending from the gesturer's hand (cf. Lücking et al., 2015; Schauerte et al., 2010).

This category of gesture captures gestures that Sauppé and Mutlu (2014) would have categorized as *grouping gestures*, as well as the multi-target subset of those gestures Sauppé and Mutlu would have categorized as *pointing*. As above, then, the difference in our categorization schemes is grounded in a difference in focus on *motion* vs *purpose*.

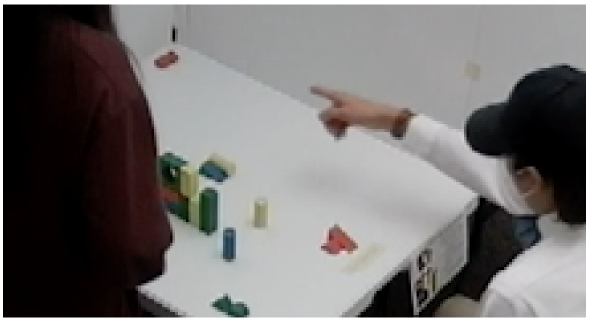


Figure 3: "Precise Deictic Gesture": These are direct and unambiguous spatial gestures to a target referent. This figure shows the participant directly pointing at a block, demonstrating a "Precise Deictic Gesture" to directly specify a red rectangle as the referent of the gesture.

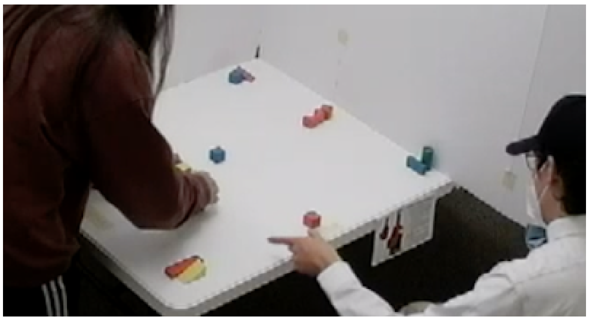


Figure 4: "Small Region Deictic Gesture": These are direct and spatial gestures to a target referent or referents, but do not have a clear unambigous target. This figure shows the participant using a "Small Region Deictic Gesture" to point in the direction of a red rectangle and a yellow triangle, where the exact target referent of the gesture underspecified by the gesture without additional context or linguistic accompaniment.

4.3 Large Region Deictic gesture

The final category of obviously Deictic gesture we observed what we term "Large Region Deictic Gestures". These gestures were very rare (1.2%), and were only used by some participants. Yet they were distinct enough in purpose and form to warrant separate consideration. The apparent purpose of this gesture was to direct attention to a large number of objects comprising multiple clusters, or to a large, general region of the task environment. This type of gesture physically manifested as a large, potentially full-arm gesture in the direction of the objects of interest, but without attempting to fit those objects into a saliency cone.

This category of gesture captures gestures that Sauppé and Mutlu (2014) would have categorized as grouping or sweeping gestures. Our categorization is more broad, however. Because we focus on on purpose rather than motion, we do not restrict this category to those that manifest as literal whole-arm sweeps, but rather include any gesture whose *intent* is to highlight a large region. For example, we include observed instances in which a speaker waves their hand across multiple clusters of blocks.

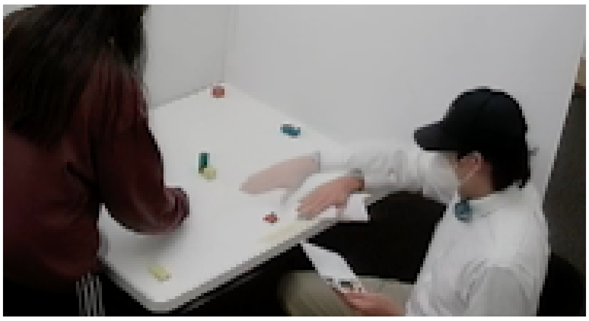


Figure 5: "Large Region Deictic gesture": These are large spatial gestures used to refer to many target referents. This figure shows the participant sweeping their hand over the top of the table demonstrating a "Large Region Deictic gesture" to refer to the red cube, the yellow cube, and the green cylinder.

4.4 Abstract Pointing Gesture

The next category of gesture we observed does not clearly fit into deictic gestures, nor does it clearly fit into another category, like metaphoric gesture. Rather, it represents a spectrum of gestures that fall somewhere between these categories. We term these gestures "Abstract Pointing Gestures". While not as common (around 9.7% of observed gestures) as deictic gestures, these gestures were consistently used at least once by all participants whose data was analyzed.

The apparent purpose of this gesture was to indicate that a target referent was "elsewhere", and *possibly* also to convey the direction in which the target referent was to be found. This gesture manifested in a variety of ways. In some cases, the gesture manifested as a point in the vague region of the target object, or in the direction of where the listener would need to go in order to begin traveling to the target. In either case, the gesture could be construed as casting an incredibly wide deictic cone; but we believe that the lack of precision in the gesture suggested that the speaker did not have a genuine expectation that the listener would follow their gaze, limiting the utility of modeling such a gesture as a cone.

Another reason for distinguishing this type of gesture from deictic gestures is other ways this could physically manifest that did not appear in this task-based dataset is due to the close distances between the speaker and all target referents, but which could manifest in other tasks with more varied out-of-context environments. A speaker trying to refer to something in another room, might wave generally or emblematically jerk their thumb over their shoulder. We would view these gestures as falling along the continuum of abstract gestures due to their shared intent.

This category of gesture captures those discussed by Stogsdill et al. (2021), McNeill et al. (1993)'s abstract deictic gestures, and some of the S-point gestures observed by Enfield et al. (2007). Our characterization of these gestures differs from that of Stogsdill et al. (2021), McNeill et al. (1993), and Enfield et al. (2007), however, in that we characterize them within a referential context, and ground them relative to other referential gestures within a comprehensive taxonomy. Additionally, they are grounded through our analysis of experimental data demonstrating how they are used in open-world task-based environments.

4.5 Iconic Reference Gesture

Thus far, we have been discussing gestures that are either clearly deictic, or that appear deictic in physical manifestation. We will now describe *iconic* gestures that we term "Iconic Reference Gestures", which we observed to also play a key role in referential communication within the open-world task-based interactions that we analyzed. This category

Figure 6: “Abstract Pointing Gesture”: These are non-spatial pointing gestures used to create a gestural representation of a referent. This figure shows a participant using an “Abstract Pointing Gesture”, pointing away from the table to refer to a block which has not seen before.

Figure 7: “Iconic Reference Gesture”: These are non-spatial gestures used to mimic properties of the target referent. This figure shows the participant using an “Iconic Reference Gesture” to refer to a rectangular block by making a rectangular shape with their hands.

comprises a subset of the iconic gestures as delineated by McNeill and Levy (1982). However, we believe they are worth highlighting here as a separate category due to the referential purposes they achieve.

These gestures accounted for over one-tenth (11.8%) of the total gestures used. The apparent purpose of this gesture was to provide semantic content regarding the referent, to help disambiguate the semantic content of the speaker’s speech. As such, this type of gesture physically manifested as mimicry of the shape of a referenced block to make the shape of a block (such as a rectangle, or semi-circle), or tracing out the shape of the block in the air with an index finger.

These gestures play an especially important role in open-world communication. In fact, these gestures were the most common method we observed when speakers intended to refer to objects that were not currently visible. Despite the observed priamcy of these gestures for open-world task-based reference, these iconic gestures are not well studied in task-based HRI.

5 Discussion

5.1 What types of referring gestures did we observe?

Our first research question was “How might we better categorize the types of referring gestures used in open-world task-based environments?” Our results show that a variety of different referring gestures are used in open-world task-based environments, including multiple types of deictic gestures, iconic gestures, and abstract gestures that trouble the previously delineated boundaries of traditional gesture categories.

As demonstrated above, our results suggest that deictic gestures may be best split into three categories, based on the intended specificity of the gesture. This suggests a need to shift from a focus on the physical form of different types of deictic gestures, to the way that speaker intent shapes gesture specificity.

Our results also highlight the need to consider iconic gestures when analyzing referring gestures. As demonstrated above, participants used a variety of iconic gestures to help communicate the properties of referents. We argue that these types of gestures are uniquely important to open-world task-based interactions, both to help describe the properties of previously-seen or as-yet-unseen objects – or to signal through the use of iconic gesture that those objects are not currently visible.

Finally, our results demonstrate the importance of Abstract Pointing Gestures; how these gestures are uniquely used in open-world task-based environments; and how imprecision and abstractness serve as tools to communicate this open-world status. These gestures, which do not fit cleanly into traditional gesture categories, demonstrate a need to think differently about gestures in open-world task-based environments, in a way that moves beyond traditional frameworks for categorizing gestures.

5.2 With what prevalence were different referring gestures observed?

Our second research question was “How prevalent are these different categories of gestures?”, Table 1 shows the distribution of how the gestures observed in the analyzed dataset. While “Deictic Precise

Gestures" are nearly the only type of referring gesture explored in the task-based human-robot interaction literature, they comprised only one-third of the gestures we observed (36.6%), making them only the second most common gesture type observed. In contrast "Deictic Small Region Gestures" were the most common gesture, constituting nearly half of observed gestures (40.6%), and one-fifth (21.5%) of gestures used were non-deictic gestures. This troubles the dominant perspective that nearly exclusively focuses on precise deictic gestures.

5.3 Is the HRI community's focus on Precise Deictic Gestures warranted?

Our final research question was "Is the overwhelming focus of the HRI research community on precise, deictic, B-point gestures justified by this distribution of observed gestures". Based on the types of gestures observed in the experiment, and their frequency of observation, we believe this overwhelming and myopic focus on precise deictic gestures is not justified. While the 37% frequency of use for "Precise Deictic Gestures" represents a meaningful quantity, it does not justify near-exclusive focus. These results suggest that the HRI research community should dramatically expand its scope of work to consider a wider variety of referring gestures.

Specifically, we recommend focusing on narrowing down the broader category of deictic gestures into different categories of deixes based on target specificity, and to explore the use of non-deictic gestures and the role they play in open-world referential communication. Similarly, while saliency cones are a good way to detect the target of a pointing gesture, as we start to incorporate non-pointing gestures, it is no longer sufficient. Understanding and generation of gestures which cannot be modeled by a saliency cone may be difficult, but it is also necessary for effective communication the open-world tasks that are prevalent within HRI. Overall, we suggest the HRI community should move away from the coarse, generic gesture categorization schemes they have relied on for so long.

6 Conclusion

Clear communication is critical for both human-human and human-robot task-based interaction; and clear communication in many task-based domains requires effective referential gestures. While current gesture research in task-based HRI is nearly exclusively focused on precise deictic gestures, as we transition to more realistic open-world tasks, we will need to move beyond these gestures.

Our results show that precise deicitic gestures only play a small part in task-based communication, suggesting that HRI research perhaps needs to be "pointed" in a new direction for both gesture generation and gesture understanding. We propose a new gesture taxonomy that can refocus the conversation about gesture to include gestures that are more suitable for open-world environments. We argue that this taxonomy is more effective at capturing the types of gestures used in tasks for both open-world and closed-world environments.

Limitations

While our research shows the need to reevaluate gesture categorization, our results were limited due to the constraints of the experiment. One major limitations was the difficulty of seeing the direction of deictic gestures from the camera's perspective, and thus, of differentiating "Precise Deictic Gestures", "Small Region Deictic Gestures" and sometimes "Abstract Pointing Gestures". Another major limitation was that certain gestures may have been under- or over-represented in the dataset we analyzed. Specifically, large deictic gestures may have been under-represented due to the task requiring only a single item at time, reducing the need for simultaneous reference for multiple blocks.

Ethics Statement

While this research itself may not have overt ethical implications, work on gesture understanding and generation does present ethical implications. Specifically, work on gesture in robotics brings along risks to privacy via perception, and risks of over-trust due enhanced anthropomorphic morphology. A major requirement for gesture generation and understanding is advanced capabilities in robotic perception. These perception methods bring along risks of data privacy and security as it can be unclear if and how perceptual data can be used during and after the robotic interaction. Additionally generation of gestures enhances the anthropomorphic morphology of the robot, which can lead to a higher and potentially false perception of the robot's intelligence. This can lead to an overtrust in the robot's intelligence and capabilities, and can lead to potentially dangerous outcomes (Robinette et al., 2016).

Acknowledgements

This work has been supported in part by the Office of Naval Research grant N00014-21-1-2418.

References

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41:273–287.

Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*, pages 708–713. IEEE.

Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. 2011. Learning actions from human-robot dialogues. In *2011 RO-MAN*, pages 125–130. IEEE.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Will Culpepper, Thomas A Bennett, Lixiao Zhu, Rafael Sousa Silva, Ryan Blake Jackson, and Tom Williams. 2022. Ipower: Incremental, probabilistic, open-world reference resolution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Nele Dael, Marcello Mortillaro, and Klaus R Scherer. 2012. The body action and posture coding system (bap): Development and reliability. *Journal of Nonverbal Behavior*, 36:97–121.

Antonella De Angeli, Walter Gerbino, Giulia Cassano, and Daniela Petrelli. 1998. Visual display, pointing, and natural language: the power of multimodal interaction. In *Proceedings of the working conference on Advanced visual interfaces*, pages 164–173.

Jan de Wit, Paul Vogt, and Emiel Krahmer. 2022. The design and observed effects of robot-performed manual gestures: A systematic review. *ACM Transactions on Human-Robot Interaction*.

Nick J Enfield, Sotaro Kita, and Jan Peter De Ruiter. 2007. Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics*, 39(10):1722–1741.

Brian Gleeson, Karon MacLean, Amir Haddadi, Elizabeth Croft, and Javier Alcazar. 2013. Gestures for industry intuitive human-robot communication from human observation. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 349–356. IEEE.

Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11):419–429.

Charles Goodwin. 2003. Pointing as situated practice. In *Pointing*, pages 225–250. Psychology Press.

Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The reliability of non-verbal cues for situated reference resolution and their interplay with language: implications for human robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 189–196.

Zhao Han, Polina Rygina, and Thomas Williams. 2022. Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 1–14.

Zhao Han and Tom Williams. 2022. A task design for studying referring behaviors for linguistic hri. In *2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-Breaking Report (LBR)*.

Adam Kendon. 1997. Gesture. *Annual review of anthropology*, 26(1):109–128.

Aelee Kim, Jooyun Han, Younbo Jung, and Kwanmin Lee. 2013. The effects of familiarity and robot gesture on user acceptance of information. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 159–160. IEEE.

Sotaro Kita. 2003. *Pointing: Where language, culture, and cognition meet*, chapter Pointing: A Foundational Building Block of Human Communication. Psychology Press.

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2005. Deixis: How to determine demonstrated objects using a pointing cone. In *International Gesture Workshop*, pages 300–311. Springer.

Andy Lücking, Thies Pfeiffer, and Hannes Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

David McNeill. 1985. So you think gestures are nonverbal? *Psychological review*, 92(3):350.

David McNeill. 2003. Pointing and morality in chicago. In *Pointing*, pages 301–314. Psychology Press.

David McNeill, Justine Cassell, and Elena T Levy. 1993. Abstract deixis.

David McNeill and Elena Levy. 1982. Conceptual representations in language activity and gesture. *Speech, place, and action*, pages 271–295.

Raedy Ping and Susan Goldin-Meadow. 2010. Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, 34(4):602–619.

Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 101–108. IEEE.

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217.

Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 342–349. IEEE.

Boris Schauerte and Gernot A Fink. 2010. Focusing computational visual attention in multi-modal human-robot interaction. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, pages 1–8.

Boris Schauerte, Jan Richarz, and Gernot A Fink. 2010. Saliency-based identification and recognition of pointed-at objects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4638–4643. IEEE.

Adam Stogsdill, Grace Clark, Aly Ranucci, Thao Phung, and Tom Williams. 2021. Is it pointless? modeling and evaluation of category transitions of spatial gestures. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 392–396.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1).

Stefan Waldherr, Roseli Romero, and Sebastian Thrun. 2000. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9:151–173.

Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. 2016. Situated open world reference resolution for human-robot dialogue. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 311–318. IEEE.

Tom Williams and Matthias Scheutz. 2015a. A domain-independent model of open-world reference resolution. In *CogSci*.

Tom Williams and Matthias Scheutz. 2015b. Power: A domain-independent algorithm for probabilistic, open-world entity resolution. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1230–1235. IEEE.

Poster Abstracts

Topic and genre in dialogue

Amandine Decker[1,2], **Ellen Breitholtz**[2], **Christine Howes**[2] and **Staffan Larsson**[2]
[1]Université de Lorraine
amandine.decker@loria.fr
[2]University of Gothenburg
{ellen.breitholtz,staffan.larsson}@ling.gu.se, christine.howes@gu.se

1 Introduction

In this paper we argue that *topic* plays a fundamental role in conversations, and that the concept is needed in addition to that of genre to define interactions. In particular, the concepts of genre and topic need to be separated and orthogonally defined. This would enable modular, reliable and controllable flexible-domain dialogue systems.

In communicative activities, genre and topic tend to be interleaved in the sense that the manner in which a particular topic is addressed can differ significantly across various genres. For instance, a conversation about politics may unfold differently in a formal debate compared to a casual conversation among friends, and a recipe for a dish can be the topic of an instructional dialogue, or a discussion among participants as how to best prepare the dish.

Analysing the influence of genre on the treatment of topics would allow us to understand how general features of interaction are adapted to specific conversations. In this paper we discuss the treatment of topics and genres in different linguistics theories and how studying the way they influence each other may help designing reliable and controllable open-domain dialogue systems that could be adapted to task-oriented conversations in many different domains.

2 Topic and genre in linguistic theories

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* (Wong and Ginzburg, 2018), etc.

When defining genres, a frequently used concept is that of *activity* in the context of which language occurs. 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". These sub-goals could be topics in a discussion carrying out the activity type "Buying/selling coffee in a café" but they could be organised in many ways with for example all the topics following each other linearly or on the contrary being all embedded in each other.

Similarly, genres can be seen as a set of actions that must be realised, or a set of questions under discussion that must be resolved (Ginzburg, 2016), to make a certain interaction successful. In that sense, the genre sets the minimal requirements in terms of outcome for a conversation but it does not say anything on the content of it beyond these requirements. Besides, while genre constrains the surface structure, content plays an important role in the detailed one. Formalising topics and understanding the way they can be articulated could help modelling a hierarchical structure of content.

Topics have been discussed in different ways in the literature, the definitions mostly vary by their granularity. A sentence topic (Bolinger, 1952; Firbas, 1964; Halliday, 1967; Givón, 1983) is an element of the sentence, usually a noun phrase, that the sentence comments on (Hockett, 1958). *Discourse topics*, on the other hand, are not necessarily explicit. They refer to what a piece of discourse is about, though the formalisation of this "aboutness" is debated. Discourse topics have been defined as based on the "question of immediate concern" (Keenan et al., 1983), explicitly stated or not, or as

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

"the proposition or set of propositions that the question of immediate concern presupposes" (Schieffelin and Keenan, 1976). Discourse topics are also considered in the frame of certain discourse modelling theories such as Segmented Discourse Representation Theory (SDRT) (Asher, 2004). Coming up with a theory organising these different levels of granularity would enable us to come up with a hierarchical modelling of topics (Teh et al., 2006).

3 Variability in dialogue

The fulfilment of a conversation goal can be achieved in many different ways, encompassing linguistic and extra-linguistic elements, and their various combinations. Consider a scenario at a café where a customer wishes to order a drink. This goal can be accomplished by pointing at the desired drink, providing a verbal description, employing both actions simultaneously, or in some cases, no action may be necessary if the customer is a regular one with a well-known preference. The diverse range of methods exemplifies the flexibility inherent in achieving conversation goals.

The straightforwardness of attaining conversation goals also varies. Sometimes, intermediate questions need to be resolved before reaching a final decision. For instance, a customer may inquire about the type of milk used in the café and only place their order once they know which drinks are lactose-free. In such cases, the fulfilment of the conversation goal is contingent upon gathering additional information and resolving relevant queries. Such examples show how a straightforward request for action can sometimes turn into something more complex, where information is requested and different alternatives can be discussed and compared.

Conversations may also deviate from a strictly goal-oriented path, allowing for detours and tangential discussions. For instance, while inquiring about a specific product, an individual might share an anecdote related to the product itself. Questions about lavender cookies could trigger memories of holidays in Provence and lead to a spirited debate about the finest variety of lavender or even spark a discussion about the seller's vacation plans. Such diversions from the primary topic rely on the participants' freedom and inclination to explore different avenues within the conversation.

While many conversational goals are associated with a default genre (and related dialogue structures), it sometimes happens that dialogue participants deviate from these defaults. The extent to which default structure diverge from the original goals of a conversation is likely influenced by the genre of the conversation and its level of formality or standardisation. Additionally, the social aspect of the interaction also plays a central role. It appears that conversations with a greater social orientation tend to afford participants more freedom to deviate from the central goal.

4 Application to dialogue systems

Variability in dialogue is a challenge for general-purpose dialogue models. There may well be an open-ended universe of dialogue genres (language games, dialogue types), which we cannot hope to map out (Wittgenstein, 1953). In any case only a limited number of dialogue genres has so far received attention from the dialogue systems / conversational AI community (including industry). Having a better understanding of the way topics and genre interact could help creating a more modular and general framework that could be fine-tuned for more specific tasks.

Different dialogue genres will be associated with different kinds of dialogue patterns. In a sense, a notion of dialogue genre is not strictly necessary for dialogue systems. What is needed in each domain is dealing with the dialogue patterns that appear there. However, we believe that the notion of genre can serve as a powerful abstraction, allowing dialogue designers to understand which dialogue patterns are relevant in a domain.

5 Discussion

Distinguishing genre and topic and treating the two as orthogonal contributing factors could provide insights regarding the structural analysis of conversation. In terms of dialogue systems it would improve the adaptation of the model's interventions based on the current topic and its links to the previous ones as well as the type of conversation. However, topic modelling is a complex task even for human annotators (Purver, 2011) and creating guidelines to annotate dialogues based on the topics they discuss and the hierarchy between them is not a trivial problem. Such an annotation guide would make it possible to analyse the differences in terms of topical structure between different types of conversations and make dialogue systems adaptable to specific genres of conversations following this analysis.

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.

Nicholas Asher. 2004. Discourse topic. *Theoretical Linguistics*, 30:163–201.

Dwight L. Bolinger. 1952. Linear modification. *PMLA*, 67(7):1117–1144.

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

Jan Firbas. 1964. On defining the theme in functional sentence analysis. *Linguistiques de Prague*, 1:267–280.

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

Jonathan Ginzburg. 2016. *Semantics of dialogue*, Cambridge Handbooks in Language and Linguistics, page 130–170. Cambridge University Press. Edited by Maria Aloni and Paul Dekker.

Talmy Givón. 1983. Topic continuity in discourse: An introduction. *Topic continuity in discourse: A quantitative cross-language study*, 3:3–41.

Michael A.K. Halliday. 1967. Notes on transitivity and theme in english part i. *Journal of linguistics*, 3(1):37–81.

C.F. Hockett. 1958. *A Course in Modern Linguistics*. Macmillan.

Elinor Ochs Keenan, Bambi B. Schieffelin, and Martha Platt. 1983. *Acquiring conversational competence*, chapter Questions of immediate concern. Routledge.

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.

Matthew Purver. 2011. *Topic Segmentation*, pages 291–317. Wiley.

Bambi B. Schieffelin and Elinor Ochs Keenan. 1976. Topic as a discourse notion: A study of topic in the conversations of children and adults. In *Subject and topic*, pages 335–384. Academic Press.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. New York, NY, USA: Wiley-Blackwell.

Kwong-Cheong Wong and Jonathan Ginzburg. 2018. Conversational types: a topological perspective. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Aix-en-Provence, France. SEMDIAL.

One Signer at a Time? A Corpus Study of Turn-Taking Patterns in Signed Dialogue

Dan Green and **Arash Eshghi**
Heriot-Watt University
Edinburgh, Scotland
dan.green@hw.ac.uk and a.eshghi@hw.ac.uk

Abstract

It is not contentious that spoken dialogue is organised as a rapid exchange of turns with very minimal gap or overlap; underpinned by the real-time and highly predictive nature of human language processing. By contrast, research on patterns and mechanisms of turn taking in *signed* interaction is very scarce, to the extent that there isn't even broad consensus on whether signed dialogue is best characterised under a one-signer-at-a-time model. In this paper, we present a preliminary corpus study of turn-taking patterns in signed dialogue in British Sign Language (BSL) using the BSL Corpus Project. Our results are broadly compatible with one-at-a-time signing, albeit obscured by non-semantic signer movements. However, we also identify examples that do not fit this model which require further study.

1 Introduction

Sacks et al.'s (1974) seminal paper presented an abstract model of turn-taking in spoken conversation, capable of organising an orderly exchange of turns at talk in a flexible way, bottom up, with two or more participants. There are several corpus studies that confirm the prevalence of this *one-speaker-at-a-time* model mostly by demonstrating just how short gaps and overlaps between turns are (see e.g. Brady, 1968; Weilhammer and Rabold, 2003; Heldner and Edlund, 2010). There have been objections too (see e.g. Heldner and Edlund, 2010), but see Levinson and Torreira (2015) for strong counter arguments. This model of everyday conversational organisation has also been shown to be strongly universal (Stivers et al., 2009; Enfield et al., 2010).

By contrast, there is a paucity of research on both turn-taking patterns and mechanisms for projecting the end of turn in *signed* interaction. The lack of direct signal interference from simultaneous signing (c.f. overlapping audio signals in speech) raises the question of whether signed interaction is more tolerant to overlap. Coates and Sutton-Spence (2001) introduce the possibility that signed dialogue may be organised into both one-at-a-time signing and the use of a "collaborative floor". Subsequent work has explored the former (e.g. de Vos et al., 2015; De Vos et al., 2016; Lepeut, 2022; de Vos et al., 2022) but the latter remains largely ignored, in favour of drawing direct parallels with spoken dialogue.

de Vos et al. (2015) propose ignoring preparatory movements at the start of utterances, signers holding signs in place at the end of utterances and the signer retracting their hands; they term these as "stroke-to-stroke" (STS) timings as opposed to "sign-naive" (SN) timings (which include all movement). By discounting these segments of dialogue, they demonstrate that Dutch signers' turn-taking (in Nederlandse Gebarentaal) follows broadly the same patterns as spoken dialogue which therefore means that they can be characterised under a one-signer-at-a-time model. However, they restrict their study to question-answer sequences only, which limits the scope of their study – crucially, for example, there is no analysis of the function or form of overlaps in other types of sequence.

In this pilot study, we aim to investigate turn-taking patterns in BSL by examining data from the BSL Corpus Project (BSLCP) (Schembri et al., 2013). Our findings are consistent with those of de Vos et al. (2015), but we also find – as yet anecdotal – evidence that overlaps are less disruptive in signed interaction; and point forward to some future research directions.

2 Materials: The BSL Corpus Project

The data for this study comes from the BSLCP, collected between 2008-2011. The conversation section[1] of the BSLCP consists of 122 30-minute, unscripted dialogues between pairs of deaf signers of various backgrounds from different parts of the UK totalling approximately 60 hours. For each dialogue, there are three video recordings: a

[1]Other sections, with the same participants, include interviews, monological narratives and lexical elicitation.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

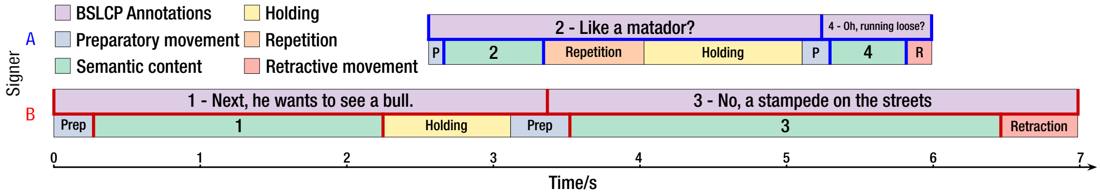


Figure 1: Example repair sequence from pilot BSLCP Study

close-up of each participant and a wide shot of both. Of these, 29[2] dialogues are annotated with roughly 500, precisely-timed glosses for each participant, along with a free translation of each utterance, yielding 4-5 minutes of annotated dialogue per pair. We use a subset of these for this study.

3 Procedure

Videos and annotations were downloaded from the BSLCP website with a research licence. A Python library - Pympi (Lubbers, 2015) - for working with ELAN data was used to combine annotations into a single file for each conversation with uniform tier names. This restructuring allows the files to be processed automatically. It proved necessary to manually adjust the timings for each file, as well as clipping the annotations at the end of the last annotated turn[3]. This produced a subcorpus of 37 minutes of dialogic data across 8 conversations.

Pympi then allows the automatic detection of turn transition times from the comparison of two tiers of turn data. This identifies gaps, overlaps, pauses and '`within-overlaps`' in the data. For this investigation, pauses were ignored, as the two turns either side of a pause can be considered as a single turn. Overlaps were ascribed a negative time value and gaps retained a positive time value. `Within-overlaps`, where there was no swap between the signers, were also given a negative time value but were kept separate from other overlaps.

4 Results

The conversations that were analysed as part of this study and the observed timings are summarised in Table 1. When considering only transitions where the primary signer changed, the mean transition time was -551ms. When considering all turns, the mean transition time was -968ms. The distribution of timings can be seen in Figure 2.

5 Discussion

As Table 1 and Figure 2 show, the timings obtained are, on the whole, consistent with the findings from de Vos et al. (2015). They observed a median of -607ms and a mean of -812ms when using SN

[2]Where annotations are available for *both* participants.
[3]To discount sections of annotations with just one signer.

	Examples	Mean Duration
Conversations	8	4m 38s
Turns	480	5.40s
Gaps	53	566ms
Overlaps	117	1056ms
Within-Overlaps	211	1304ms

Table 1: Turns and turn transitions in pilot BSLCP study

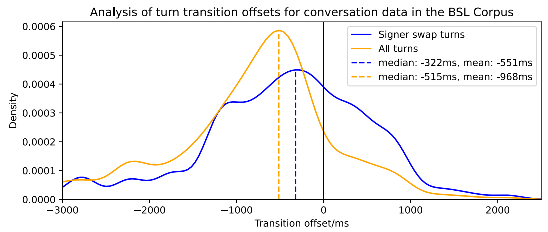


Figure 2: Turn transition times from pilot BSLCP Study

timings, comparable with the annotations in the BSLCP. Using STS timings, their results shift positively (i.e. gaps rather than overlaps) to be much closer to the universal averages found by Stivers et al. (2009). We expect that a similar STS analysis on our data would yield comparable results.

These findings support the hypothesis that BSL signers adhere to one-at-a-time turn taking norms. However, the significant number of (what appear to be non-interruptive) `within-overlaps` suggests that signed interaction may be more resilient to overlaps. Further research is needed into both the form and function of the `within-overlap` turns to establish how much of these might be characterised as *backchannels* or *interjections*. We illustrate this issue with a repair sequence from our BSL data with added STS annotations (Figure 1).

Utterances 1-3 now appear to occur sequentially, with gaps between each turn. However, even with an STS analysis, utterance 4 is still in complete overlap with utterance 3. This demonstrates the problems with SN timings but also that even using STS timings, there remains `within-overlaps` without an explanation. It is not clear, in this example, what effect the overlap has on the interaction.

Can these overlaps be characterised as backchannels? Or more generally, how disruptive (or not) are they? How are they sequentially integrated? What effect do the 'non-semantic' movements (ignored by STS timings) have on turn taking?

References

Paul T. Brady. 1968. A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal*, 47(1):73–91.

Jennifer Coates and Rachel Sutton-Spence. 2001. Turn-taking patterns in deaf conversation. *Journal of Sociolinguistics*, 5(4):507–529.

Connie De Vos, Marisa Casillas, Onno Crasborn, and Stephen C. Levinson. 2016. Stroke-to-stroke turn-boundary prediction in Sign Language of the Netherlands. In *the 12th International Conference on Theoretical Issues in Sign Language Research (TISLR12)*.

Connie de Vos, Marisa Casillas, Tom Uittenbogert, Onno Crasborn, and Stephen C. Levinson. 2022. Predicting conversational turns: Signers' and nonsigners' sensitivity to language-specific and globally accessible cues. *Language*, 98(1):35–62. Publisher: Linguistic Society of America.

Connie de Vos, Francisco Torreira, and Stephen C. Levinson. 2015. Turn-timing in signed conversations: coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*, 6.

N.J. Enfield, Tanya Stivers, and Stephen C. Levinson. 2010. Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42(10):2615–2619. Question-Response Sequences in Conversation across Ten Languages.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Alysson Lepeut. 2022. When hands stop moving, interaction keeps going: A study of manual holds in the management of conversation in French-speaking and signing Belgium. *Languages in Contrast*, 22(2):290–321.

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

Mart Lubbers. 2015. Pympi. Available at: https://dopefishh.github.io/pympi/index.html.

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.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the british sign language corpus. *Language Documentation & Conservation*, 7:136–154. Publisher: University of Hawaii Press.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592. Publisher: Proceedings of the National Academy of Sciences.

Karl Weilhammer and Susen Rabold. 2003. Durational Aspects in Turn Taking. In *15th International Congress of Phonetic Sciences (ICPhS-15)*, pages 2145–2148, Barcelona, Spain.

From position to function: Exploring word distributions within intonation units in American English conversation

Ryan Ka Yau Lai, Lu Liu, Haoran Yan and John W DuBois
University of California, Santa Barbara
{kayaulai, lu20, haoranyan, dubois}@ucsb.edu

1 Introduction

Traditionally, Firthian semantics (Firth 1957) examines meanings of linguistic forms through co-occurrences with other forms. This distributional method has enjoyed tremendous success in computational approaches, yet there has been less attention to how forms are distributed within larger units. Discourse markers' functions are often linked to positions in interactional units like turns and sequences (e.g. Sato 2008, Kim 2022, Fuentes-Rodríguez et al. 2016), but other form classes or prosodic units like the intonation unit (IU; DuBois 1992, Chafe 1994, Wahl 2015) are less frequently investigated. In this study, we examine the length of IUs in which words appear and position of words within IUs in the Santa Barbara Corpus of Spoken American English (DuBois et al. 2000), which is manually annotated for IUs based on acoustic cues (DuBois 1992). We find strong systematicity in word distributions across the lexicon, modellable with simple probabilistic models.

2 Exploring prosodic profiles

We first plot the distribution of words within the IU in heatmaps (Figure 1). Most words display clear tendencies as to where they appear in IUs, with three types of patterns. Firstly, words have different length preferences: Interjections prefer very short IUs and prepositions typically prefer longer ones. Secondly, some distributions are centred around a fixed place value, e.g. subject pronouns tend to be first and auxiliaries second. Finally, some distributions are centred around a fixed value from the *end* of an IU: accusative pronouns tend to come last, while determiners and prepositions are typically 1-2 places from the next IU boundary. Some words display bimodal distributions: conjunctions often have one mode near the front of an IU and another, smaller one near the end.

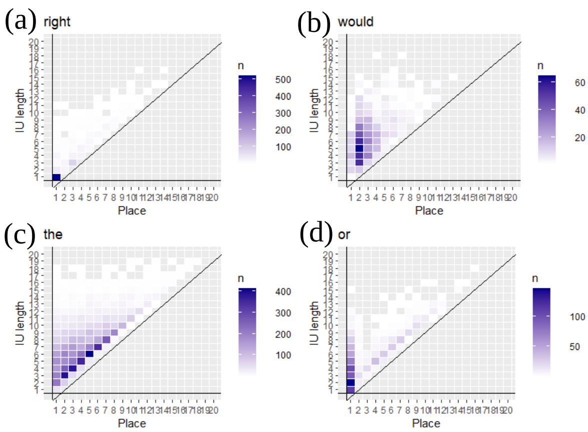


Figure 1: Heatmaps of place and length for the short-biased *right* (a), front-biased *would* (b), end-biased *the* (c) and bimodally distributed *or* (d). The *y*-axis gives the length of the IU where a word appears; the *x*-axis gives the *place*, i.e. sequential position of a word within an IU. The darker a position in the heatmap, the more tokens found in it.

Hierarchical clustering on the joint distributions of the 200 word-types with highest Juilland's *U* (Gries 2008) values, based on Tai & Pham-Gia's (2010) measure of cluster width, reveals syntacto-semantically interpretable clusters. Results at 22 clusters are in the Appendix. Interjections take up two clusters, typically occupying one-word IUs (and occasionally the ends of longer IUs), consistent with their often strong associations with intonation contours (Norrick 2009). At initial positions of longer IUs are conjunctions and other words relating different stretches of discourse, often serving as prefaces (Kim & Kuroshima 2013) to turns. *Wh*-words also tend to come first in an IU and modal-evidential verbs (main and auxiliary) second – words typically described as constituting recognisable turn beginnings in turn-initial position (Schegloff 1996), but the IU-initial tendency remains even in turn-medial positions, e.g. after filtering out uppercase-initial instances. One cluster contains words like *know* and *think*

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

preferring final positions of two-word IUs, reflecting their role in stance-marking chunks like *I think* (Thompson 2002). Words attracted to IU ends include nouns and non-nominative pronouns, projected by words attracted to (ante)penultimate positions like prepositions and determiners.

3 Modelling prosodic profiles

To go beyond exploratory analysis to predictive modelling, we model the words' prosodic profiles with a Bayesian approach, focusing on words with unimodal distribution. We adopt a parametric approach so the distributions can be summarised using a small number of interpretable parameters.

For each word, we first modelled the length of IUs that it appears in using a negative binomial distribution. We use the parametrisation standard in negative binomial regression (Ver Hoef et al. 2007) with the following probability mass function:

$$f(y;\mu,\phi) = \binom{y+\phi-1}{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}$$

where μ is the mean and ϕ a dispersion parameter; the variance is $\mu(1+\mu/\phi)$. Since 0 places are impossible, we truncated the distribution at 0.

To obtain the joint distribution of place and length, we then modelled the distribution of the place conditional on the length. For the front-biased words, we modelled the place values directly. Since back-biased words tend to be consistently the same number of places from the end of the IU, we model the *back* values of those words by subtracting place from IU length and adding one. The conditional distributions of the place and back values were modelled as Poisson distributions with rate parameter λ, and values below 1 and above the length truncated.

The models were fit in a Bayesian framework in Stan through RStan (Stan Development Team 2023a, 2023b). Priors were set on the parameters as follows: $\lambda \sim Gamma(3,3)$, $\phi, \mu \sim Gamma(1,1)$. The means of the posterior distributions of λ and μ, along with the 'variance' of IU length $\mu(1+\mu/\phi)$, are shown in Table 1 and Table 2 for eight words.

From μ values, which reflect length preferences, clearly *yes* and *right* are much more biased towards short IUs than the rest. This is expected from their functions as interjections: They can function alone to express stance alignment (DuBois 2007) and, for *right*, as backchannels. *Right* has great variance in IU length considering how short the length usually is, reflecting *right*'s secondary use as an adjective.

λ values reveal *yes* and *he* to be most attracted to the edges of IUs, followed by *right*, whereas *the* and *an* are the farthest from IU edges. The interjections' attraction to front edges may allow for early action ascription in the IU, considering their stance alignment functions (cf. Levinson 2012 for similar discussions in the context of turns), and the attraction of *he*, a highly accessible (Ariel 2001) referential expression, to IU beginnings reflects general preferences for producing highly accessible elements first (Levshina 2022). The articles' relatively long distance from the IU edge allows them to project lengthy, inaccessible referential expressions in English.

word	λ	μ	$\mu(1+\mu/\phi)$
yes	1.76	0.22	0.49
he	1.99	6.07	8.80
just	3.20	6.03	11.7
would	3.24	6.65	9.17

Table 1: Parameter estimates for front-biased words. Note that these are not true estimates of means and variances because the distributions are truncated.

word	λ	μ	$\mu(1+\mu/\phi)$
right	2.67	0.58	3.85
an	3.86	6.68	10.23
little	3.73	7.09	10.60
the	4.59	7.01	9.83

Table 2: Parameter estimates for back-biased words.

4 Conclusion and future directions

Words in English conversation reliably pattern as to where they occur in IUs of what length. Some of these distributions can be modelled with simple probability distributions with parameters revealing of the words' functions. This shows location within IUs as a promising avenue for examining linguistic function distributionally, adding to analyses based on collocations and interactional units, perhaps even suggesting refinements of traditional syntax-based word classes like nouns and verbs, while incorporating interjections/discourse markers that do not fit neatly into sentence-based analyses.

We plan to extend these models to account for special words, e.g. those like *'re* or *'m* where initial positions are much less likely than Poisson-like models predict. We also plan to model words with clearly bimodal distributions like *or*. Finally, we hope to compare word distributions within IUs with other units like the turn, turn-constructional unit and sequence, to determine how much additional information IUs capture.

References

Ariel, Mira. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord & Wilbert Spooren (eds.), *Text representation: Linguistic and psycholinguistic aspects*, 29–87. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.8.04ari.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

DuBois, John W. 1992. Discourse transcription. *Santa Barbara Papers in Linguistics* 4. 1–225.

DuBois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson & Nii Martey. 2000. *Santa Barbara Corpus of Spoken American English*. CD-ROM. Philadelphia: Linguistic Data Consortium. Linguistic Data Consortium.

DuBois, John W. 2007. The stance triangle. In Robert Englebretson (ed.), *Pragmatics & Beyond New Series*, vol. 164, 139–182. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.07du.

Firth, J.R. 1957. A synopsis of linguistic theory, 1930-1955. In J.R. Firth (ed.), *Studies in Linguistic Analysis*. Oxford: Basil Blackwell.

Fuentes-Rodríguez, Catalina, María Elena Placencia & María Palma-Fahey. 2016. Regional pragmatic variation in the use of the discourse marker pues in informal talk among university students in Quito (Ecuador), Santiago (Chile) and Seville (Spain). *Journal of Pragmatics* 97. 74–92. https://doi.org/10.1016/j.pragma.2016.03.006.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. https://doi.org/10.1075/ijcl.13.4.02gri.

Kim, Hye Ri Stephanie & Satomi Kuroshima. 2013. Turn beginnings in interaction: An introduction. *Journal of Pragmatics* 57. 267–273. https://doi.org/10.1016/j.pragma.2013.08.026.

Kim, Mary Shin. 2022. Identical linguistic forms in multiple turn and sequence positions in Asian languages. *Journal of Pragmatics* 200. 1–7. https://doi.org/10.1016/j.pragma.2022.06.007.

Levinson, Stephen C. 2012. Action Formation and Ascription. In Jack Sidnell & Tanya Stivers (eds.), *The Handbook of Conversation Analysis*, 101–130. https://doi.org/10.1002/9781118325001.ch6.

Levshina, Natalia. 2022. *Communicative Efficiency: Language Structure and Use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108887809.

Norrick, Neal R. 2009. Interjections as pragmatic markers. *Journal of Pragmatics* 41(5). 866–891. https://doi.org/10.1016/j.pragma.2008.08.005.

Sato, Shie. 2008. Use of "please" in American and New Zealand English. *Journal of Pragmatics* 40(7). 1249–1278. https://doi.org/10.1016/j.pragma.2007.09.001.

Schegloff, Emanuel A. 1996. Turn organization: one intersection of grammar and interaction. In Elinor Ochs, Emanuel A. Schegloff & Sandra A. Thompson (eds.), *Interaction and Grammar*, 52–133. Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.002.

.Stan Development Team. 2023a. Stan Modeling Language Users Guide and Reference Manual, 2.32. https://mc-stan.org.

Stan Development Team. 2023b. RStan: the R interface to Stan. R package version 2.21.8. https://mc-stan.org/.

Tai, Vo Van & T. Pham-Gia. 2010. Clustering probability distributions. *Journal of Applied Statistics* 37(11). 1891–1910. https://doi.org/10.1080/02664760903186049.

Thompson, Sandra A. 2002. "Object complements" and conversation towards a realistic account. *Studies in Language* 26(1). 125–163. https://doi.org/10.1075/sl.26.1.05tho.

Ver Hoef, Jay M. & Peter L. Boveng. 2007. Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* 88(11). 2766–2772. https://doi.org/10.1890/07-0043.1.

Wahl, Alexander. 2015. Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics* 13(1). 191–219. https://doi.org/10.1075/rcl.13.1.08wah.

Appendix

Interpretation	Examples	Concentrated in
interjections	hm, oh, right, unhunh	one-word IUs
interjections and vocatives	god, mom, sure, uh, why	one-word IUs and, secondarily, other final positions of shorter IUs
time-/choice-related	after, before, every, or	mostly beginnings of short IUs + sometimes next-to-last positions of longer IUs

conjunction and conjunction-like words	and, so, which, but	strongly initial, well spread across IU sizes
subordinators and modals	how, maybe, what, where	strongly initial, well spread across IU sizes (more short-biased than 10)
modal-evidential verbs	know, mean, think, wanted	second position of two-word IUs
semantically light verbs	came, gon, wan, told	second to third positions of moderate-sized IUs
contractions and modal-evidential verbs	's, goes, guess, should	second positions of short IUs
temporal and modal adverbs	always, just, never, not	third word from the beginning of moderate-sized IUs
semantically light verbs	go, want, went, have	2-4 positions of moderate-sized IUs
light (pro)nouns	day, lot, me, anything	final positions of IUs, well spread out across IU lengths
(diverse)	around, back, time, say	final position across a range of IU lengths
semantically light nouns	everything, something, here	final positions, spread across IU lengths
(diverse)	four, kinda, really, remember	final to penultimate positions of shorter IUs
(diverse)	about, big, long, her	last or penultimate word of moderate-sized IUs
determiners, light content words, some prepositions	an, tell, very, real, call	penultimate position of moderate-sized IUs, highly concentrated
semantically light content words	good, years, great, like	penultimate to antepenultimate positions of short IUs

mostly determiners and prepositions	all, as, by, first, these	penultimate to antepenultimate words of IUs
modal and semantically light verbs	be, even, getting, take	penultimate to fourth-from-last positions of moderate IUs
prepositions and quantitative determiners	any, in, three, through	antepenultimate and penultimate positions of moderate-sized IUs
genitive pronouns and other determiners and semantically light adjectives	another, my, our, than	antepenultimate position across a range of IU lengths
nominative pronouns and modal verbs	are, does, is, it	well spread out or bimodal distribution of positions, short to moderate IUs

Common Strategy Patterns of Persuasion in a Mission Critical and Time Sensitive Task

Claire To, Setareh Nasihati Gilani, David Traum
Institute for Creative Technologies
clairt2@uci.edu, sngilani@ict.usc.edu, traum@ict.usc.edu

1 Introduction

We investigate different styles of approach to persuasion in high-stakes, time-critical interactions. In human dialogue, there are generally multiple motivations underlying choices of specific utterances and higher-level strategies in approaching an interaction. These can include achieving the speakers' own goals, helping the interlocutor achieve theirs, opening, and maintaining conversation, and maintaining interpersonal relations. People differ in terms of their weight on each of these goals, but decisions about what and how to say things also depends on the situation itself, e.g., what is at stake in the conversation and how urgent is a resolution needed. Differences in these factors may result in very different kinds of dialogues even when undertaken for the same purposes.

We examine a set of short dialogues (2-16 turns, average 7.62) all concerned with the same high-stakes, urgent goals. A disaster relief manager needs to communicate with people in the town who are in danger from an out of control forest fire. The manager wants to convince them to leave, and if necessary, offer resources to help them accomplish that. We look at how different experimental participants playing the manager role approach this situation, specifically what kinds and ordering of speech acts they perform in the initial stages of the dialogues. We look at whether and how proposals to act are presented, for example do they get right to the point, or first greet the other and ask after their interests before presenting their proposal.

We annotated the manager's turns with a high-level set of speech acts (Searle, 1969; Bunt et al., 2012) (a turn can realize multiple acts). We then categorized the dialogues with respect to position of greetings and proposals, looking at the trade-off between politeness and focusing on addressee's concerns vs getting to the point quickly. Finally, we looked at the number and types of proposals that were made. We are currently examining which situational factors are related to different types of approaches, taking into account factors of the participant, their interlocutor type and style, and urgency of the situation.

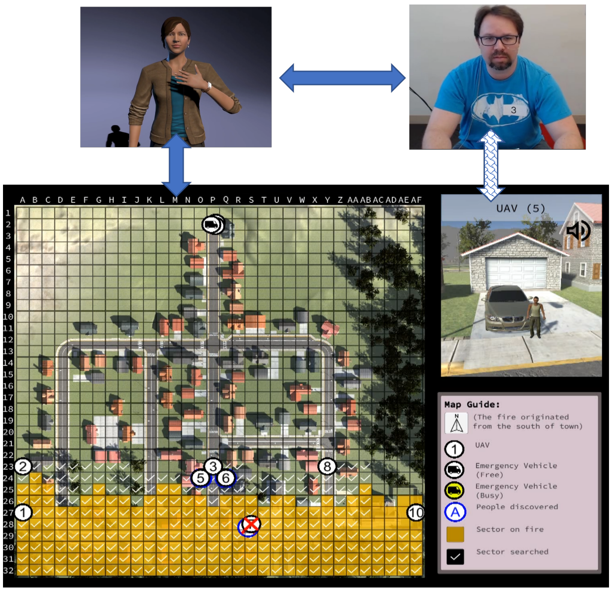


Figure 1: Overview of the simulation environment, the spokesperson, and the operator (aka the user)

2 Data

The data used is from an experiment first introduced by Chaffey et al. (2019), and illustrated in Figure 1. In this simulation, the human participant (shown top right) plays the role of a disaster relief manager, operating a swarm of robots and assisted by a virtual human spokesperson (Julie) for the swarm (shown top left). The manager (also called "operator") must deploy robots to track a forest fire that is spreading towards the town, search for residents within the town, establish communications with the residents, and rescue residents. Robots are of two types: flying drones, and ground transport vehicles. The spokesperson can be seen as an assistant, who can inform the operator about the situation with the swarm, but can also autonomously take on some tasks to relieve the operator's burden. In the lower part of the figure, the operator's view of the simulation is shown. They have a high-level map of the town, broken into grids that can be used for communicating locations to the drones. The current state of the fire is shown in orange. Robots are represented as circles (with numerical ids for individual drones). When residents are located (as shown in top right pane of the operator view), the operator tries to save them, sometimes engaging in dialogues with them or sometimes delegating this task to the virtual spokesperson.

There were five different residents in the simulation, representing different individuals or small groups, with different concerns about leaving, and different requirements to be able to leave (e.g. needing guidance or a transport vehicle). They were placed randomly within the environment, and the fire spread following a stochastic distribution. Resident utterances were pre-recorded by actors, and triggered using a Wizard of Oz interface by an experimenter, following a protocol for which concerns would be brought up and what would convince them to comply.

31 participants each participated in two runs of the simulation. Thus, the maximum possible number of distinct dialogues between the operator and a resident was 310. However, not all residents were discovered in each simulation run, and some residents were handled by the spokesperson rather than the operator. Eight participants delegated all interactions to the spokesperson. Only one participant had all 10 possible resident interactions. A total of 104 dialogues (average length of 34.68 seconds and 85% success rate) between a participant and a resident were identified and transcribed.

2.1 Speech Act Annotation

We annotated operator turns for the presence or absence of each of the following speech acts:

Greeting refers to the initiation of conversation. Opening remarks serve as a polite and friendly way to acknowledge the presence of the resident and establish the beginning of the conversation. (e.g., "Hello." "Are you there?" "Yes...")

Statement refers to providing insight, reason, justification, or information to the resident. (i.e., "It's an emergency." "There is an evacuation." "The vehicle is on the way.")

Question refers to inquiring the current status or information from the resident. (i.e., "Are you okay" "How are you?" "Do you need assistance?")

Proposal refers to presenting a course of action or plan to the resident. (i.e., "We need you to leave right now." "Can you guys please just get out of there as quickly as possible?" "You should probably try to get out there as quickly as possible.")

Concession refers to withdrawing a proposal. (i.e., "Okay, that's your choice." "Do understand that I did try to evacuate you.")

Closing refers to end of the conversation. (i.e., "Okay. Thank you." "Bye.")

We classified proposals based on who would do the proposed action and the strength of the commitment or obligation, yielding five types: command, request, suggestion, offer, and commitment.

3 Analysis

We identified 4 initiation patterns, based on the combination and positioning of greetings and proposal speech acts in the dialogues. These are from most to least urgent (or least to most polite):

1. **proposal in the first turn, no greeting** (14 dialogues, 8 with just a proposal, 6 also including a statement).
2. **proposal in the first turn with a greeting or question** (40 dialogues, some also including questions or statements in the 1st turn).
3. **proposal occurring after an initial greeting exchange** (44 dialogues)
4. **no proposal presented** (6 dialogues)

Of the 98 dialogues with proposals, 33 contained only a single proposal, 40 contained 2 proposals, 21 contained 3 proposals, 3 contained 4 proposals, and 1 contained 5 proposals. Concerning the first proposal type, 43 were commands, 25 commitments, 16 offers, 8 suggestions, and 6 requests.

3.1 Next Steps

We are currently examining patterns involving the relationships between proposals and other actions, and how they are distributed across the above initiation patterns. We will also look at correlations between types of patterns and several factors including individual operator participants and residents, correlations with facial expressions (Nasihati Gilani and Traum, 2023), success at saving the residents, and the simulation state to see whether the proximity of the fire to the resident makes a difference in the distribution of patterns.

4 Acknowledgements

The first author was supported by the NSF REU program, award 1852583, during her internship at the USC Institute for Creative Technologies. The second and third authors were sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Patricia Chaffey, Ron Artstein, Kallirroi Georgila, Kimberly A Pollard, Setareh Nasihati Gilani, David M Krum, David Nelson, Kevin Huynh, Alesia Gainer, Seyed Hossein Alavi, Rhys Yahata, and David Traum. 2019. Developing a virtual reality wildfire simulation to analyze human communication and interaction with a robotic swarm during emergencies. In *Workshop on Human Language Technologies in Crisis and Emergency Management*.

Setareh Nasihati Gilani and David Traum. 2023. Analyzing user's mental state and facial expressions in interaction with different personalities in a critical situation. In *1st International Multimodal Communication Symposium (MMSYM 2023)*.

John R. Searle. 1969. *Speech Acts*. Cambridge University Press, New York.

Referring as a collaborative process: learning to ground language through language games

Dominik Künkele[1] and **Simon Dobnik**[1,2]
Department of Philosophy, Linguistics and Theory of Science[1]
Centre for Linguistic Theory and Studies in Probability (CLASP)[2]
University of Gothenburg, Sweden
dominik.kuenkele@outlook.com and simon.dobnik@gu.se

Abstract

How do artificial agents based on neural networks coordinate on a new language through referential games over 3-d scenes? We extend a popular CLEVR dataset to control for different combinations of features of target and distractor objects and examine the success of referential grounding learned by the agents.

1 Introduction

Agents interact with the physical world through their actions and perception, and with other agents through language. Their sensors and actuators allow them to sample the world and their own state using measures that are continuous in nature such as intensity of light, distance, angles, velocity and others which can be measured with a high degree of accuracy. On the other hand, the language that is used to communicate with other agents is based on representations that are composed of a limited set of discrete and arbitrarily chosen symbols. How can both domains and representations arising from these interactions be combined? How are the ranges of measurements expressed in a continuous domain mapped to discrete linguistic labels? How is ambiguity and underspecification resolved? How can agents achieve it through interactive grounding (Regier, 1996; Roy, 2005; Cooper, 2023)?

In this paper we explore how agents based on artificial neural networks learn referential grounding of entities in images of 3-dimensional scenes through language games (Clark, 1996; Bartlett and Kazakov, 2005; Kirby et al., 2008; Steels and Loetzsch, 2009; Zaslavsky et al., 2018). One agent is describing the entities represented as features within bounding boxes of objects, inventing new vocabulary as necessary. The other agent learns to interpret the reference of symbols by identifying one of the bounding boxes based on object attributes such as shape, colour and size. Both agents learn through the success of interaction. The novelty of our work, compared with the previous work with this setup (Kharitonov et al., 2019; Lazaridou et al., 2017), consists the extension of the popular CLEVR dataset (Johnson et al., 2016) with new artificially generated 3-d scenes of objects. These can be referred to based on attributes such as *shape*, *colour* and *size* and discriminated based on different overlaps of these attributes between the target and the distractor objects.

2 CLEVR-Dale-2 and Dale-5

We extend the CLEVR dataset (Johnson et al., 2016) by dividing the objects into one *target object* and *distractors* and by controlling for the number of shared attributes between these groups as in the GRE algorithm in (Dale and Reiter, 1995). The target object is always unique, because at least one attribute is different from the distractors. Each distractor can share a maximum of two attributes with the target object. There is no restriction on the relation between distractors, hence it is possible to have multiple identical distractors in one image. Given the ranking of features in the original GRE algorithm, the target object is therefore identifiable either by *shape* (1), *shape* and *colour* (2) or *shape*, *colour* and *size* (3). For each image, fixed-size bounding boxes are extracted around the centre-point of each object. The *Dale-2* dataset contains one target object and one distractor, while the *Dale-5* dataset contains one target object and four distractors. Both datasets contain 10.000 images. Examples are shown in Appendix A.

3 Language games

The language games were developed and run in the EGG framework (Kharitonov et al., 2019).[1] Both our sender and receiver have a similar architecture to the *agnostic sender* and *receiver* of (Lazaridou et al., 2017), as shown in Appendix B. One central difference is the production of the message.

[1] https://github.com/DominikKuenkele/MLT_Master-Thesis

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

As we focus on sequences of referring expressions, made-up of different attributes, our models produce sequences of symbols for the message instead of a single symbol to refer to an image. This is done by using an encoder LSTM (sender) and a decoder LSTM (receiver) to encode language descriptions. Another difference is that both sender and receiver receive visual input as segmented objects rather than as two images. The order of the objects is random, except that the first object for the sender is always the target object to be referred to. For the sender, the images are passed through *ResNet101* (He et al., 2016) and a following linear layer that reduces the dimensions to an embedding size e_s. All embedded images are concatenated and passed through another linear layer to reduce the dimensions to the hidden size h_s. This is then used as the initial state of the encoder LSTM. After, the sequence of labels is generated through Gumbel-Softmax relaxation (Jang et al., 2017). The receiver also encodes all images using *ResNet101* with a following linear layer, reducing it to e_r. The sequence, received from the sender is the input for its decoder LSTM, where a hidden state with a dimension of h_r is randomly initialised. After each step of the LSTM, the receiver calculates a dot product between the hidden state and all of its image encodings. The receiver then 'points' to one of the images by applying a softmax function over the results of the dot products. The loss is calculated using the NLL-loss. Following, the losses for all steps are summed up, and all weights of the receiver as well as the sender are updated based on this summed loss.

4 Experiments and results

There are five variables in the experiments that are adjusted: (1) the image embedding size for the sender e_s, (2) the LSTM hidden size for the sender h_s, (3) the image/message embedding size for the receiver e_r, (4) the LSTM hidden size for the receiver h_r and (5) the size of the vocabulary $|V|$. Table 1 shows the accuracy of the models calculated on the basis of the success of communication if the receiver can identify the target object. A random guess corresponds to 50% in the *Dale-2* dataset and 20% in the *Dale-5* dataset.

For the *Dale-2* dataset it can be clearly seen that an embedding size and hidden size that are as high as the vocabulary size are beneficial for identifying the correct object. The receiver identifies almost every sample correctly with all sizes of 10. When the hidden and embedding sizes are increased, the guesses by the receiver are random. Interestingly, a vocabulary size of 10 is enough to communicate a meaningful message for the *Dale-2* dataset. Using *Dale-5* with four distractors and with low hidden, embedding and vocabulary sizes, the agents barely pass the random baseline at 23%. Only increasing the vocabulary size to 100 raises the accuracy by almost 20% to 43% which is still considerably lower than the 95% of the *Dale-2* dataset.

Dataset	h_s	e_s	h_r	e_r	$\|V\|$	**Acc.**
Dale-2	10	10	10	10	10	95%
Dale-2	50	50	128	128	10	50%
Dale-5	10	10	10	10	10	23%
Dale-5	10	10	10	10	20	23%
Dale-5	10	10	10	10	100	41%

Table 1: Results: h are different hidden sizes, e embedding sizes and $|V|$ vocabulary sizes.

5 Discussion and future work

Unsurprisingly, the agents have a much higher difficulty to discriminate a target object from four instead of one distractor. Since we discriminate objects based on properties that are also distinguished in human cognition (colour, size, shape), we expect that the vocabulary onto which the agents converge reflects these categories and is therefore close to human vocabulary. There are 48 possible combinations of attributes. Still, for Dale-2, a vocabulary size of only 10 is enough for an almost perfect accuracy with two objects. This hints to the fact that the agents don't describe the complete target object, but only rely on discriminative attributes between the objects. The need for a more detailed description of discriminative attributes is higher when more distractors are involved. Therefore, the models must learn more combinations of symbols in order to attest to this higher level of detail and how to relate them to features in the images.

In our ongoing work we are extending our analysis of features and agent configurations as well as we are investigating the emerged language and the new vocabulary, whether it consists of similar categories as human language and how its words are combined to form complete messages. In our future work we will extend the learning to relations between entities which introduce a high level referential underspecification.

References

Mark Bartlett and Dimitar Kazakov. 2005. The origins of syntax: from navigation to language. *Connection Science*, 17(3-4):271–288.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Robin Cooper. 2023. *From Perception to Communication: A Theory of Types for Action and Meaning*, volume 16 of *Oxford Studies in Semantics and Pragmatics*. Oxford University Press Press.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. *arXiv*, arXiv:1611.01144 [stat.ML]:1–13.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of language in games. *arXiv*, arXiv:1907.00852 [cs.CL]:1–6.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. *arXiv*, arXiv:1612.07182v2 [cs.CL]:1–11.

Terry Regier. 1996. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, Massachusetts, London, England.

Deb Roy. 2005. Semiotic schemas: a framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A. Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

A Extended CLEVR datasets

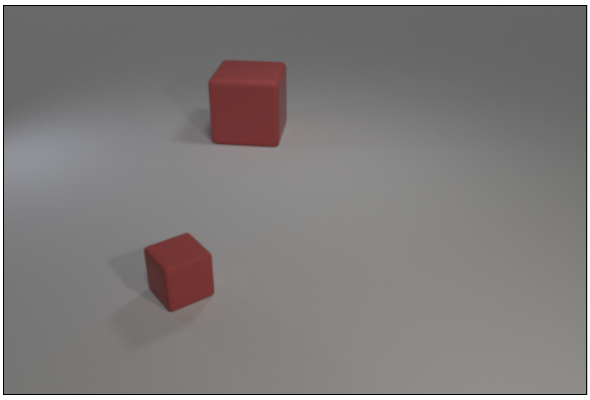


Figure 1: An example from the Dale-2 dataset

In Figure 1, the small red cube is the target object. Since all attributes except for the size are shared with the distractor, all three attributes are necessary, to identify it following Dale and Reiter (1995)'s rules, namely the *small red cube*.

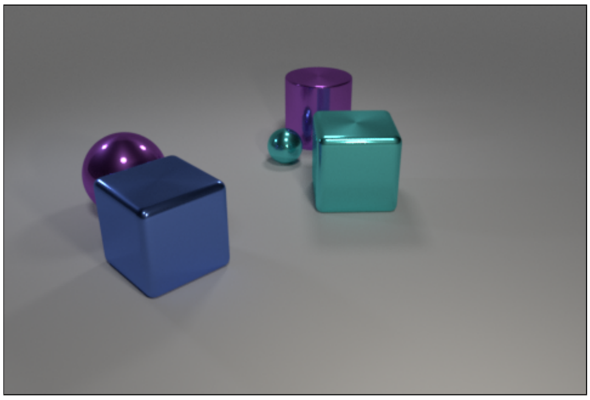
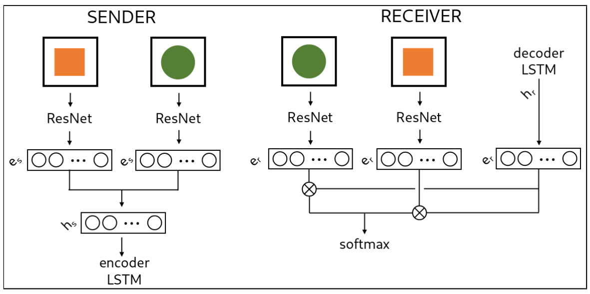


Figure 2: An example from the Dale-5 dataset

The target object in Figure 2 is the purple cylinder. It shares the same colour and size with the purple sphere, the same size with the two cubes and no attribute with the turquoise sphere. It can be uniquely identified as the *cylinder*.

B Setup of the language game



Humour in early interaction: what it can tell us about the linguistic, pragmatic and cognitive development of the child.

Chiara Mazzocconi and **Béatrice Priego-Valverde**
Institute of Language, Communication and the Brain, Aix-Marseille University,
Laboratoire Parole et Langage, Aix-en-Provence, France
chiara.mazzocconi@univ-amu.fr

Abstract

We analyse longitudinally humour episodes appreciation and production in 4 North-American children while interacting freely with their mums at home (Providence Corpus, Demuth et al. (2006)), at 12, 18, 24, 30, 36 months. We annotate humourous episodes combining resources from the General Theory of Verbal Humor (Attardo and Raskin, 1991), namely the construct of *Script Opposition* (im/possible, ab/normal, non/actual), with a further characterisation of those in terms of the knowledge domain such opposition is related to (e.g. Natural World, Social Conventions, Meta-linguistic). We observe significantly different distributions in the types of SO and domains between mothers and children and a developmental trajectory in the emergence of SOs and domains in children. We discuss how these patterns reflect the child linguistic and cognitive development and how they can inform us about the general principles of reasoning acquired and developing.

1 Introduction

Humour is inherently interactive and relies deeply on shared knowledge, conventions, and cultural norms (Priego-Valverde, 2003), being often context-dependent (Cunningham, 2005). Humour appreciation has indeed been shown to correlate and be informative about pragmatic and mentalising abilities (Aykan and Nalçacı, 2018; Bischetti et al., 2019). 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). The ability to appraise (and eventually enjoy) an incongruity entails the acquisition and knowledge of a typical pattern. Therefore when looking at child development, humour appreciation can be informative about their pragmatic development and considered as a marker of what children are learning about the world, their culture and language (Martin, 2007; Mireault and Reddy, 2016; Loizou and Recchia, 2019; Telli and Hoicka, 2022) (and about their current models). Piaget (1945) considered laughter in relation to humour as a sign of cognitive mastery: humour being mostly appreciated when the stimulus involves concepts that the child has just acquired or is in the process of learning, placed therefore at the *zone of proximal development* (Vygotsky, 1980), when it is neither too hard nor too easy to grasp the incongruity (Zigler et al., 1966; McCall, 1972; McGhee, 1979). While some cross-sectional studies have been conducted (e.g. Hoicka and Akhtar (2012); Telli and Hoicka (2022)), structured longitudinal investigations of humour development are still scarce.

2 Current Study

We investigate humour appreciation and production in spontaneous mother-child interaction longitudinally from 12 to 36 months of age. We analyse humour episodes occurring in 4 American English mother-child dyads (Providence Corpus, Demuth et al. (2006).) during 30 minutes of spontaneous interaction at home at 12, 18, 24, 30 and 36 months of child age. We integrated the speech annotations available from the CHILDES database with those related to laughter occurrences (N=287) and pragmatic functions publicly shared in Mazzocconi and Ginzburg (2022) and Mazzocconi and Ginzburg (2023) respectively[1].

Following the methodology used by Archakis and Tsakona (2005) in adult conversation, we established two criteria for humourous episodes identification: (1) the occurrence of laughter and (2) the identification of an incongruity appraised or intended as pleasant (Mazzocconi et al., 2020) in what the laughter is related to. The current study is based on the analysis of 271 humourous episodes:

[1]All transcriptions and audio/video files can be found on the CHILDES database. Laughter annotations are available at https://osf.io/48fmd/ and https://osf.io/8enf3/.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

113 identified through laughs produced by children and 158 through laughs produced by mothers.

Each humourous episode is annotated in terms of the **Script Opposition(s)** involved (Raskin, 1985; Attardo and Raskin, 1991), following the hierarchical step-wise methodology by Hempelmann and Ruch (2005): (1) *Possible-Impossible*, (2) *Normal-Abnormal*, (3) *Actual-Non Actual* (i.e., when the overlap/clash is between two possible and typical scripts, and the incongruity relies on having initially considered one instead of the other). Each SO is further characterised by describing which knowledge **Domain** it is related to: (1) *Natural World*: Human scheme, Physical Laws, Use and properties of objects; (2) *Social Domain*: Default action sequences, Moral rules, Conversational rules; (3) *Meta-linguistic Domain*: Phonetics, Phonology, Semantics, Pragmatics (i.e. less probable meaning, e.g. irony or scare-quoting). Annotations were conducted using the software ELAN[2] (Brugman and Russel, 2004) and the statistical analyses using R (R Core Team, 2022).

3 Results

The most frequent type of SO is *ab/normal*, being in proportion more frequent in mothers than in children (Fisher's Exact Test, $p=.011$). The SO *im/possible* is significantly more frequent in children than mothers ($p<.001$), while the SO *non/actual* is more frequent in mothers than in children ($p=.016$). While the SOs *ab/normal* and *im/possible* are present over all the time points analysed, in children we observe the SO *non/actual* only from 24 months of age. In terms of Domain, SOs related to Natural World are significantly more frequent in children than in mothers ($p<.001$); those related to the Meta-Linguistic domain are more frequent in mothers than in children ($p=.004$), while SOs pertaining to Social Conventions are more balanced ($p=0.26$). When looking at the longitudinal patterns, the tendency for children to appreciate SO related to Natural World more than mothers is constant over time. On the other hand, we observe that mothers produce more laughs in relation to Social Convention violation especially at the first time points, while towards 36 months the percentages observed are more balanced for this domain. Looking at the sub-types, in children we observe the emergence of laughter related to humourous episodes involving violations of Conversational Conventions only from 18 months, and to the violation of Moral Rules from 24 months (being more frequent in children than in mothers). Humour episodes in the Meta-Linguistic domain are more frequent in mothers at all time-points. Children appreciate SOs related to Phonetics aspects of speech and vocal production similarly to mothers from 12 months of age, while we see SOs related to Phonology and Semantics to be rarer in children. Humourous episodes related to the Pragmatic subdomain, are observed only in mothers when the child is 36 months, while absent in children.

4 Discussion

We observed developmental trajectories both for the type and the pertaining Domain of SOs involved in humourous episodes appreciated by children. The significantly higher frequency of laughter related to *im/possible* SOs in children than in mothers might be related to the fact that funniness is best found at the zone of proximal development (Piaget, 1945; Vygotsky, 1980). Children might especially appreciate this kind of SO, since relying on the ontology of the world that they are in the process of building, while for mothers such oppositions might be less amusing. The observation of the *non/actual* SO only from 24 months might be due to the fact that it involves the ability to co-activate two potentially possible and normal scripts for a specific context and switch between the two, implying more complex cognitive processes (e.g. executive functions and inhibition) still developing during childhood (Best and Miller, 2010). Similarly, we observe that SOs related to different knowledge domains are appreciated over time and important differences can be observed between mothers and children. Our data invite a refinement of the humour developmental stages proposed by McGhee (1979), showing that some types of humour, at least in interactional ecological contexts, are accessible to children earlier than previously postulated: we observe misnaming of objects and actions already from 18 months (rather than between 2 and 4 years) and playing with word sounds already from 12 months (rather than between 3 and 5 years). Our results show that laughter in relation to humourous episodes can give us important insights into early cognitive, linguistic and pragmatic development, as well as into the general principles of reasoning acquired and developing in children.

[2] Inter-annotator agreement details in Appendix.

References

Argiris Archakis and Villy Tsakona. 2005. Analyzing conversational data in GTVH terms: A new approach to the issue of identity construction via humor. *Humor*, 18(1):41–68.

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.

John R Best and Patricia H Miller. 2010. A developmental perspective on executive function. *Child development*, 81(6):1641–1660.

Luca Bischetti, Irene Ceccato, Serena Lecce, Elena Cavallini, and Valentina Bambini. 2019. Pragmatics and theory of mind in older adults' humor comprehension. *Current Psychology*, pages 1–17.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/ multi-modal resources with elan. In *LREC*.

J Cunningham. 2005. Children's humor. *Children's play. SAGE publications*.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

Christian F. Hempelmann and Willibald Ruch. 2005. 3 WD meets GTVH: Breaking the ground for interdisciplinary humor research. *Humour*, 18(4):353–387.

Erika Hoff. 2006. How social contexts support and shape language development. *Developmental review*, 26(1):55–88.

Elena Hoicka and Nameera Akhtar. 2012. Early humour production. *British Journal of Developmental Psychology*, 30(4):586–603.

Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in psychology*, 10:123.

Eleni Loizou and Susan L. Recchia, editors. 2019. *Research on Young Children's Humor: Theoretical and Practical Implications for Early Childhood Education*, volume 15 of *Educating the Young Child*. Springer International Publishing, Cham.

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.

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.

Rod A Martin. 2007. *The psychology of humor: An integrative approach*. Elsevier.

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, 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*.

Robert B McCall. 1972. Smiling and vocalization in infants as indices of perceptual-cognitive processes. *Merrill-Palmer Quarterly of Behavior and Development*, 18(4):341–347.

Paul McGhee. 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.

Jean Piaget. 1945. *Play, dreams, and imitation in childhood*. New York: Norton.

Béatrice Priego-Valverde. 2003. L'humour dans la conversation familière: description et analyse linguistiques. *L'humour dans la conversation familière*, pages 1–248.

V. Raskin. 1985. *Semantic mechanisms of humor*, volume 24. Springer.

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.

Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes*. Harvard university press.

Francisco Yus. 2017. Incongruity-resolution cases in jokes. *Lingua*, 197:103–122.

Edward Zigler, Jacob Levine, and Laurence Gould. 1966. Cognitive processes in the development of children's appreciation of humor. *Child development*, pages 507–518.

Limitations

The conclusions from our study should be taken cautiously given the small sample size analysed (4 mother-child dyads over 5 time-points) due to the chronophagus method applied, requiring manual annotations. Moreover, our study is focused exclusively on middle-class American English speaking dyads and we cannot therefore scale our conclusions to any other language and culture given the impact that those factors have both on parenting interactional dynamics (Hoff, 2006) and laughter and humour production and perception (Martin, 2007; Jiang et al., 2019).

Acknowledgements

This work, carried out within the Labex BLRI (ANR-11-LABX-0036) and the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

5 Appendix

5.1 Inter-annotator agreement (IAA)

IAA was conducted among 2 coders in two steps. The first step was aimed at testing agreement on humour episodes identification (i.e. Pleasant incongruity laughables: start- and end-time boundaries), while the second was aimed at testing agreement on the specific classifications in terms of Script Oppositions (SO) and domain. The first phase (Pleasant incongruity/humourous episodes identification and segmentation) was conducted on 20% of the laughter annotations applying the *Staccato* algorithm implemented in ELAN (Lücking et al., 2011).[3] The average degree of organization between annotators is of 0.74. The raw percentage of agreement on whether each laugh (n=47) was related to a humourous laughable, or not, is 93.6% (3 disagreements). For the second step, looking at the specific classification of each humourous laughable in terms of SO and Domain, we asked the second annotator to analyse all the laughables annotated by the first annotator for children and mothers. An *Other* category was offered to all coders, whenever specific humour episodes could not be classified according to the proposed framework. Overall, for SO we obtain a percentage agreement of 92.8 ($\pm$ 1.33) and a total Krippendorff's α of 0.79; for the Domains we obtain a percentage agreement of 90.1 ($\pm$ 5.09) and an overall Krippendorff's α of 0.78. IAA on sub-domains is 95.4% with a Krippendorff α of 0.82. Regarding the IAA on SO classification within each sub-domain , we observe an overall percentage of agreement of 91.9 ($\pm$2.1) and a Krippendorff's α of 0.77. After discussion, annotators came to unanimous agreement on the annotations and those values retained for the current analysis.

[3]We ran the analysis with 1000 Monte Carlo Simulations, a granularity for annotation length of 10, and a = 0.05.

5.2 Distribution of Script Oppositions

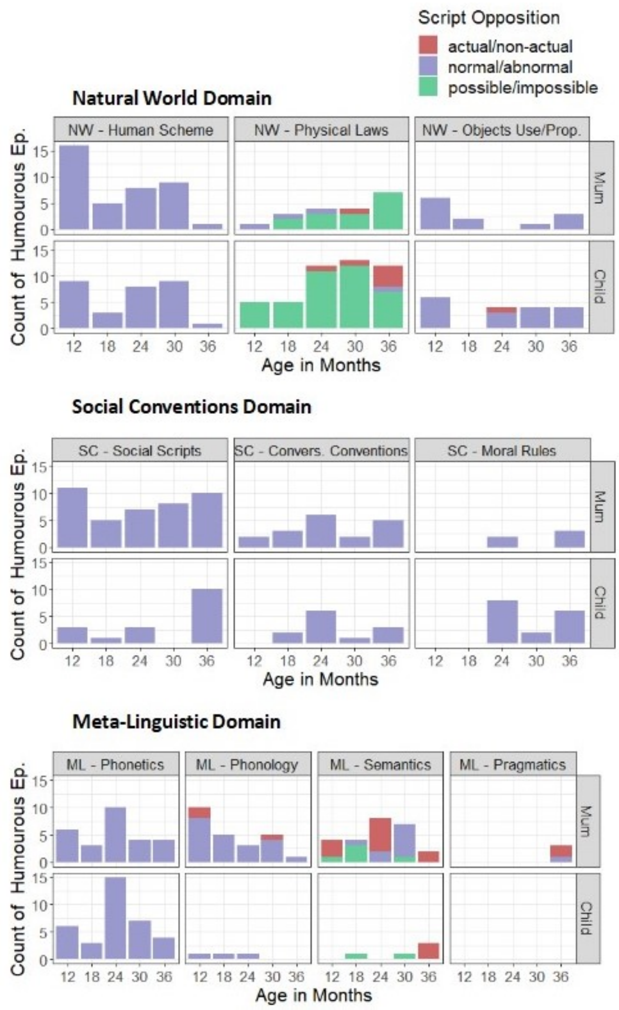


Figure 1: Count of Script Oppositions (*Im/Possible, Ab/Normal, Non/Actual*) over time as a function of knowledge Domain (*Natural World, Social Conventions, Meta-Linguistic*) in Mothers and Children

Characterization of Discourse Salience in English Social Dialogs and its Application to Assessing Interactional Competence of Social Dialog Systems

Alex Lưu
Brandeis University
alexluu@brandeis.edu

1 Characterization of Discourse Salience in English Social Dialogs

[1]To gain better insights into the co-constructed nature of meaning in social conversation, I conducted an empirical study of **discourse salience** in naturally occurring English casual dialogs. First, expert annotators are asked to put themselves in conversational participants' shoes and rely on their communicative competence to recognize the main point (most salient content) in the arguments of *2529* discourse relations annotated in NEWT-SBCSAE, a publicly accessible corpus of naturally occurring casual dialogs in American English (Du Bois et al., 2000; Riou, 2015; Lưu and Malamud, 2020), taking into account the interlocutors' shared social goal as defined in Lưu (2022b)[2]. In addtion, they annotate different linguistic aspects characterizing the salient content of utterances including its directionality (i.e. whether it is backward- or forward-looking) and information packaging (i.e. the given-new ordering of information and syntactic variations for realizing that ordering). The detailed annotation guidelines and outcomes are publicly accessible at https://alexluu.flowlu.com/hc/6/274--discourse-salience/. In this paper, I use the annotated data to systematically characterize discourse salience in English social dialogs, which directly relevant to social dialog system evaluation (Section 2) and modeling (future work).

1.1 Distribution of Discourse Relations

Figure 1 shows the distribution of all annotated discourse relation types. It is clear that social dialog is by no means dominated by question–answer pairs (category 'interaction–query'). In fact, it is full of 'feedback' and utterances functioning across multiple sociocultural dimensions such as 'prominence-2', 'emphasis-repetition', 'positioning-evaluation', 'evaluation', 'positioning' and 'alignment'. These observations has several implications:

- Human users who wants to test a social dialog system should diversify their conversation moves instead of adhering to the question–answer pattern and informational dimension.
- Social dialog systems' conversational strategies should cover all sociocultural dimensions and leverage the power of simple 'feedback'.

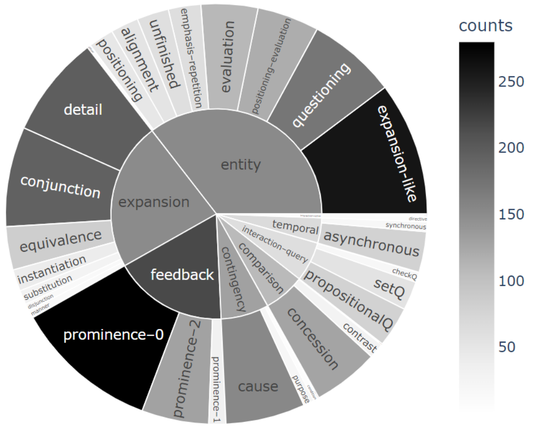


Figure 1: Distribution of discourse relations.

1.2 Directionality of Discourse Salience

The directionality of discourse salience is showed in Figure 2, revealing that backward-looking salient content is much less popular than salient content that is both backward- and forward-looking. In addition, the 2nd argument of non-prominent 'feedback' and Q-relations ('interaction–query' and 'interaction–other') are more probable to be only backward-looking salient content. Hence, salient content that is both backward- and forward-looking is preferred; and if such content is not available, non-prominent 'feedback' is a safe choice.

[1]This paper's live version is located at https://osf.io/yvjgb/.

[2]This goal is to create a coherent experience of together making sense of Self, the Other, and the relationship between them. Within this shared goal, performing a conversational move implies taking a public social act of simultaneously (1) evaluating the subject matter discussed in that move, (2) positioning interlocutors, and (3) aligning with other interlocutors, with respect to any salient sociocultural dimension such as informational, affective or normative.

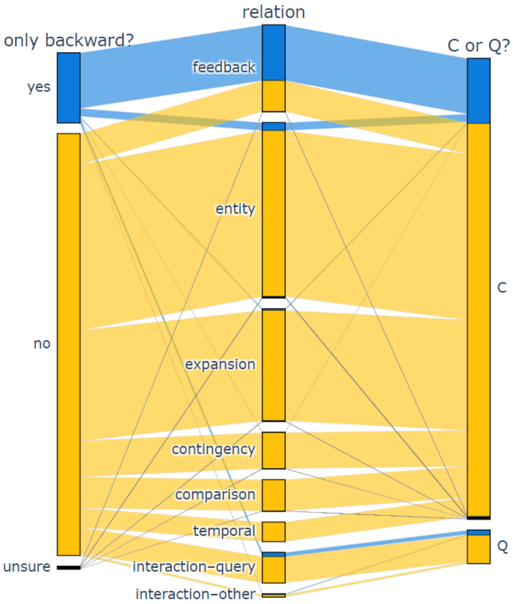


Figure 2: Directionality of at-issue content in the second arguments of informational coherence relations (C and Q are C-relations and Q-relations respectively.

1.3 Information Packaging of Discourse Salience

Figure 3 shows information packaging of discourse salience. The minor portions of new-before-given and noncanonical word order cases confirm the preference of given-before-new information ordering and canonical word order (CWO) in naturally occurring discourse (Prince, 1992; Birner, 2012, inter alia). In addition, all new-before-given cases are realized in CWO, and can be classified in two categories (see more detail in Lưu, 2022a):

- dialogic resonance (Du Bois, 2014)
- non-informational emphasis (Lưu, 2022b)

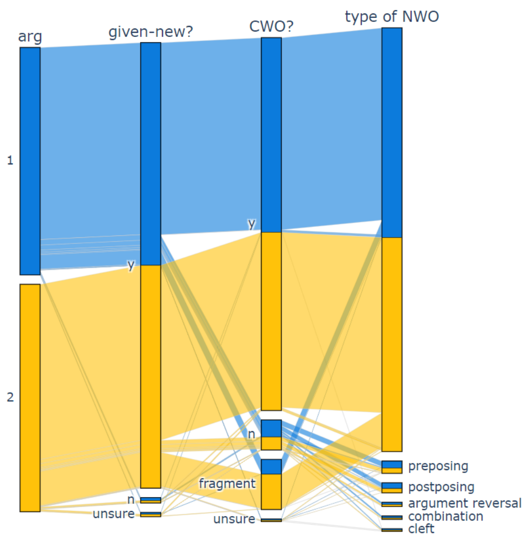


Figure 3: Information packaging of discourse relations.

2 Application to Assessing Interactional Competence of Social Dialog Systems

Questioning the status quo of research on human–computer communication, Kopp and Krämer (2021) argue that we should prioritize modeling the key aspects of mutual understanding in conversation, instead of surface-level behaviors learnable from data. Consequently, adequate evaluation of dialog systems should take into account their **interactional competence** (IC) (e.g. Galaczi and Taylor, 2018, inter alia), which captures the real-time context-sensitivity of interlocutors' meaning interpretation and production.

Based on the characterization of discourse salience presented in Section 1, we can identify a social dialog system's IC by whether its responses:

- pick up on forward-looking salient content in prior discourse
- contribute new content which
  - can be forward-looking salient content or simple feedback
  - is relevant to and consistent with prior discourse with respect to different socio-cultural dimensions (see detailed discussion in Lưu, 2022b, pp.155–157)

In addition, to create an adequate setup for the interaction between human evaluators and social dialog systems, we can adopt the concept of **scaffolding conversation** (Imberi-Olivares, 2012), originally referring to an important learning avenue for children in social interaction and based on the notion of scaffolding in developmental psychology (Vygotsky, 1978; Bruner, 1975; Wood et al., 1976). Being applied to human–computer communication, scaffolding conversation[3] is conducted in such a way that human interlocutors, as the more competent conversants, actively adjust their conversational moves to increase dialog systems' IC. To successfully converse with humans in scaffolding conversation is a realistic goal of social dialog systems, and the analysis of problematic conversational moves can directly inform the systems' improvement. Moreover, scaffolding conversation allows human interlocutors to raise the bar in a systematic and constructive manner when social dialog systems become more and more competent.[4]

[3]This is comparable to a specific strategy of inquiry in the communication game in the Question Under Discussion framework (Roberts, 1996/2012), except for the fact that social conversation is not only about information exchange or inquiry.

[4]Based on the principles proposed in this section, I develop an expert human evaluation protocol publicly accessible here.

Acknowledgements

I am extremely grateful to my annotators, Eben Saveson and Tali Tukachinsky, for their curiosity, diligence and creativity. My deepest gratitude goes to Sophia A. Malamud for her active encouragement and thorough feedback on the project reported in this paper.

References

Betty J Birner. 2012. *Introduction to Pragmatics*. John Wiley & Sons.

Jerome S. Bruner. 1975. The ontogenesis of speech acts. *Journal of Child Language*, 2(1):1–19. Publisher: Cambridge University Press.

John W. Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410. Publisher: De Gruyter Mouton.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa Barbara corpus of spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Evelina Galaczi and Lynda Taylor. 2018. Interactional Competence: Conceptualisations, Operationalisations, and Outstanding Questions. *Language Assessment Quarterly*, 15(3):219–236.

Kristi Ann Imberi-Olivares. 2012. *The process of scaffolding in conversation: who does it best and why?* Ph.D. thesis, UC Merced.

Stefan Kopp and Nicole Krämer. 2021. Revisiting Human-Agent Communication: The Importance of Joint Co-construction and Understanding Mental States. *Frontiers in Psychology*, 12.

Alex Lưu. 2022a. Edge cases of discourse salience in american english casual dialogs: A new window into the co-constructed nature of social conversation. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Dublin, Ireland. SEMDIAL.

Alex Lưu. 2022b. Sketching a linguistically-driven reasoning dialog model for social talk. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 153–170, Dublin, Ireland. Association for Computational Linguistics.

Alex Lưu and Sophia A. Malamud. 2020. Annotating coherence relations for studying topic transitions in social talk. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 174–179, Barcelona, Spain. Association for Computational Linguistics.

Ellen Prince. 1992. The ZPG letter: Subjects, definiteness, and information status. In William C. Mann and Sandra A. Thompson, editors, *Discourse Description: Discourse Analyses of a Fundraising Text*, pages 295–325. Amsterdam: John Benjamins.

Marine Riou. 2015. *The Grammar of Topic Transition in American English Conversation. Topic Transition Design and Management in Typical and Atypical Conversations (Schizophrenia)*. Ph.D. thesis, Université Sorbonne Paris Cité.

Craige Roberts. 1996/2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:6–1.

L. S. Vygotsky. 1978. *Mind in society: the development of higher psychological processes*. Harvard University Press, Cambridge, MA, USA.

David Wood, Jerome S. Bruner, and Gail Ross. 1976. The Role of Tutoring in Problem Solving*. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

`rezonateR`: An R package for analysing coherence in conversation

Ryan Ka Yau Lai
University of California, Santa Barbara
`kayaulai@ucsb.edu`

1 Introduction

`rezonateR` is an R (R core team 2022) package working with complex data annotations, geared towards discourse and interactional linguists examining topics like dialogic resonance, turn-taking, and reference tracking. It aims to bridge the gap between data from modern multilayer corpus annotations, which usually take on complex graph formats, and features arranged in a tabular format used in visualisation, statistical analysis, and machine learning environments, which linguists need to answering particular research questions. `rezonateR` takes annotations from the visual annotation environment Rezonator (DuBois 2019, DuBois et al. 2020), transforms the graph into a relational database-like format, and offers a wide range of functions for generating features used in discourse research.

2 Features

The first step of working with Rezonator annotations in `rezonateR` is to import Rezonator's native `.rez` format using the `importRez()` function. This creates an object that contains, among other information, a series of data frames, each of which corresponds to a node type in Rezonator's underlying graph structure. Semi-automatic annotations can be added to these data frames by first guessing the values in R, then using `rez_write_csv()`, `rez_read_csv()` and `updateFromDF()` to export it as a `.csv`, edit it in a spreadsheet, and incorporate the edits in R.

After import, `rezonateR` contains numerous functions for deriving features from the imported annotations. Two sets of generic functions are available for data wrangling (e.g. combining information from different node types in the annotations): the EasyEdit series for base R users, and the TidyRez series for `tidyverse` (Wickham et al. 2019) users.

Beyond these basic features, `rezonateR` contains functions for analysing more specific discourse questions. Figure 1 shows three main structures in Rezonator: stacks (annotations of text segmentations), tracks (coreference chains), and dialogic resonance (DuBois 2014).

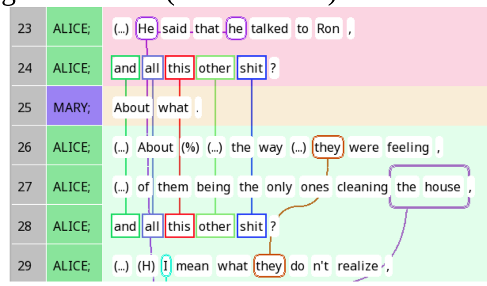


Figure 1: Sample Rezonator text (SBC007) with stacks (background colours indicating turns), resonances (straight-line connections between words), and tracks (curved lines between mentions).

Stacks represent discourse units (e.g. turns). `rezonateR` can compute values like positions of tokens within stacks, optionally excluding non-word tokens like pauses and punctuation; this is useful when e.g. investigating a form's function through its position within large structures (e.g. Kim 2022). For dialogic resonance, `rezonateR` can find resonances between parts of a sequence (e.g. between first and second assessments), and calculate resonance-related statistics used in studies like Tantucci & Wang (2021). For tracks, it contains a rich set of functions for deriving predictors for coreference-related issues like referential choice, e.g. extracting the distance to or a property of the last mention or counting recent mentions (possibly conditionally, e.g. subjects only) within a window of lines. The case study below demonstrates how `rezonateR` deals with the first two annotation types.

3 Sample analysis

To demonstrate the use of various functions in `rezonateR`, this sample analysis examines responsiveness in the seventh conversation from the Santa Barbara Corpus of Spoken American English (SBC007; DuBois et al. 2000). Question-answer sequences are perhaps the clearest examples of responsiveness, since a question socially obligates a response. I began by identifying all the question-answer sequences using stacks, and tagged the stacks for the action they implement (e.g. information-seeking question, confirmation request, other-initiation of repair). Three types of questions were identified as the most common in the text: Information-seeking questions, ritualised expressions of disbelief (Wilkinson & Kitzinger 2006), and soliciting the recognition of a reference (Heritage 2007).

To analyse the formal correlates of responsiveness, I examined two linguistic devices: discourse markers (DMs) and resonance. I annotated all resonance in the text in Rezonator. `rezonateR`'s functions for combining various parts of the annotations (`rez_left_join()`, `findResonancesBetween()`, `stackToToken()`) were used to produce the following graphs: the number of resonance chains associated with each Q-A sequence type (Figure 2), and word tokens found in Q-A sequences of each category (Figure 3).

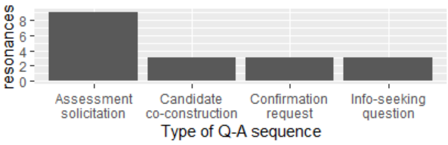


Figure 2: Resonance count for four Q-A types; remaining types have no resonance. Only 18 resonances were in Q-A sequences, out of 234 total.

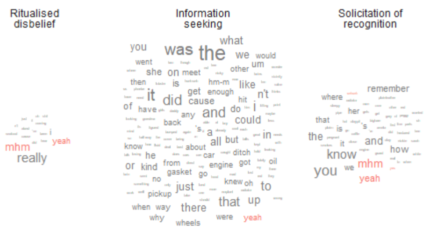


Figure 3: Word clouds by Q-A type (DMs in red).

Since little resonance is associated with Q-A pairs in this dataset, I then focused on analysing the DMs, including outside Q-A sequences. The two most common discourse markers in Q-A sequences were *yeah* and *mhm*. Since *yeah* seems to be used more for information-seeking questions and *mhm* for the other two with more regulatory functions, this may hint at a more general pattern regarding the distribution of *yeah* vs *mhm*, further supported by fact that *yeah* seems to appear more frequently in longer turns (Figure 5). To further investigate this, each instance of the DMs was tagged according to the epistemic gradient between asker and answerer (HIERARCHY, after Gadanidis et al. 2023); whether it was elicited by the other party (e.g. with interrogative syntax or rising intonation), responding to previous speech with no explicit invitation for a response, or simply pointing back to one's own speech (RESPONSIVENESS); whether the speaker was expressing affiliation with or understanding of the speech she was responding to, or some other stance (STANCETYPE); the DM's position in a sequence (second pair part (SPP), sequence-closing third, other; SEQPOSITION). The text was also annotated for turns using stacks. The DM's position in the intonation unit (IU) and the IU's position in the turn were automatically derived in `rezonateR` using data wrangling functions and `getOrderFromSeq()`. Hierarchical clustering with complete linkage revealed two layers of interpretable clusters. The first (k = 2) divides DMs with substantive semantic contribution from those with primarily regulatory functions. The second (k = 5) divides regulatory cases into backchannels and follow-ups to one's own prior talk, and substantive cases into SPPs to information-seeking questions, non-information-seeking questions, and thirds. Figure 4 shows the distribution of *yeah* and mhm within these categories, supporting the above-mentioned association of *yeah* with greater substantiveness.

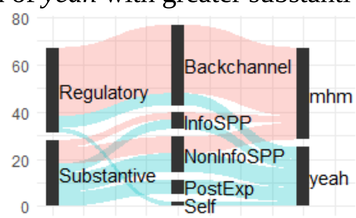


Figure 4: Sankey diagram of the two clusterings and the distribution of *yeah* and *mhm* within each.

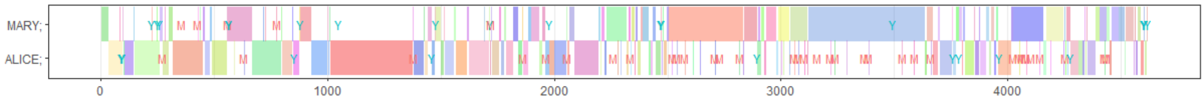


Figure 5: Gantt chart, produced with `rezonateR`'s `getGantt()`, of *yeah* (Y) and *mhm* (M)'s locations.

References

DuBois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson & Nii Martey. 2000. Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium. Linguistic Data Consortium.

DuBois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics* 25(3). 359–410. https://doi.org/10.1515/cog-2014-0024.

DuBois, John W. 2019. Rezonator: Visualizing Resonance for Coherence in Dialogue. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue. London: SEMDIAL*.

DuBois, John W., Terry DuBois, Georgio Klironomos & Brady Moore. 2020. From answer to question: Coherence analysis with Rezonator. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue*.

Gadanidis, Timothy, Angelika Kiss, Lex Konnelly, Katharina Pabst, Lisa Schlegl, Pocholo Umbal & Sali A. Tagliamonte. 2023. Integrating qualitative and quantitative analyses of stance: A case study of English that/zero variation. *Language in Society* 52(1). 27–50.

Heritage, John. 2007. Intersubjectivity and progressivity in person (and place) reference. In N. J. Enfield & Tanya Stivers (eds.), *Person Reference in Interaction*, 255–280. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486746.012.

Kim, Mary Shin. 2022. Identical linguistic forms in multiple turn and sequence positions in Asian languages. *Journal of Pragmatics* 200. 1–7. https://doi.org/10.1016/j.pragma.2022.06.007.

R Core Team. 2022. *R: A language and environment for statistical computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tantucci, Vittorio & Aiqing Wang. 2021. Resonance and engagement through (dis-)agreement: Evidence of persistent constructional priming from Mandarin naturalistic interaction. *Journal of Pragmatics* 175. 94–111.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43). 1686. https://doi.org/10.21105/joss.01686.

Wilkinson, Sue & Celia Kitzinger. 2006. Surprise As an Interactional Achievement: Reaction Tokens in Conversation. *Social Psychology Quarterly* 69(2). 150–182. https://doi.org/10.1177/019027250606900203.

A Framework for Confusion Mitigation in Task-Oriented Interactions

Na Li
Technological University Dublin
na.li@tudublin.ie

Robert Ross
Technological University Dublin
robert.ross@tudublin.ie

Abstract

Confusion is a mental state that can be triggered in task-oriented interactions and which can if left unattended lead to boredom, frustration, or disengagement from the task at hand. Previous work has demonstrated that confusion can be detected in situated human-robot interactions from visual and auditory cues. Therefore, in the next step, we propose appropriate interaction structures in this study, which should be used to mitigate confusion. We motivate and describe this dialogue mechanism through an information state-style dialogue framework and policies, and also outline the approach we are taking to integrate such a meta-conversational goal alongside core task-oriented considerations in modern data-driven conversational techniques.

1 Introduction

While we have a keen common sense intuition of what it means to be confused, the concept theoretically has only had some study in affective sciences: From a positive perspective, confusion is an effective response that occurs in people willing to explore new knowledge or tasks, but it is also an epistemic emotion that is associated with cognitive impasses when people try to solve problems (Lodge et al., 2018). The effects of the confusion state have been studied in online learning and driver assistance (Grafsgaard et al., 2011; Atapattu et al., 2020; Hori et al., 2016), but to date, the amount of research on confusion focused on the dialogue domain has been limited. One potential reason for the limited systematic study of confusion in the dialogue community may be that confusion is arguably better detected and more relevant in physically embodied interactions such as with robotic systems, although in this domain, research to date has been limited. In previous research (Li and Ross, 2023a), we have shown that it is possible to systematically identify users in a state of confusion, at least in a controlled study. If we can directly detect confusion as a cognitive state in interactions, the question then becomes: How should we train or otherwise adjust our dialogue policy to mitigate that confusion? Certainly, some of this mitigation would factor into the design cycle where we measure user confusion during initial interactions and adjust task designs to reduce the potential for confusion, but we also need to allow for the fact that confusion will occur (particularly in educational or training settings (D'Mello et al., 2014)) and that the conversational policies deployed must be able to dynamically adjust to the user in a confused state.

Given this challenge, in this paper we present a policy framework for the mitigation of confusion in task-oriented interaction. The policy framework builds on some design concepts from classical information state (IS) and dialogue acts representations from Dynamic Interpretation Theory (DIT), and Dialogue Act Markup in Several Layers (DAMSL). Our intuition for designing our dialogue framework builds on IS dialogue models and related toolkits such as TrindiKit, which is a dialogue move engine toolkit and the IBiS system (issue-based dialogue system) (Traum and Larsson, 2003; Larsson, 2002).

We first outline a set of relevant atomic information state and dialogue acts specifications; we then outline an information state structure including dialogue moves, and formalise the detailed dialogue policies corresponding to the dialogue acts. Following that, we illustrate the proposed approach using several scenarios as case studies. While the approach is very much a classical perspective, this is simply a stepping stone for us to providing aligned behaviours in data-driven policies.

2 Information State & Framework Design

In this study, the information state represents cumulative additions from previous actions in dialogue, and also the mechanisms to trigger dialogue moves for activating a corresponding dialogue act.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue,
August, 16–17, 2023, Maribor.

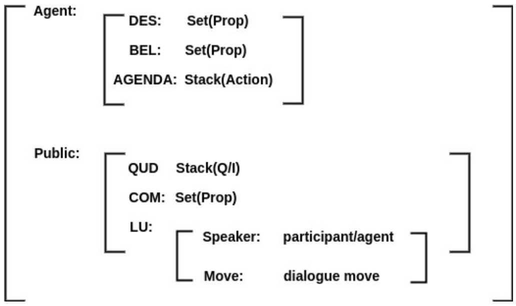


Figure 1: IS structure for confusion mitigation

Figure 1 presents an overview of the information state structure that we assume. The information state structure is typical of many other information state proposals such as IBiS, but for the sake of clarity, we briefly summarise for the unfamiliar reader. At a high-level the information state is split between a private grouping of state variables (Agent) which are internal to the agent and a public grouping of variables (Public) which the dialogue model assumes are shared between both agents. Within the private entities, the field /Agent/DES (desire) is a set of propositions (prop) that are used to capture the goals that the agent wishes to achieve. The field /Agent/BEL (believe) is a set of propositions that are directly correlated to the task that is taken to be true. Finally, the field /Agent/Agenda is a stack of plans which the agent intends to enact in order to achieve dialogue goals or otherwise lead to manipulate the mental state.

Turning to the public elements of the information state, the field /Public/QUD is a stack of questions under discussion (QUD). The QUD encompasses the ordering of unresolved questions or tasks to be confirmed that have been raised within the dialogue. The field /Public/COM includes a set of propositions that the user and the agent have committed to in the dialogue. It is not necessary for discourse participants to genuinely believe in those propositions, but discourse participants should have made a commitment to those statements for the objectives of the conversation. Finally, the field /Public/LU simply captures the last utterances in terms of the speakers and the specific dialogue moves associated with the utterances and the specific dialogue moves associated with the utterance.

Building on Larsson (2002)'s IBiS1 model, our dialogue moves are coarse-grained operations that trigger updates to the information state and the selection of relevant dialogue acts. Therefore, we designed ten dialogue moves and nine dialogue acts in our technical report (Li and Ross, 2023b), which can be applied across four information types (*i.e.*, statement, feedback, generic, and interface), to operationalise a policy to mitigate user confusion states. Our technical details include two tables (*i.e.*, Table 1 and Table 2) in Li and Ross (2023b) outline the general form of communication updates associated with these dialogue acts and the specific updates related to confusion states, respectively.

In that report, we detail a dialogue management process that is based on these definitions. A confusion detection model is assumed and integrated into the dialogue framework for real-time detection of the user's confusion states. Our model assumes semantically distinct levels of productive confusion, unproductive confusion, and non-confusion. When a confusion state is detected, this aspect of the dialogue policy becomes active. This structuring is in accordance with similar elements of communicative management in those moves and acts are selected to achieve the interaction goal of mitigating the user's confusion state. When an interlocutor is not manifesting confusion behaviours, the dialogue policy proceeds with those moves and acts associated with task progression as outlined. Moreover, we also present a task-oriented dialogue scenario in that report with associated updates of dialogues to help elucidate the policy presented.

3 Discussion & Outlook

In this paper, the proposed models and the underlying report have been designed and applied at a conceptual and empirical level in part of our human-avatar and human-robot studies. While the key motivators for these earlier studies were (a) whether confusion states can be induced; and (b) whether it is possible to detect confusion states extraverbally.

The policy presented here is to highlight one way in which we can identify and mitigate confusion as a pragmatic phenomenon that can be identified. While the benefit of a controlled dialogue flow remains important, we do recognise the importance of folding in the goals of embodied structured conversation with the naturalness and task-oriented appeal of integration with large language model-based solutions. Although the current proposal is still embryonic and not in a state where it can be systematically evaluated, we believe that the study of pragmatic effects in embodied systems presents an important next step for the study of the semantics and pragmatics of dialogues.

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

Thushari Atapattu, Katrina Falkner, Menasha Thilakaratne, Lavendini Sivaneasharajah, and Rangana Jayashanka. 2020. What do linguistic expressions tell us about learners' confusion? a domain-independent analysis in moocs. *IEEE Transactions on Learning Technologies*, 13(4):878–888.

Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170.

Joseph F Grafsgaard, Kristy Elizabeth Boyer, and James C Lester. 2011. Predicting Facial Indicators of Confusion with Hidden Markov Models. Technical report.

Chiori Hori, Shinji Watanabe, Takaaki Hori, Bret A. Harsham, JohnR. Hershey, Yusuke Koji, Yoichi Fujii, and Yuki Furumoto. 2016. Driver confusion status detection using recurrent neural networks. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Staffan Larsson. 2002. Issue-based dialogue management.

Na Li and Robert Ross. 2023a. Hmm, you seem confused! tracking interlocutor confusion for situated task-oriented hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 142–151, New York, NY, USA. Association for Computing Machinery.

Na Li and Robert Ross. 2023b. Technical report: A framework for confusion mitigation in task-oriented interactions.

Jason M. Lodge, Gregor Kennedy, Lori Lockyer, Amael Arguel, and Mariya Pachman. 2018. Understanding Difficulties and Resulting Confusion in Learning: An Integrative Review. *Frontiers in Education*, 3.

David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. *Springer Netherlands eBooks*.

“Are you telling me to put glasses on the dog?” Content-Grounded Annotation of Instruction Clarification Requests in the CoDraw Dataset

Brielen Madureira[1] **David Schlangen**[1,2]
[1]Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
{madureiralasota,david.schlangen}@uni-potsdam.de

Abstract

Instruction Clarification Requests are a mechanism to solve communication problems, which is very functional in instruction-following interactions. Recent work has argued that the CoDraw dataset is a valuable source of naturally occurring iCRs. Beyond identifying when iCRs should be made, dialogue models should also be able to generate them with suitable form and content. In this work, we introduce CoDraw-iCR (v2), extending the existing iCR identifiers with fine-grained information grounded in the underlying dialogue game items and possible actions. Our annotation can serve to model and evaluate repair capabilities of dialogue agents.

Introduction If someone requests you to put glasses on a dog, you may doubt yourself: *Is that really what I am supposed to do?* Before attempting that, you’d likely seek confirmation, for instance, by posing a clarification request. In real life, dogs do fine without glasses, but, as we see in Figure 1, that is indeed a correct action in the context of a scene construction dialogue game.

In instruction following settings, ambiguous or underspecified instructions may elicit clarification requests when the instruction follower realises they cannot act properly without further information. These are Instruction Clarification Requests (iCRs), namely CRs that occur in Clark’s 4th level of communication (Clark, 1996), when an utterance (here, an instruction) is understood generally, but not at the level of uptake (Schlöder and Fernández, 2014).

We have recently argued that the CoDraw dataset (Kim et al., 2019) is a rich and large source of spontaneous iCRs (Madureira and Schlangen, 2023). We identified iCRs among all instruction follower utterances and proposed using the annotation to model the tasks of knowing *when* to ask and to reply to an iCR. However, knowing *what* and *how* to ask are also topical devices for a competent instruction follower dialogue model. To account for that, we continue this initiative by adding information about the *content* and *form* of iCRs, in order to allow modelling and evaluating the subsequent task of *generating* iCRs, not yet explored in this corpus.

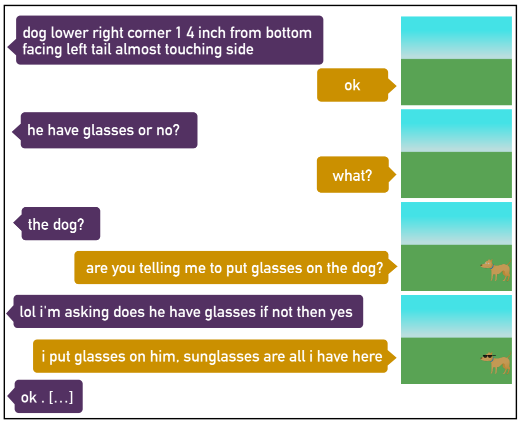
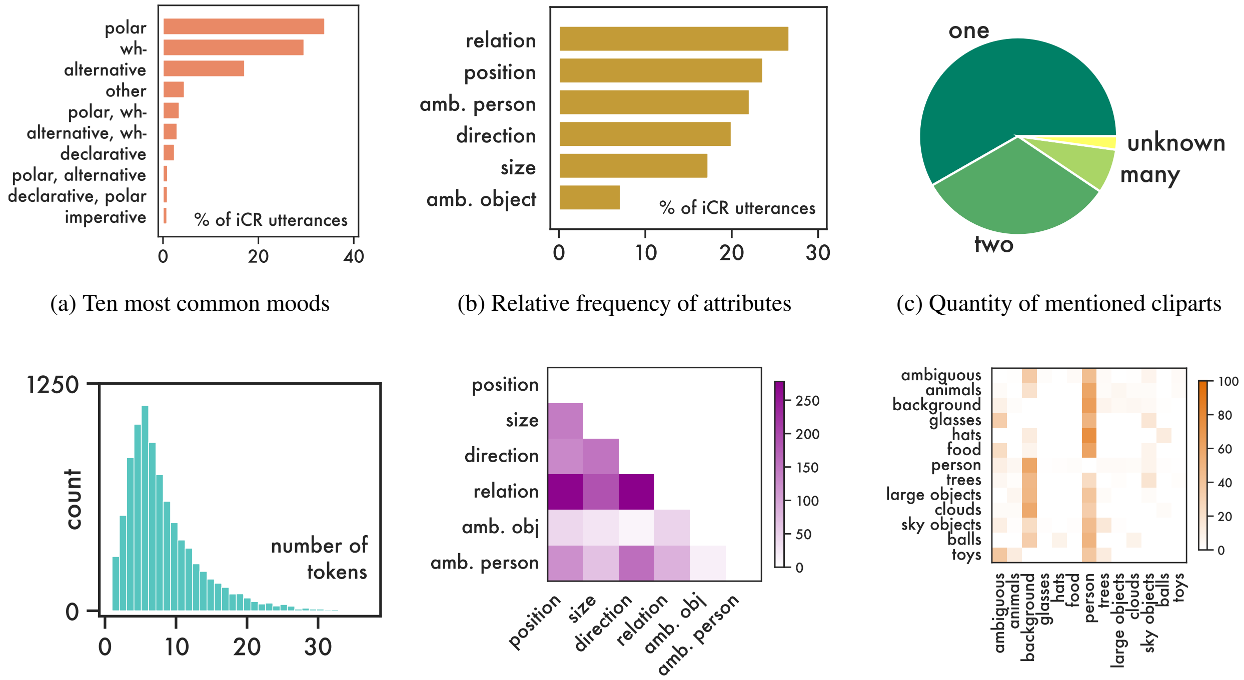


Figure 1: A communication problem occurring and being resolved with the aid of clarification requests in an instruction following interaction (CoDraw, ID 9429, CC BY-NC 4.0, scene from Zitnick and Parikh (2013)). When an instruction is not clear enough, the instruction follower asks for clarification, in order to act accordingly (here, placing cliparts in the scene).

Our annotation complements CoDraw-iCR (v1) by adding mood categories and by mapping each utterance to its corresponding objects and action-related attributes. We show that this sample is an appealing ensemble of mostly unique surface forms through which interesting relations in co-occurring objects and attributes emerge, making it a handy resource for further CR research. The data and documentation is available for the community at `https://osf.io/gcjhz/`, which also contains a link to an extended version of this summary.

Background Clarification Requests are a multifaceted phenomenon in dialogue, with vast literature on categorising, documenting and modelling their various realisations as well as their relations to other utterances and to the context (Purver et al., 2003; Gabsdil, 2003; Rodríguez and Schlangen,



(a) Ten most common moods

(b) Relative frequency of attributes

(c) Quantity of mentioned cliparts

(d) Distribution of length

(e) Frequency of co-occurring attributes.

(f) Co-occurrence of thematic entity categories, distribution by row in %.

Figure 2: Overview of the distributions of annotated categories in CoDraw iCR utterances.

2004, *inter alia*). Still, it remains an open research area; in particular, we cannot delineate yet to what extent CR mechanisms can be learnt via data-driven methods (Benotti and Blackburn, 2021), and dealing with underspecifications is still hard for pre-trained language models (Li et al., 2022).

Benotti and Blackburn (2021) have recently raised awareness to the different world modalities upon which clarifications can be *grounded*, like vision, movement or physical objects. Still, few works exist that systematically map the content of CRs to elements related to the context where they occur (Gervits et al., 2021). Some examples are Benotti and Blackburn (2017), who use a methodology to classify CRs according to why they make implicated premises explicit (*e.g.* wrong plan, not explainable plan or ambiguous plan in instruction giving), in a corpus that is further analysed in Benotti and Blackburn (2021) with a recipe to detecting *grounded clarifications*. Gervits et al. (2021) propose a fine-grained annotation schema for CR types related to the environment (object location, feature, action, description, etc). The small size of these corpora, however, does not meet the needs of current data-driven methods.

Corpus Overview 8,765 utterances (7,710 types) were identified as iCRs in CoDraw. Figure 2 presents an overview of the annotation. The immediately preceding instruction giver utterance is the source utterance (*i.e.* the utterance where the communication problem manifests) for 80.26% of the iCRs and 78.49% of the iCR utterances get a response from the instruction giver in the immediately following turn. For 63.85% of them, both conditions are true. They are realised in many surface forms, ranging from short and generic (*sorry?*), to very specific (*owl is med?*), to long and verbose (*is the girl sitting or standing i need to know as there are multiple options and her expression as well*). Besides, the iCRs cover all available objects and are well distributed among actions.

Outlook Given the need for large scale corpora for data-driven methods, trading some of the ecological validity in the annotation process for machine-learnability was necessary. Still, even in its controlled environment with a limited number of actions and objects, the resulting iCR utterances are very diverse in surface form and very fertile in content. With the release of the annotation, the community gains a larger resource with sequential, spontaneous iCRs in turn-based dialogues. We aim to encourage more research on modelling CRs in instruction following interactions, and also to enable detailed evaluation of iCR generation.

Acknowledgements

We are very thankful to our student assistants, Sebastiano Gigliobianco, for performing the initial identification and implementing the GUI and Sophia Rauh, for doing the fine-grained annotation. We also thank the anonymous reviewers for their feedback.

References

Luciana Benotti and Patrick Blackburn. 2017. Modeling the clarification potential of instructions: Predicting clarification requests and other reactions. *Computer Speech & Language*, 45:536–551.

Luciana Benotti and Patrick Blackburn. 2021. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

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

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35.

Felix Gervits, Antonio Roque, Gordon Briggs, Matthias Scheutz, and Matthew Marge. 2021. How should agents ask questions for situated learning? an annotated dialogue corpus. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 353–359, Singapore and Online. Association for Computational Linguistics.

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

Haau-Sing Li, Mohsen Mesgar, André FT Martins, and Iryna Gurevych. 2022. Asking clarification questions for code generation in general-purpose programming language. *arXiv preprint arXiv:2212.09885*.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2303–2319, Dubrovnik, Croatia. Association for Computational Linguistics.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*.

Julian Schlöder and Raquel Fernández. 2014. Clarification requests at the level of uptake. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Edinburgh, United Kingdom. SEMDIAL.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016.

Menstruating vampires: What talk about taboos can tell us about dialogue

Christine Howes, Vladislav Maraev and Ellen Breitholtz
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
christine.howes@gu.se, vladislav.maraev@gu.se, ellen.breitholtz@ling.gu.se

1 Introduction

We argue that humour cannot be understood – or analysed – without considering the interactional and cognitive resources, including low-level repair mechanisms and higher-level inferences, which underpin any conversational exchange. To account for incongruity – often argued to give rise to humorous effect – we use common-sense inferences linking background knowledge, beliefs and context to the ongoing dialogue. How any utterance should be interpreted relies on underpinning assumptions warranting these inferences, principles of reasoning called *topoi* (Aristotle, ca. 340 B.C.E./2007; Ducrot, 1988). Topoi are cultural affordances accessible to members of a community which licence certain inferences. Accessing or accommodating an appropriate topos to interpret an utterance is crucial for successful dialogue (Breitholtz, 2020). As there is usually more than one potentially applicable topos this can lead to a mismatch between interlocutors' interpretations (Breitholtz et al., 2017). This potential for mismatch is exploited in the case of humour where it may result in incongruity (Attardo and Raskin, 1991; Maraev et al., 2021).

2 Taboo

In any community, there are subjects which it is not normally considered acceptable to talk about. These may be repulsive (e.g. faeces, vomit) or actions that are considered morally deviant (e.g. cannibalism, incest). One such taboo that we will focus on in this paper is menstruation.

What counts as a taboo depends heavily on the context of the interaction. For example, bodily functions may be the legitimate subject under discussion between a doctor and their patient. Taboos are also gradient with certain topics more or less improper depending on the situation, including the (social) identities of the participants.

For example, discussion of menstruation may be unproblematic or humorous between women, but embarrassing or offensive with men present. How we negotiate the contexts which situate how mentions of taboo subjects are interpreted demonstrates the complex interplay of personal relationships and identities that we navigate in all interactions.

3 Menstruation and humour

One way in which talk about taboos can be licensed is through humour, and this can be facilitated by using readily available topoi which share some underlying features. Menstruation is no exception, and can be constructed as dirty/mysterious in jokes: *"Confucius says never turn your back on anything that can bleed for five days straight and still live"* ' (Bemiller and Schneider, 2010).

As with any domain which is as essential to the human condition there are a large number of topoi associated with menstruation. These are culturally and context specific. For example, in some contexts menstruation signals fertility (a woman on her period has not reached menopause), but in other contexts signals a lack of fertility (getting one's period when one is hoping to conceive). In situations which require one such topos to be accommodated for the discourse meaning to be correctly interpreted humans generally have no problems identifying and accommodating a relevant topos.

4 Talk about tampons

Our first example, taken from the British National Corpus 2014 (BNC2014; Love et al., 2017) occurs in a family home, between a mother (F1) and her daughters (F2 and F3), along with a male family friend (M1). Also present are the father of the family and a 32 year old male. The dialogue at the start between the mother and her daughters, is quite matter of fact about the tampons. It is only when two of the men join the conversation with non-sequitur questions and bad puns that F1 and F2 laugh to defuse any potential awkwardness.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

(1) From BNC2014 SE68. F1 (female, 49), F2 (female, 24), F3 (female, 21), M1 (male, 53)

1 **F1** why have we got a packet of this Tampax here?
2 **F2** it's not mine <F3/>'s
3 **F1** are those yours <F3/>?
4 **F3** yeah
5 **F1** on the table <laugh/>
6 **UNKMALE** what are they for then?
7 **M1** are we are we eating in there?
8 **F3** yeah Tampax
9 **UNKMALE** yeah
10 **F3** for your periods
11 **UNKMALE** oh
12 **F2** <laugh/>
13 **M1** well you did ask
14 **F1** box of Tampax on the table I take it <F3/>'s on her period
15 **UNKMALE** do you get them periodically?
16 **F2** <laugh/> oh that's a good one
17 **F1** <laugh/>
18 **F3** funny
19 **F2** <laugh/>

4.1 Vampires and menstruation

The aim of the dialogue in (2) is to be funny, but without using obviously scripted jokes. To understand the joke, you have to know that vampires are associated with Transylvania, that Vlad the Impaler was also known as Vlad Dracula (and that Dracula is a vampire), and about the infamous hacked phone conversation (then) Prince Charles had with his (then) mistress Camilla ("tampongate").

(2) BBC Radio 4 Friday Night Comedy "The News Quiz" 5th May 2023 discussing the coronation of King Charles

1 **Ria Lina:** ...Charles likes to holiday in Transylvania and he's paid for this brand new really eco friendly water treatment for this village <laugh> not that we have a problem with our waterways at all, do we?
2 **Audience:** <laughter>
3 **Ria Lina:** Erm I'm not saying that it proves he wants to be a vampire because it's in Transylvania
4 **Audience:** <laughter>
5 **Ria Lina:** [He goes every year to Transylvania]
6 **Rachel Cunliffe:** [He he] goes every year to Transylvania and he is distantly related to Vlad the Impaler
7 **Ria Lina:** Is he?
8 **Rachel Cunliffe:** Yeah
9 **Ria Lina:** Well I'm not saying it proves he wants to be a vampire but he did once want to be a tampon, so.
10 **Audience:** <laughter and groaning>

In contrast to our previous examples, the context of example (3) is a specific joke-telling one, between teenagers. Adolescents gain knowledge about menstruation and related experiences from their peers through talking, storytelling and joking (Fingerson, 2012).

(3) BNC KPG 2498-2529 Josie (14, F), Shelley (15, F), Sean (12, M)

1 **Josie** Right, three vam , a vampire walks into a pub and goes erm
2 **Shelley** Oh yeah. I know.
3 **Josie** excuse me, <mimicking Romanian accent> I want a pint of blood.
4 **Shelley** Yeah.
5 **Josie** And the man goes sorry mate we don't do blood. And he goes, I want a pint of blood! So the man goes <pause> ah, chops the dog's head off.
6 **Unknown** <laugh>
7 **Josie** Sticks it in the cup, goes and gives it to him, he goes, <mimicking Romanian accent> thank you. And goes and sits in the corner. Second vampire comes in, <mimicking Romanian accent> I want a pint of blood.
8 **Unknowns** <laugh>
9 **Josie** He goes alright. Gives it to him. He goes, <mimicking Romanian accent> thank you , and go and sit down. Third vampire comes in, right, the other one goes and sits down, the third one comes in, he goes <pause> yo! What's going down man? I want a pint of water. He goes, pardon? He goes, I want a pint of water.
10 **Shelley** <laugh>
11 **Josie** So he gives him a pint of water, he goes and sits with the other ones. And the other ones look at him, and they sort of look in their cups and going <pause> er, how comes we got blood <pause> and you got water? <laugh> <pause> He goes, nah mate! Ain't you lot ever heard of tea bags? And he puts a Tampax in the water.
12 **Unknowns** <laugh>
13 **Unknown** Very good.
14 **Sean** How comes your jokes are sick?

5 Conclusions

In this paper we have provided examples of dialogues in which menstruation is discussed. This topic can cause a range of responses from embarrassment to enjoyment which are not because of the topic per se but rather a complex interplay of the context and purpose of the dialogues and the interlocutors and their inter-relationships across a range of dimensions. These include factors about the speaker (are they a member of an in-group or out-group? What is the projected persona?), relationships between speakers and their roles (How intimate are they? Are they performing roles associated with particular rights and obligations, e.g. teacher-student)

Although there is an increased interest in incorporating such socio-cultural knowledge and beliefs in semantic analyses of language (see Burnett, 2020; Davis and McCready, 2020; Noble et al., 2020) there is, as yet, no formal theory which encompasses all of these factors. As ever, much remains to be done.

Acknowledgements

This work was supported by an ERC Starting Grant, *DivCon: Divergence and convergence in dialogue: The dynamic management of mismatches* (10977927) and a Swedish Research Council Grant for the establishment of the *Centre for Linguistic Theory and Studies in Probability (CLASP)* (2014-0039).

References

Aristotle (2007). *On Rhetoric, a theory of civic discourse* (translated by George A. Kennedy). Oxford University Press, Oxford. (original work published ca. 340 B.C.E.).

Attardo, S. and Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3-4):293–348.

Bemiller, M. L. and Schneider, R. Z. (2010). It's not just a joke. *Sociological Spectrum*, 30(4):459–479.

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

Breitholtz, E., Howes, C., and Cooper, R. (2017). Incrementality all the way up. In *Proceedings of the Computing Natural Language Inference Workshop*.

Burnett, H. (2020). A persona-based semantics for slurs. *Grazer philosophische studien*, 97(1):31–62.

Davis, C. and McCready, E. (2020). The instability of slurs. *Grazer Philosophische Studien*, 97(1):63–85.

Ducrot, O. (1988). Topoï et formes topique. *Bulletin d'études de la linguistique française*, 22:1–14.

Fingerson, L. (2012). *Girls in power: Gender, body, and menstruation in adolescence*. State University of New York Press.

Love, R., Dembry, C., Hardie, A., Brezina, V., and McEnery, T. (2017). The spoken bnc2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

Maraev, V., Breitholtz, E., Howes, C., Larsson, S., and Cooper, R. (2021). Something old, something new, something borrowed, something taboo: Interaction and creativity in humour. *Frontiers in Psychology*, 12:1443.

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

Are metadiscourse dialogue acts a category on their own?

Darinka Verdonik, Simona Majhenič, Andreja Bizjak
University of Maribor, Facutly of Electrical Engineering and Computer Science
{darinka.verdonik, simona.majhenic, andreja.bizjak1}@um.si

Abstract

Annotation of dialogue acts was stimulated by computational work in building dialogue systems, even though theoretically the notion has its roots in the pragmatic speech act theory. We will be dealing with a special type of dialogue acts that cannot be described in terms of expressed intention, and can be described as non-topical linguistic material, for example, acts like *Well listen* or *I don't know how to say this*. We propose to use the cover term metadiscourse dialogue acts for this type of acts. We present empirical experiences from dialogue act annotation on Slovenian data.

1 Introduction

The meanings of utterances can be interpreted in terms of their functions, like, for example:

- A: *Omake še imaš kaj?* 'Do you have some more sauce?' can be interpreted as eliciting information or as a request for more sauce,
- B: *Ja.* 'Yes,' can be interpreted as confirmation,
- B: *Eee Ana saj je žlica tam pa si vzemi.* 'Uhm the spoon is there, Ana, take some,' can be interpreted as directive, etc.

One of the most known and early theories that has drawn attention to this level of meaning was Austin's (1975) speech act theory, in which an illocutionary act is considered as the "performance of an act in saying something" (Austin 1975: 99); for example, apologising, offering help, stating information, etc. However, when faced with real-life data, the five basic speech act categories of the speech act theory—representatives, directives, commissives, expressives, declarations—(Searle 1979) turned-out to be insufficient (Levinson 2017). In data annotation, alternative classifications like DAMSL (Allen, Core, 1997), SWBD-DAMSL (Jurafsky et al. 1997), AMI (2005), ISO 24617-2 (2012) and DART (Weisser 2019b) have therefore developed. Along with that the term changed from speech act to dialogue act, and the core notion was expanded significantly (Jurafsky 2004). Dialogue acts are, nowadays, usually defined in terms of dialogue functions (Jurafsky 2004) or communicative functions (ISO 24617-2 2012: 13) that an utterance performs. However, the existing schemes suffer drawbacks such as ambiguous distinction between the semantic and pragmatic meaning of utterances, lack of appropriate tags, poor informativeness of very general tags such as inform, and unsystematic annotation of metadiscourse acts (Verdonik 2022).

2 Data and methodology

Our approach is corpus-based. We have selected data for annotating dialogue acts in the total length of one hour. The data were selected mainly from the Slovene reference speech corpus (Verdonik et al. 2013) and represent diverse communicative settings. Detailed information on the data is provided in Table 1. The data were annotated by two independent annotators, both linguists. They have annotated the main dialogue act categories as identified in Verdonik (2022): information-providing acts, information-seeking acts, action acts, social acts and metadiscourse acts.

Both corpus annotators worked independently. The units of annotation were pre-annotated in order to avoid different interpretations of what is the basic unit of annotation. A minimal semantic and prosodic unit in the given context was annotated as the basic unit. The notion of context is crucial here, and it includes non-verbal, especially prosodic aspects, which we find most important for any interpretation of spoken language use.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

Speech event	Duration h:mm:ss
TV news	0:02:05
Lecture	0:09:37
Telephone sale	0:09:27
Family conversation at lunch	0:09:51
At home, friends planning a common vacation	0:10:03
Online counselling in the form of an interview	0:06:05
Entertaining TV talk show, three participants, humour	0:13:00
Total	**1:00:08**

Table 1: Data for annotation.

3 Metadiscourse dialogue acts annotation

Based on the annotated data we have identified typical metadiscourse dialogue acts which are very frequent, and the annotators had no problem recognizing them:

- signalling comprehension, e.g., backchannels like *ja* 'yeah', *mhm* or *aha* both 'mhm'
- signalling attention, e.g., Ja? 'Yes?'
- signalling production processes, e.g., *Aaa kako naj rečem.* 'Uhm, how should I say this.' or A: *Eee n() mislim*. 'Uhm n() I mean.' or *Kaj jaz vem*. 'I don't know.' or *Ne vem kako bi ti rekla.* 'I don't know how to say this.'
- closings and transmissions, e.g., *No to je to.* 'Well this is it.', or *Ja no prav.* 'Yes okay right.', or *Dobro.* 'Alright.', or *In to je bistvo ne*. 'And that's the point, y'know.'
- initiations, e.g., *Veš kaj*. 'You know what.', or *No v glavnem glej*. 'Well look.'
- referring backward, e.g., A: *Kot si rekel.* 'As you've said.'
- referring forward, e.g, *Glej jaz bom tako rekel.* 'Look I will say like this.', or *No pa še enega imam za vas*. 'Well, there is one more thing.'

The listed examples can be recognised fully for their dialogue act functions in the context of their use. Here, we do not have enough space to describe the context in detail. Furthermore, the functions of these acts in the context are typically more complex than, e.g. "signalling comprehension", since such an act can, at the same time, be signalling attention, interest, agreement, etc. Nevertheless, differences between the defined types are significant and all types can be recognised.

Along with the defined metadiscourse dialogue act types borderline cases were identified in our data. Those were:

- Expression of attitude or emotion towards the discourse content with (a) Phrases such as: *Huhu, super je*! 'Wow, it's awesome!', or *Kaj si ti nor, ej!* 'Is this crazy or what!' *Fenomenalno*! 'Phenomenal!' *Fajn!* 'Nice!', (b) Laugh, (c) Non-verbal sounds like *mmm*, expressing pleasure when the speaker eats something very tasty, (d) Swear words.
- Checking the collocutor's comprehension, e.g., *You understand what I mean?* or checking one`s own comprehension, e.g., *Like this?*
- Discussing discourse flow with subtypes (a) Committing the speaker's future discourse behaviour or dialogue act, e.g., *I will explain it to you later;* (b) Directing the collocutor`s discourse behaviour or dialogue act, e.g., *Comeon, be quiet!,* (c) Consulting the discourse flow, e.g., *Do we now have a serious moment?*, (d) Evaluating the discourse flow, e.g., *I said to myself that I will practice how to pronounce this. | But I'm not doing very well.*
- Repetitions can be a subtype of signalling comprehension type if their primary function is to express how the speaker comprehends the collocutor, or a subtype of expression of attitude or emotion type.
- Rhetorical questions can be a subtype of referring forward type.

The data we have used for the present research were limited in their size, and we should expect additional types and subtypes of metadiscourse dialogue acts when annotating more data.

Acknowledgments

The research presented in this paper was conducted within the research project titled Basic Research for the Development of Spoken Language

Resources and Speech Technologies for the Slovenian Language (J7-4642) funded by the Slovenian Research Agency.

References

James Allen and Mark Core. 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers*. https://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/

AMI. 2005. *Guidelines for Dialogue Act and Addressee Annotation Version 1.0.* http://groups.inf.ed.ac.uk/ami/corpus/Guidelines/dialogue_acts_manual_1.0.pdf

John L. Austin. 1975. *How to do things with words*. Oxford University Press, Oxford. Second edition.

ISO 24617-2. 2012. *ISO DIS 24617-2 Language resource management – Semantic annotation framework (SemAF), Part 2: Dialogue acts*. Geneva.

Dan Jurafsky. 2004. Pragmatics and computational linguistics. In Laurence R. Horn and Gregory Ward (Eds.), *The Handbook of Pragmatics*. Blackwell Publishing Ltd, pages 578–604.

Dan Jurafsky, Elizabeth Shriberg and Debra Biasca. 1997. *Switchboard SWBD-DAMSL shallow-discourse-function annotation*. *Coders manual, draft 13*. University of Colorado at Boulder & +SRI International. https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

Stephen C. Levinson. 2017. Speech acts. In Yan Huang (Ed.), *The Oxford Handbook of Pragmatics*. Oxford University Press, Oxford, pages 199–216.

John R. Searle. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge, UK.

Darinka Verdonik. 2022. Annotating dialogue acts in speech data: problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*.10.1075/ijcl.20165.ver.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek and Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation* 47(4): 1031–1048.

Martin Weisser. 2019b. The DART annotation scheme: form, applicability & application. *Studia Neophilologica*, 91(2):131–153. 10.1080/00393274.2019.1616218