SEMDIAL 2021
PotsDial

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue

Ellen Breitholtz, Kallirroi Georgila and David Schlangen (eds.)

Potsdam & The Internet, 20–22 September 2021

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.illc.uva.nl/semdial/

PotsDial Website

https://semdial2021.ling.uni-potsdam.de/

PotsDial Sponsors

PotsDial Endorsements

SIGdial

SIGSEM

Preface

PotsDial brings the SemDial Workshop on the Semantics and Pragmatics of Dialogue to the University of Potsdam for the second time; the first time was as the 10th meeting (brandial) in 2006. PotsDial, and the SemDial workshop as a whole, offers a unique cross section of dialogue research including experimental studies, corpus studies, and computational and formal models.

This year we received 28 full paper submissions, 15 of which were accepted after a peer-review process, during which each submission was reviewed by a panel of three experts. The poster session hosts 3 of the remaining submissions, together with 9 additional submissions that came in response to a call for late-breaking posters and demos. All accepted full papers and poster abstracts are included in this volume.

The PotsDial programme features three keynote presentations by Ryuichiro Higashinaka, Ruth Kempson and Stefan Kopp. We thank them for participating in SemDial and are honoured to have them at the workshop. Abstracts of their contributions are also included in this volume.

PotsDial has received support from the University of Potsdam, Department of Linguistics. We have also been given endorsements by the ACL Special Interest Group SIGdial and SIGSEM.

We are grateful to Jana Götze and Brielen Madureira for organising the compiling and editing of these proceedings. We would also like to extend our thanks to our Programme Committee members for their very detailed and helpful reviews.

Last but not least we would like to thank our local organisers from the Department of Linguistics at the University of Potsdam who have made PotsDial possible. Special mentions go to David Schlangen as the local chair, Jana Götze for coordinating the local organisation, our webmaster Brielen Madureira, Maike Paetzel-Prüsmann for managing our technical setup, Anne Beyer for managing local administration and social events, and our local helpers that made a hybrid event possible: Ines Mauer, Pelin Çelikkol, Philipp Sadler, Wencke Liermann, Luise Köhler, Karla Friedrichs and Sebastiano Gigliobianco.

Ellen Breitholtz, Kallirroi Georgila and David Schlangen

Potsdam

September 2021

Programme Committee

Ellen Breitholtz (chair)	University of Gothenburg
Kallirroi Georgila (chair)	University of Southern California
David Schlangen (chair)	University of Potsdam
Ron Artstein	University of Southern California
Claire Beyssade	CNRS / Univ. Paris 8
Eve V. Clark	Stanford University
Chris Cummins	University of Edinburgh
David DeVault	Anticipant Speech, Inc.
Barbara Di Eugenio	University of Illinois at Chicago
Simon Dobnik	University of Gothenburg
Arash Eshghi	Heriot-Watt University
Raquel Fernández	University of Amsterdam
Mary Ellen Foster	University of Glasgow
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	University of Gothenburg / King's College London
Pat Healey	Queen Mary University of London
Julian Hough	Queen Mary University of London
Christine Howes	University of Gothenburg
Amy Isard	University of Hamburg
Kristiina Jokinen	AIRC AIST Tokyo Waterfront
Ruth Kempson	King's College London
Staffan Larsson	University of Gothenburg
Alex Lascarides	University of Edinburgh
Pierre Lison	Norwegian Computing Center
Ramesh Manuvinakurike	Intel Labs
Vladislav Maraev	University of Gothenburg
Bill Noble	University of Gothenburg
Maike Paetzel-Prüsmann	University of Potsdam
Massimo Poesio	Queen Mary University of London
Matthew Purver	Queen Mary University of London
Hannes Rieser	Bielefeld University
Gabriel Skantze	KTH Royal Institute of Technology
Vidya Somashekarappa	University of Gothenburg
Matthew Stone	Rutgers University
David Traum	University of Southern California

Local Organising Committee

Anne Beyer	University of Potsdam
Jana Götze	University of Potsdam
Brielen Madureira	University of Potsdam
Maike Paetzel-Prüsmann	University of Potsdam
David Schlangen (chair)	University of Potsdam

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

Table of Contents

Poster Abstracts

Invited Talks

Human-aware conversational agents

Stefan Kopp
Faculty of Technology – Bielefeld University
skopp@techfak.uni-bielefeld.de

The development of conversational agents has made impressive advances in the last decades —ranging from statistical and neural approaches to dialogue management, to the understanding or generation of multimodal signals, to the large-scale deployment in voice assistants. This progress was made possible mainly by applying machine learning techniques to increasing amounts of training data. While these approaches yield dialogue, language or behavior models that cover larger domains, they do impose assumptions of universality and generality to structures and features of dialogue and even interlocutors. However, in many applications conversational agents need to be able to adapt rapidly and continuously to an individual user and an individual interaction, from as little data as a few verbal or nonverbal signals. I will argue that in order to achieve this goal we need to make conversational agents aware of how specific human users coordinate communication and dialogue, how they cognitively process socio-communicative behavior, and how they perceive conversational assistants. Along this line I will discuss work on conversational agents that are attentive and responsive to the interaction-relevant mental states (stance) of their human interlocutor and that can process (understand and generate) semantic and pragmatic functions of multimodal behavior. I will also present results from several studies on how different kinds conversational agents are perceived by different kinds of users.

How to Escape the Encodingism Stranglehold: Dynamic Syntax, Process and Interaction

Ruth Kempson with Ronnie Cann & Eleni Gregoromichelaki

This talk responds to the "encodingism" challenge posed by Bickhard (2009, in prep) that if models of cognition are ever to receive a naturalistic grounding, cognition, and language as a sub-discipline, will have to be seen as interactive context-dependent processes subject to ongoing change, with constructs of individual entities as emergent, thereby making cognition commensurate with a quantum theoretic perspective (Laudisa and Rovelli 2002). His attack in particular on linguistic theorising is that grammars defining fixed context-independent string-representation mappings, "encodingism" as he dubs it, cannot explain the flexibility of natural language, context-relativity, openness to change, and learnability through error detection. Hence such theories should be abandoned in favour of process-based theories.

This talk will tell the narrative of how Dynamic Syntax (DS) has increasingly managed to escape this encodingism stranglehold, as a case study of how this can be achieved. In its early days, (Kempson et al 2001), DS was planned to model logical form construction to substantiate pragmatic theorising, so a proper subpart of the encodingism methodology. Yet even initially, in this modelling of a process, DS demonstrated striking parallels with the criteria for explanatory models of language and cognition put forward by Bickhard, and the framework was successively confirmed by reanalysing phenomena previously taken to be syntactic/semantic puzzles in process–based terms, with new structural universals becoming expressible (specifically the preclusion of multiple unfixed nodes, the DS reframing of the concept of movement underpinning discontinuities in language). However, the incorporation of the basis adopted for the DS implementation in Eshghi et al. (2011, 2015) into the grammar formalism confirmed a much more radical break from the competence-performance dichotomy, a move signalled by the immediate explanation and prediction of the fluent exchange of roles in conversation, moreover allowing potential for correction, clarification etc., hence a tool for learning.

The tree-theoretic perspective of DS might seem nevertheless to retain the representationalism fiercely criticised by Bickhard. However, in turning to a composite DS-TTR framework (Purver et al 2010), DS was shown to be transformable into a multi-modal model integrating all facets of cognition in context including verbal processing as a subpart (Gregoromichelaki 2017), with even tree transitions characterisable as processes of differentiation (Bickhard, in prep) expressed by the utterance of words which offer affordances to trigger them (cf. Bruineberg and Rietveld 2019, Gregoromichelaki et al 2020). So the concept of representation is essentially secondary and emergent. And the particular structural universal, that no more than one unfixed node of a type can be introduced from a node at any point in the process, can be seen as reducible to wholly general process dynamics.

A further hurdle to overcome in the struggle to escape encodingism is then the problem of extensive systemic ambiguity, a hangover from the encodingism underpinnings of the initial DS goal. In this connection, recent work in combining DS and distributional semantic methods is providing an alternative semantic perspective able to directly reflect the indeterminacy of word meaning. And culling data from vast corpora collections is arguably one way to access cross-speaker variability and its incremental resolution in particular contexts. So in some sense, recent work in combining DS and distributional semantic methods is providing an alternative semantic perspective which is able to directly reflect the indeterminacy of word meaning, and its essentially social grounding (Purver et al, 2021, cf also Gregoromichelaki et al, 2019a,b for an affordance-based account).

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

It is with this shift into seeing both computational and lexical actions as affordances for underpinning interaction in dialogue that we get the final shift into granting the social nature of language. The criterion of success in language exchanges, on this view, is notably different from classical truth-directed assumptions. The purpose of language is to coordinate joint action: description, which involves truth, is one facet of that, not the most crucial. Instead, the normativity of action lies in criteria as to whether it has achieved its goal or whether further attempts are needed to achieve that success; and this can only be defined in some sense outside the individual themselves (Wittgenstein, 1953). The grammar is thus no longer a neutral intermediary between comprehension and production, all three to be defined independently. Nor is it some psychological competence of the individual independent of others. It is a model of what underpins participants' interactions in the social exchange, enabling fluent effects of feedback as well as drawing on the physical/social environment which provides these affordances. Universal aspects of language will then have to be constraints imposed by domain-general mechanisms guiding perception and action in the form of probabilistic generalisations over predictive, anticipatory processes, as all the rest emerges from interaction, arguably from birth (Raczaszek-Leonardi, et al 2018). And going along this route, if it can be achieved, IS the final definitive break with encodingism.

Postscript: this is of course in one sense a repetition of other papers of ours, but I hope the diachronic spin on how we came to where we are is illuminating. My closing message, given the background influence of the quantum theory process-based perspective throughout science and the force of Bickhard's critique, is that the development of appropriate grammar formalisms which directly reflect the nondeterminism, change and process at the heart of all languages is urgently required.

References

Bickhard, M. 2009. The interactivist model. *Synthese 166(3)*, 547-591.

Bickhard, M. in prep. *The Whole Person.*

Bruinenberg J & Rietveld E. 2019. What's in your head once you've figured out what your head's inside of, *Ecological Psychology* 3/13. 198-217.

Eshghi, A.; Purver, M.; and Hough, J.2011. *Dylan: Parser For Dynamic Syntax*. Technical Report Queen Mary University of London.

Eshghi, A., C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, London.

Gregoromichelaki, E. 2012. Review of J. Ginzburg (2012) The Interactive Stance. *Folia Linguistica*, Volume 47(1), 293–316.

Gregoromichelaki, E. 2017 Procedural Syntax and Interactions: ad hoc grammatical categorisation in DS-TTR. *Proceedings of FADLI 2017*. 22-26.

Gregoromichelaki, E. Howes, C. & Kempson R. 2019a. Actionism in syntax and semantics. In *CLASP Papers in Computational Linguistics* 12.

Gregoromichelaki, E. Howes, C., Eshghi, A., Kempson, R., Sadrzadeh, M., Hough, J., Purver, M., Wijnholds, G. 2019b. Normativity, Meaning Plasticity, and the Significance of Vector Space Semantics. In *Proceedings of SemDial 2019*, London.

Gregoromichelaki, E. et al 2020. Affordance competition in dialogue: the case of syntactic universals. *SEMDIAL Proceedings*2020.

Kempson, R. W. Meyer-Viol, D. Gabbay 2001. *Dynamic Syntax: The Flow of Language Understanding.* Blackwell.

Laudisa, F, and Rovelli, C. 2002. Relational Quantum Mechanics. In *The Stanford Encyclopedia of Philosophy* (Spring 2002 Edition), ed. Edward N. Zalta. http://plato.stanford.edu/archives/spr2002/entries/qm-relational/.

Purver, M., E. Gregoromichelaki, W. Meyer-Viol & R. Cann. 2010. Splitting the I's and Crossing the You's: Context, Speech Acts and Grammar. In *SEMDIAL Proceedings 2010 (PozDial)*, Poznań, Poland, June 2010.

Purver, M, Sadrzadeh, M., Wijnholds, G., Kempson R., Hough, J. 2021, Incremental composition in Distributional Semantics. *Journal of Logic, Language and Information* 30

Raczaszek-Leonardi, J. et al 2018. Language Development From an Ecological Perspective: Ecologically Valid Ways to Abstract Symbols, *Ecological Psychology*, 30:1, 39-73.

Wittgenstein, L. 1953. *Philosophical Investigations*. Blackwell Publishing.

Leveraging the wisdom of the crowd to realize a character-like chatbot

Ryuichiro Higashinaka
Nagoya University/NTT
higashinaka@i.nagoya-u.ac.jp

Having consistent personalities is important for chatbots if we want them to be believable. In this talk, I describe a technique called "role-play based question answering" in which multiple users play the role of certain characters and respond to questions posed by online users, making it possible to collect a large amount of character-associated data that can be used to build character-like chatbots with consistent personalities. I describe the technique in detail as well as a series of experiments performed to verify the usefulness of the collected data. Especially, I will talk about how much users can be motivated to voluntarily provide data, the quality of the collected data, how meta-data such as emotion can be collected and used for response generation, and how recent neural-based methods can be applied to the collected data. I will also describe a large-scale experiment, in which a chatbot based on "role-play based question answering" was used by the general public. I describe how the chatbot was received by the users and show typical errors made by the chatbot, which can give useful insights for improvements.

Full Papers

"By the way, do you like Spider Man?" — Towards A Social Planning Model for Rapport

Alafate Abulimiti[1], **Justine Cassell**[1], **Jonathan Ginzburg**[2]

[1] INRIA, Paris, France
alafate.abulimiti,justine.cassell@inria.fr
[2]CNRS, Université de Paris, Laboratoire Linguistique Formelle
yonatan.ginzburg@u-paris.fr

Abstract

Interaction takes place not only on the propositional level but also on the social level. In this paper, we consider *rapport* as an important social phenomenon in interaction. Motivated by data from the tutoring domain, we hypothesize that (i) off-task episodes are triggered by a low level of rapport and (ii) such episodes are means of raising the level of rapport. We sketch a planning model that allows off-task episodes to be triggered by (low) rapport level, which we apply to two simple examples.

1 Introduction

Pursuing multiple goals is a common phenomenon in interaction (Tracy and Coupland, 1990). E.g., people use politeness strategies (Brown et al., 1987) to pursue social goals, while simultaneously conveying propositional meanings to the recipient. In task-oriented contexts, social goal fulfillment becomes an unspoken purpose underlying the surface interaction. In the literature on the relative importance of multiple goals (Fetzer, 2003; Cassell and Bickmore, 2003), evidence is provided of the interplay of propositional, interpersonal and interactional meaning of the speech acts.

Consider (1), an excerpt from an actual algebra tutoring interaction between a teen tutor (TR): and a teen tutee (TT).

(1)

TR: It should be k equals something. [1]
TT: yeah that's right, I forgot about the k. [2]
TR: yeah so it wouldn't be that it be this. [3]
TR: so the reason why I'm so tired today is that I was up until like twelve thirty working on this script all right... [4]
TT: oh... [5]
: (... 5 minutes of off-task talk)
TR: all right so let's get back on topic to see what you're doing here [6]
TT: yeah. [7]

In this example, the tutor digresses during the tutoring session to discuss why he was looking tired. He does not apologize, and returns to the task by reintroducing the original task—related topic. This interlude where the tutor engages in self-disclosure (Derlega et al., 1993) might help the two interlocutors to establish a closer social relationship and thereby helps the tutor achieve his task goals (Sinha and Cassell, 2015b).

This interleaving of on-task and off-task moves is common and is not random (Coupland, 2014) in human-human interaction and reflects the phenomenon of individuals pursuing multiple goals while interacting (Tracy and Coupland, 1990; Fishbach and Ferguson, 2007).

Rapport (Spencer-Oatey, 2005; Tickle-Degnen and Rosenthal, 1990; Zhao et al., 2014) also serves as an important factor of social communication. This off-task mode (e.g., referring to shared experience, deep self-disclosure) also helps to enhance and maintain rapport (Zhao et al., 2014), which has also been shown to have a positive relationship with task performance (Sinha, 2016).

The application of the phenomenon of human interactivity to agent design is an important philosophy. But how to use these off-task moves at the right time and in the right place is a problem that has not been sufficiently investigated to the best of our knowledge. So this paper attempts to present a planning model that helps an agent to perform the

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

interleaving of task and social moves, in instances where we believe that the interlocutor may have picked up on low rapport accumulation over time.

The paper starts with a review of literature about rapport and its integration in dialogue systems. This leads us to formulate certain hypotheses about the relationship between off-task talk (OTT) and rapport. We then test these hypotheses using data from the peer tutoring domain. Building on this data, we sketch a model which provides rules for triggering off-task talk in natural conversation. We illustrate this model by analyzing two simple examples.

2 Related work

Rapport is described most fundamentally as a feeling of connection or harmony with another, and it has been shown to have important positive effects on communication and collaboration in a number of domains (Drolet and Morris, 2000; Bronstein et al., 2012; Bernieri and Rosenthal, 1991; Madaio et al., 2018). Some describe it as a calculus based on three essential components—mutual attentiveness, coordination and positivity (Tickle-Degnen and Rosenthal, 1990), while Spencer-Oatey (2005) describe it as based on behavior expectations, face sensibilities and interactional wants. Tickle-Degnen and Rosenthal (1990) shows that rapport is highly related to non-verbal moves. Moreover, verbal expression also influences rapport management (Zhao et al., 2014). Zhao et al. (2014) present a computational model for rapport management based on prior work and analysis of conversational data, and introduce different conversational strategies that enhance or maintain the rapport level. These conversational strategies include: self-disclosure, praise, violation of social norm, adherence to social norm, and hedging.

Cassell et al. (2007) presents a model where raising rapport may allow a conversational agent to model how people build friendships. Gratch et al. (2006); Huang et al. (2011) present a model that allows a virtual agent to produce verbal and non-verbal behavior appropriately to indicate rapport. Madaio et al. (2017) use such information to build a rapport estimator with temporal association rules. In (Pecune et al., 2018) the rapport scale is divided into seven levels from low to high (from 1 to 7); in the following discussion, we will follow this rapport scale to formulate our model.

Romero et al. (2017) create a social reasoner which takes a task reasoner's intention and rapport level as input and outputs a phrase that both attempts to achieve the task goals, and is phrased as one of the aforementioned conversational strategies, by using a spreading activation network (Maes, 1989). However, this model relies on a unidirectional relationship between the task and social reasoner—the social reasoner cannot affect the task reasoning. Concretely, such a model could generate an example similar to (1), such as "I'm doing a bad job because I'm so tired from staying up all night…" —a negative self-disclosure to mitigate the recipient's negative face in order to improve *rapport* and to continue the task plan without shifting to another topic and a resulting side sequence (one that ends normally with "let's go back to this question now" in order to continue the pending question). However, it could not generate the interleaving of utterances that only deal with social matters, such as those shown in (1).

A high rapport level contributes to effective task performance (Sinha and Cassell, 2015b), but when the current sub-task is performed with low rapport and this phenomenon lasts a certain period of time, it may not be viable to continue this sub-task. In contrast to the small talk function of (Cassell and Bickmore, 2003) (i.e., small talk helps the user modeling process with gaining the user's trust), in (Cassell and Bickmore, 2003) switching the topic is used to initiate an off-task mode characterized by a pleasant environment and positive valence in phatic communication like small talk (Coupland, 2003), jokes, and gossip, so that rapport can calibrate to a normal level to help the recipient comfortably continue the pending task. Zhao et al. (2014) introduce diverse conversational strategies which include referring to shared experience and self-disclosure. These strategies are the means of switching to off-task mode with off-task utterances that last longer than a clause. However, although off-task talk appears frequently, it is excluded from consideration given the architecture of the SARA system (Pecune et al., 2018). An as yet unresolved issue is whether off-task talk happens after low rapport occurs. Hence, we make the hypothesis:

H1: Off-task talk occurs after low rapport is observed within a certain window.

Off task moves are extensively used in real-life interaction (Jaworski, 2014; Holmes, 2014; Endrass et al., 2011). Kopp et al. (2005)'s museum

guide agent can interact with the visitors using small talk, but these are controlled by rules triggered by specific interactional contents. By contrast, the REA model of Cassell and Bickmore (2003) uses an activation network to construct the discourse planner, which can choose new topics with entire off-task utterances, and these new topics are designed to gather the user's information in order to shape the user model (i.e., user's goal, plan and knowledge) to identify the users' cooperation level and establish the close social relationship — trust. This alternative planning serves the ultimate task goal. However, REA is not based on an investigation of the distribution of off-task episodes during the interaction (i.e., where people often use small talk). Sinha and Cassell (2015a); Sinha (2016) explain the relationship between the time series of rapport and task performance in peer tutoring setting. They conduct an in-depth analysis relating different conversational strategies (i.e., self-discourse, praise and reference to shared experience), different gender pair and learning gains across all periods (social and task periods) of interaction. The relationship between the OTT and the final learning gain is also a question worth exploring. Different conversational strategies have different effects on learning gains (Sinha et al., 2015). If OTT is widely present in real conversations, there is a natural question whether OTT can contribute to the completion of the task? This motivates a second hypothesis:

H2: Off-task talk helps to improve task performance (e.g., in the peer tutoring domain—learning gain).

3 Data Analysis

Our data consists of video recordings of reciprocal peer tutoring interaction taken from 14 dyads in which the age of the participants was on average 13, half were male and half were female. All dyads consisted of identically gendered participants. Each dyad interacted over two sessions. Before and after each session all participants were tested in order to measure their learning gains. A total of 24 interaction transcripts were manually annotated with tags relating to the off-task talk which we explain in more detail shortly. This yielded a total of 22709 clauses. Their interactions were split into 30 second slices whose rapport had been estimated by Amazon Mechanical Turk annotators.

3.1 Data Annotation

We classify the data into 4 types with respect to their content: 1) on-task talk, 2) off-task talk—side sequences involving topics that are distinct from the task, 3) clauses that reject or ignore off-task talk (e.g. P2's utterance in the following, which rejects P1's off-task talk invitation: P1: I watched Spider-Man yesterday. P2: let's add 5 to both sides.), 4) meta-task talk—clarification sequences concerning the task itself.

The session is divided into 4 periods, first period is the first social period usually about the self-introduction, second period is the task period about the algebra tutoring, one becomes the tutor, another becomes the tutee, the third period is a five minute long break after which the participants switch roles to start the second task period. We focus on the off-task talk in the task period and do not annotate the social period clauses. The entire data annotation work was carried out by two annotators, and all disagreements were discussed and resolved.

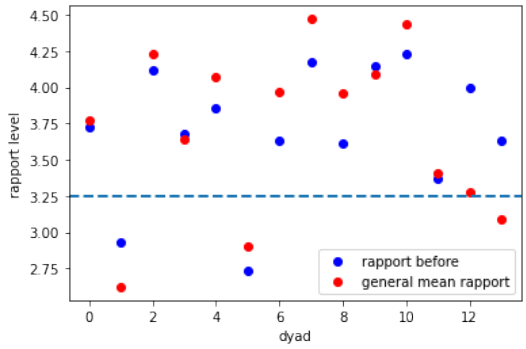
3.2 Results

The Pearson Correlation between the time spent per session off-task talk and the dyad average learning gain is 0.3558, $(p = 0.08 > 0.05)$. This suggests that there is a moderate positive correlation between the duration of the off-task talk and the learning gain. As off-task talk increases, it may help a dyad's task completion. However, due to the small size of our data set, our results are not statistically significant. Nonetheless, this result does raise another reasonable question that needs to be investigated in future work: in a temporally-bounded task-oriented conversation, is too much off-task talk detrimental to task completion?

We mark the slices that appear to be off-task talk (OTT) and use the ten slices that do not appear to be OTT before these slices as reference objects. Means were calculated for each dyad and these were compared with the general rapport mean to perform a paired t-test. The result is: $t = -1.4587, p = 0.1470 > 0.05$. This result shows that rapport is lower compared to general rapport levels before the onset of the OTT, but the result is not significant. So we examined the data for each dyad as shown in the figure1.

We observe that 9 out of 14 dyads satisfy our **H1** hypothesis, that is, indicating that the rapport level is lower than the general level when OTT occurs. Moreover, when we filter out dyads with

Figure 1: Comparison by dyad between general rapports and rapports before OTT



lower levels of general rapport (less than or equal to 3.25), 9 out of 10 satisfy **H1**. This provides some justification for the truth of hypothesis **H1**, which drives our model of *social repair*, a concept we now turn to.

4 Tutoring: the interplay between task goals and social goals

4.1 A simple example

From the earliest days of AI to the present, Intelligent Tutoring Systems have remained as an important area, with its purpose to enable learners to learn better and more efficiently by using existent computer technologies to simulate what seems to be the most effective way to tutor, namely private tutors. Ritter et al. (1998) helps to improve a student's algebra performance based on the student's prior knowledge and experience. The Rapport-Aware Peer Tutor Assistant[1] is an anthropomorphic embodied social robot that maximizes student performance by improving rapport with students. This assistant incorporates a rapport estimator (Madaio et al., 2017) and uses a *social reasoner* to deduce the next conversational strategy and based on the communicative intent its task reasoner assigns to the recipient.

The assistant uses a top-down structure where the social reasoner accepts the task reasoner's intentions, but has no alternative means of influencing the manner in which the task reasoning proceeds, often causing the assistant to maintain rapport in a circumlocutory way, while fulfilling the tyrannical "commands" of the task reasoner. This approach tends to cause low rapport to build up slowly, causing the assistant's rapport level to slip in the opposite direction at some point. This is exemplified in (2):

(2)

AGENT: No that's not it. [1]

TEENAGER: I don't know. [2]

AGENT: So why did you do it that way sometimes I have a tough time explaining my thinking. [3]

TEENAGER: I wasn't right, so like. . . I don't know how to do it. [4]

AGENT: You want to get the x term by itself I hope I'm explaining this okay. [5]

TEENAGER: I. . . still don't know. [6]

This is a small extract from a lesson where the agent tries to help a student to complete a given algebra problem. The student is not able to complete the problem successfully perhaps because the difficulty of the question was beyond his current proficiency, though the agent uses different conversational strategies to perform face work (e.g., negative self-disclosure) in order to complete the task reasoner's "assignment" for this part and to maintain or enhance rapport.

Two shifts could take place at this point. The first one (*knowledge tracing*) is to change the current question to an easier one. This serves two purposes, one is not to remain blocked with the current question, and by doing the easier one, to accumulate experience which will enable solving the pending question. The second shift is to alleviate the teenager's negative face, so that the rapport may rise. This second shift involves changing the current direction of the dialogue by initiating unrelated topics (e.g., weather, hobbies, etc.), where small talk is proven to pull the relationship closer (Coupland, 2014). In this paper, we focus on the social reasoning part with the second type of shift rather than on knowledge tracing.

When the current discussion is not appropriate to be continued given a low level of rapport, it is imperative to change the current topic in order to *raise* the current rapport to an acceptable level. The assistant can digress (e.g., switch to talking about hobbies, sports and favorite movie stars) and wait until the rapport level returns to a baseline, then resume the questions that were just pending.

We dub these kinds of behaviors as *social repair*.

[1] http://articulab.hcii.cs.cmu.edu/projects/rapt/

4.2 Social repair

The purpose of communicative repair (Jefferson et al., 1977; Purver, 2004; Ginzburg, 2012) is to enable one interlocutor to fully understand their interlocutor's initially incompletely comprehended utterance and the associated intentions. In such cases, the problematic utterance is set aside until reference or mishearing etc are resolved. By analogy, what we call *social repair* concerns the need to restore social relations such as rapport, power, trust, to "appropriate" levels.

Social repair can happen in a single utterance that also attempts to achieve task goals, by using the different conversational strategies supported by rapport theory (Zhao et al., 2014; Spencer-Oatey, 2005; Tickle-Degnen and Rosenthal, 1990). However, people also perform off-task moves constantly during an interaction.

Zhao et al. (2014) introduce several conversational strategies as rapport enhancement and rapport maintenance, in order to keep the rapport level high enough to allow the task to proceed smoothly. They are initiated when the rapport falls below a certain level, or in the beginning of the dialogue to raise it to a certain level. Some of these strategies can be delivered within a single clause (e.g., praise, adherence to social norms, etc.) But here we focus on strategies that can involve more than one clause without mentioning the task (e.g., referring to shared experience, deep self-disclosure).

With the existence of a social repair function, the question arises when to trigger the social repair process. When we return to consider example (2), we observe that the student repeats his feeling that he doesn't know what to do next, and this statement also appears after the assistant gives him a hint how to solve the problem ([5]). The low rapport in the current interaction which accumulates over time as he moves from a simple "I don't know" to a self-defeating [4] to a direct statement that he still does not know.

4.3 Rapport Accumulation

As we already mentioned above, following (Madaio et al., 2017)'s rapport estimator, rapport is determined as a time series function where the value ranges from 1 to 7 (1 for lowest rapport level and 7 for the highest). We assume rapport persists and accumulates in one's *Cognitive State*, tempered by some notion of *decay*. Rapport is, then, given by a function over time $rp(t)$ modulated by γ, a postulated decay rate.[2] So we obtain the following for *rapport accumulation*:

$$r_a(t) = rp(t_0) + \gamma * rp(t_1) \ldots + \gamma^{(\Delta t-1)} * rp(t_0 - \Delta t)) = \sum_{t=t_0-\Delta t}^{t=t_0} \gamma^{t-t_0} * rp(t) \quad (1)$$

5 Social Planning

In this section, we introduce a computational model that adjusts the current plan to maintain a reciprocal goal relationship based on the different low rapport accumulation to the task encountered, in order to maintain an efficient pattern of pursuing goals in tandem.

5.1 Cognitive States

We use KoS (Larsson, 2002; Ginzburg, 2012; Ginzburg et al., 2020) as a framework for representing the cognitive states of dialogue participants; the individual's perspective on the public aspects of the interaction are represented in the *dialogue gameboard* (DGB) whereas the private projection and interpretation of current events are presented in *private*, on which more below. Rapport estimation is mostly based on information originating in the DGB e.g., gaze, linguistic delivery, smiling etc, though we do not offer an account of how this gets computed here.

(3)

a. Total Cognitive State $=_{def}$

$$\begin{bmatrix} \text{dialoguegameboard} : \text{DGBtype} \\ \text{private} : \text{Private} \end{bmatrix}$$

b. DGBType $=_{def}$

spkr: Ind	turn
addr: Ind	owner-
utt-time: Time	ship
c-utt: addressing(spkr,addr,utt-time)	
Facts: Set(Proposition)	shared assumptions
VisSit: [InAttention : Ind]	visual field
Pending: list (locutionary Proposition)	ungrounded utts
Moves: list (illocutionaryProposition)	grounded utts
QUD: poset (Question)	qs under disc
Mood: Appraisal	face

[2] Thanks for an anonymous reviewer for SemDial for suggestions concerning the decay rate.

5.2 Plan Modification for Social Repair

A plan p can be represented by a sequence of episodes: $p = stack(\{ep_1, \ldots, ep_n\}) = stack(EP)$.

We hypothesize that low rapport accumulation is a trigger of task plan change. As the plan represents the dynamic representation of pursuing the goal(s), for simplicity we restrict the goal set here to two goals with same hierarchy: a task goal and a social goal. The task goal represents the task completion state, whereas the social goal is constituted by the end state attained by a social actor (e.g., maintaining high rapport).

We introduce Ω as a set of weights relating to the goals: $\Omega = \{\omega_1, \ldots, \omega_k\}$, Depending on the parameters, different goals have different levels of importance, which affects the judgment made by the agent based on the own threshold values. We simplify here to assume a simple goal set: $G = \{g_t, g_s\}$ (i.e., one task goal and one social goal) and the $\Omega = \{\omega_t, \omega_s\}$. Furthermore, we define r_{th} as the accumulated rapport threshold. From this, we obtain $\omega_s * r_{th}$ as the weighted rapport accumulation threshold. We also assume the existence of a *repair set* $REP = \{ep^r_1, \ldots, ep^r_n\}$ which is composed of several *repair actions*, which include actions construed as OTT.

Putting all this together, we assume the PRIVATE part of Total *Cognitive State* is typed as in (4).

(4) PRIVATE =

$$\begin{bmatrix} \text{Agenda:OpenQueue(Action)} \\ \text{Plan: OpenStack(PlanConstruct)} \\ \text{BEL:} \begin{bmatrix} \text{Rapport} = \begin{bmatrix} \text{Cur} = rp(t) \\ \text{Accu} = r_a(t) \\ \text{Trd} = r_{th} \end{bmatrix} \end{bmatrix} \\ \text{Goals:} \begin{bmatrix} \text{GoalsSet: List(Prop)} \\ \text{GoalsIpt} = \Omega\text{: List(Float)} \end{bmatrix} \\ \text{RepairSet: Set (Plan)} \end{bmatrix}$$

Rapport, as we have said above, is given by a function over time: $rp(t)$. r_{th} is the rapport threshold that triggers the task plan changes if $r^j_i < \omega_s * r_{th}$. r_{th} implies the agent's social sensibility during the interaction. If the threshold condition is reached, we infer that the next episode in the plan ep_{j+1} can be *deferred*. If one episode is *deferred*, an element of the repair set ep^r_i can be *inserted* before ep_{j+1}.[3]

[3]There are also other operations we could envisage such as *replace*, *delete* or *swap*, but we will restrict ourselves to a simple account in the current paper.

Calculation of all the expected rapport raising actions for the episodes in the *repair set* and the subset $EP_{j+1,n}$ takes place according to the current *Total Cognitive State* s_i then a choice is made for the maximal one. This calculation process occurs after the precondition is triggered (i.e., $r^j_i < \omega_s * r_{th}$). The formal update rule (**Off Topic Triggering**) is given in (5):

(5) **Off Topic Triggering**

$$\begin{bmatrix} \text{Pre:} \left[\text{BEL.Rapport.Accu} < \text{BEL.Rapport.trd} \right] \\ \text{Eff:} \left[insert(ep^r_i, Plan.cur) \right] \end{bmatrix}$$

Where ep^r_i is an episode selected from EP^r assumed to be maximal in rapport raising. We do not explicate this selection process in the current paper.

6 Examples

In this section, we apply our model of planning to example (2) and to a variant thereof.

In example (2), we assume that:

- The decay rate is 0.8.
- This conversation takes place in the j th episode.
- The rapport accumulation threshold is $r_{th} = 5$.
- We assume that at the completion of utterance [2], the rapport score (in ascending order from 1 to 7) is 4, and after [6], it is 2.
- The assistant's $\Omega = (\omega_t, \omega_s) = (0.5, 0.5)$. This means that equal importance is assigned to the task and social goals.
- Δt = 1. This means the turn span is one. We calculate the accumulation from [2] to [6].

From (1), r[2]: $r^j_2 = \sum_{t=1}^{t=2} \gamma^{2-t} rp(t) > \omega_s * r^j_{th}$ and r[6]: $r^j_6 = \sum_{t=5}^{t=6} \gamma^{6-t} rp(t) < \omega_s * r^j_{th}$

When [6] triggers the accumulation threshold condition, the assistant defers the current episode (i.e. the agent could say:"Let's take a break."); the assistant needs to select an episode that maximizes the rapport raising among the social repair set EP^r, inserting an off task topic (i.e., agent could say:"By the way, do you like Spider-Man?").

We turn to another example. In this case we assume that the person helping the child is a strict parent. The parent's goal is to get the child to complete the given task quickly regardless of the rapport level.

Since the social relationship is established, we can naturally assume that the parent pays less attention to social needs. Hence, in the model we propose that the social repair function will be triggered less frequently in the process of interaction, which is followed by a continuation of the task plan. We consider the following constructed example:

(6)

PARENT: You should do this . . . [1]
TEENAGER: I don't know. [2]
PARENT: What do you mean, you should understand, you should do exactly like this . . . [3]
TEENAGER: Yes, but I still don't know how to do it. [4]
PARENT: You should do this and this! [5]
TEENAGER: OK. . . [6]

We assume that:

- The decay rate is 0.8.
- This conversation takes place in the j th episode.
- The accumulation threshold is $r_{th} = 5$.
- We assume that at the completion of [2], the rapport score (in ascending order from 1 to 7) is 3, and after [3], it is 1.
- The parent's $\Omega = (\omega_t, \omega_s) = (0.8, 0.2)$. This means that the parent's focus is more on task than on the social needs.
- Δt = 1. This means the turn span is one.

From (1), r[2]: $r_2^j = \sum_{t=1}^{t=2} \gamma^{2-t} rp(t) > \omega_s * r_{th}^j$ and r[5]: $r_5^j = \sum_{t=4}^{t=5} \gamma^{5-t} rp(t) > \omega_s * r_{th}^j$

After [3], since the rapport accumulation has not been lower than the parent's threshold, the social repair mechanism is less likely to be triggered than in the first example. Hence, the parent pushes the child to continue the task with [5].

7 Conclusions and Future Work

Interaction involves not only exchange resolving around propositional meaning, but also around different social phenomena. In task-oriented natural dialogue, we find that the participants not only initiate the dialogue as required by the task, but also intersperse this with different off-task moves which contribute to the achievement of social goals. In this paper, we sketch a model that places rapport as a central social phenomenon and segments tasks into episodes using the accumulated rapport as a mechanism for triggering off-task moves. We also apply this model to two examples.

We hope to spell out this model and apply it to more complex examples in future work.

Required refinements include:

- Certain conversational strategies in the rapport model proposed by Zhao et al. (2014) can improve rapport when social repair is required. How this and more generally how rapport estimation can be integrated into formal model of dialogue such as KoS remains to be worked out.
- We focus on the example of tutoring in this paper, but off-task mode exists in a wide variety of task-oriented conversational types in different forms. Furthermore, we plan to investigate how this mode differs across distinct conversational types.
- In our brief presentation, for simplicity, we assume that the individual's rapport accumulation threshold (r_{th}) is fixed. In reality, however, r_{th} changes dynamically in accordance with the dialogue context and conversational type. We would like to integrate the conversational type (Wong and Ginzburg, 2018) into the assessment of the rapport accumulation threshold.
- In our subsequent data analysis, we performed a t-test in the interval where OTT occurred and in the interval before it occurred, The result did not show a significant effect ($t = -0.5718, p = 0.5683$). This is probably because we do not have sufficient data. However, we hope to scale this up in the hope of proving hour basic hypothesis.

Acknowledgments

Many thanks to Aliah Zewail, Abdellah Fourtassi, and the other members of the ArticuLabo at INRIA Paris for their precious assistance. This work was supported in part by the the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). This work is also supported by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris - ANR-18-IDEX-0001. We also acknowledge a senior fellowship from the Institut Universitaire de France to JG.

References

Frank J Bernieri and Robert Rosenthal. 1991. Interpersonal coordination: Behavior matching and interactional synchrony.

Ilan Bronstein, Noa Nelson, Zohar Livnat, and Rachel Ben-Ari. 2012. Rapport in negotiation: The contribution of the verbal channel. *Journal of Conflict Resolution*, 56(6):1089–1115.

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

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User modeling and user-adapted interaction*, 13(1):89–132.

Justine Cassell, Alastair Gill, and Paul Tepper. 2007. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*, pages 41–50.

Justine Coupland. 2003. Small talk: Social functions. *Research on language and social interaction*, 36(1):1–6.

Justine Coupland. 2014. *Small talk*. Routledge.

Valerian J Derlega, Sandra Metts, Sandra Petronio, and Stephen T Margulis. 1993. *Self-disclosure*. Sage Publications, Inc.

Aimee L Drolet and Michael W Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50.

Birgit Endrass, Matthias Rehm, and Elisabeth André. 2011. Planning small talk behavior with cultural influences for multiagent systems. *Computer Speech & Language*, 25(2):158–174.

Anita Fetzer. 2003. 'no thanks': a socio-semiotic approach. *Linguistik online*, 14(2).

Ayelet Fishbach and Melissa J Ferguson. 2007. The goal construct in social psychology.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa*, 5(1).

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. 2006. Virtual rapport. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer.

Janet Holmes. 2014. Doing collegiality and keeping control at work: small talk in government departments 1. In *Small talk*, pages 32–61. Routledge.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual rapport 2.0. In *International workshop on intelligent virtual agents*, pages 68–79. Springer.

Adam Jaworski. 2014. Silence and small talk. In *Small talk*, pages 110–132. Routledge.

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

Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents*, pages 329–343. Springer.

Staffan Larsson. 2002. *Issue-based dialogue management*. Citeseer.

Michael Madaio, Rae Lasko, Amy Ogan, and Justine Cassell. 2017. Using temporal association rule mining to predict dyadic rapport in peer tutoring. *International Educational Data Mining Society*.

Michael Madaio, Kun Peng, Amy Ogan, and Justine Cassell. 2018. A climate of support: a process-oriented analysis of the impact of rapport on peer tutoring. International Society of the Learning Sciences, Inc.[ISLS].

Pattie Maes. 1989. How to do the right thing. *Connection science*, 1(3):291–323.

Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field trial analysis of socially aware robot assistant. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1241–1249.

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

Steven Ritter, John Anderson, Michael Cytrynowicz, and Olga Medvedeva. 1998. Authoring content in the pat algebra tutor. *Journal of Interactive Media in Education*, 1998(2).

Oscar J. Romero, Ran Zhao, and Justine Cassell. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3807–3813. ijcai.org.

Tanmay Sinha. 2016. Cognitive correlates of rapport dynamics in longitudinal peer tutoring.

Tanmay Sinha and Justine Cassell. 2015a. Fine-grained analyses of interpersonal processes and their effect on learning. In *International Conference on Artificial Intelligence in Education*, pages 781–785. Springer.

Tanmay Sinha and Justine Cassell. 2015b. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 13–20.

Tanmay Sinha, Ran Zhao, and Justine Cassell. 2015. Exploring socio-cognitive effects of conversational strategy congruence in peer tutoring. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 5–12.

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

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

Karen Tracy and Nikolas Coupland. 1990. Multiple goals in discourse: An overview of issues. *Journal of Language and Social Psychology*, 9(1-2):1–13.

Kwong-Cheong Wong and Jonathan Ginzburg. 2018. Conversational types: a topological perspective. pages 156–166.

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International conference on intelligent virtual agents*, pages 514–527. Springer.

Speech planning interferes with language comprehension: Evidence from semantic illusions in question-response sequences

Mathias Barthel
Humboldt University Berlin
mathias.barthel@hu-berlin.de

Abstract

In conversation, speakers need to plan and comprehend language in parallel in order to meet the tight timing constraints of turn taking. Given that language comprehension and speech production planning both require cognitive resources and engage overlapping neural circuits, these two tasks may interfere with one another in dialogue situations. Interference effects have been reported on a number of linguistic processing levels, including lexico-semantics. This paper reports a study on semantic processing efficiency during language comprehension in overlap with speech planning, where participants responded verbally to questions containing semantic illusions. Participants rejected a smaller proportion of the illusions when planning their response in overlap with the illusory word than when planning their response after the end of the question. The obtained results indicate that speech planning interferes with language comprehension in dialogue situations, leading to reduced semantic processing of the incoming turn. Potential explanatory processing accounts are discussed.

1 Introduction

When speakers are in conversation, they notoriously take turns at talking, switching their roles of speaker and listener within short intervals of time (Sacks et al., 1974). For the greatest part of the conversation, only one of the speakers talks while the other stays silent, and stretches of mutual silence and overlapping talk are mostly very brief (Heldner and Edlund, 2010; Stivers et al., 2009). Among the greatest driving forces for fast responses, next to the possibility of completely missing out on the turn, is the semiotics of turn-timing, whereby long gaps are interpreted to be meaningful, signalling, for example, reduced willingness to comply with a request (Kendrick and Torreira, 2014; Roberts and Francis, 2013; Roberts et al., 2011). In order to achieve this remarkably precise orchestration of speaking turns, the next speaker needs to start planning his utterance while the current speaker is still delivering her turn. Planning-in-overlap has indeed been found to be the default strategy of speech planning in conversational situations, where speakers start to plan their response as soon as they can anticipate the message of the incoming turn (Barthel and Levinson, 2020; Barthel et al., 2016, 2017; Bögels et al., 2015; Bögels, 2020; Corps et al., 2018). While planning in overlap makes seamless responses possible (Barthel, 2020; Levinson and Torreira, 2015), it comes with the cost of increased processing load during speech planning as compared to planning during the silence between turns (Barthel and Sauppe, 2019). That means that, in dialogue, processing load in next speakers usually peaks just before turn transitions, which seems reasonable for two related reasons. Firstly, turn-transitions are dual-task situations, with response planning being executed during ongoing language comprehension. And secondly, the related nature of the two tasks can create interference between them, making them less efficient as they become computationally harder. Such interference effects can occur on any possible level of language processing, from lexical selection over word form retrieval and phonetic encoding down to motor preparation (Abdel Rahman and Melinger, 2019; Barthel and Levinson, 2020; Boiteau et al., 2014; Bürki et al., 2020; Fargier and Laganaro, 2016; He et al., 2021; Jescheniak et al., 2014; Konopka, 2012; La Heij et al., 1990; Meyer, 1996; Schriefers et al., 1990, inter alia). These cross-talk effects have been assumed to be rooted in shared representations for production and comprehension and/or partly overlapping neural architecture underpinning these tasks (Buchsbaum et al., 2001; Hagoort and Indefrey, 2014; Indefrey and Levelt, 2004; Kempen

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

et al., 2012; MacKay, 1987; Menenti et al., 2011; Silbert et al., 2014). While the interference effects of language comprehension on speech production have received increasing attention during the last decades, the effects of speech planning on parallel comprehension remain comparatively understudied (Daliri and Max, 2016; Fargier and Laganaro, 2019; Levelt et al., 1991; Roelofs et al., 2007).

The present study focuses on semantic processing of the incoming speech in experimentally elicited question-answer sequences, exploiting the well-known effect of semantic illusions – the acceptance by a comprehender of a fallacious question or statement containing a word that makes the question or statement wrong but is semantically related to the correct word that was to be expected. A classic example used in the seminal study by Erickson and Mattson (1981) is the question *"How many animals of each kind did Moses take on the Ark?"*. In their study, Erickson and Mattson found that participants frequently failed to spot the illusion in the questions, even though they actually knew the correct facts, e.g., that it was not Moses but Noah who prepared himself for the great flood. The origin of the illusion is commonly explained by frugal, superficial semantic analysis and partial semantic feature-matching between the critical word and its sentence co-text – the greater the intersection of the semantics of the critical word with that of the correct word, the greater the chance that the illusion passes undetected (Song and Schwarz, 2008; Speckmann and Unkelbach, 2020; Van Oostendorp and De Mul, 1990).[1]

The conversational counterpart to the carefully constructed illusions used in experimental studies are word substitution errors, where an intended word accidentally gets replaced by an unintended word during speech planning. Such word substitutions are a very common kind of speech error (Meringer, 1908). Of the Fromkin Speech Error Corpus[2], a collection of 8673 spontaneous speech errors, 1083 errors (12.5%) are word substitutions, many of which replace the target word with a semantically related word that is generally of the same part of speech. In one example produced by Vicki Fromkin herself (and recorded by Robert Rodman), she produced "*Jack was going to build a YACHT on the 38th day*", instead of the intended and semantically related '*an ark*', when talking about a long period of rain in Oxford. As speech error corpora have traditionally mainly been used to draw inferences on the processes of speech production, the perception of these errors, including common detection rates, received far less attention (Bond, 1999).

In conversational situations, two different general language processing strategies are conceivable, predicting different effects on the detection rate of semantic illusions. As discussed above, in order to secure a timely response, speech planning needs to proceed swiftly already during the incoming turn. Therefore, delays due to capacity limits could lead to undesirably long turn-transition times that might communicate unintended meanings. Because planning in overlap is cognitively more demanding than planning the next turn in silence, next speakers might need to prioritize planning speed over comprehension accuracy after the point when response planning begins in order to secure smooth turn-transitions. We will call this the *turn-timing-prioritized hypothesis*. Equally theoretically conceivable is an alternative strategy that focuses on language comprehension in processing-heavy situations in turn taking. This strategy appears reasonable in view of the differences in the temporal dynamics of speech input and output processing. While the rate of speech input is defined by the speech rate of the current speaker, the rate of progress in speech output planning is under the control of the next speaker. That means that while the input, if not processed upon reception, is soon gone from perceptual memory, delays in response planning can be handled more flexibly by the next speaker, which is a potential incentive to prioritize the processing of incoming speech over speech planning in conversational situations, so that sufficient processing resources are available for comprehension. We will call this the *comprehension-prioritized hypothesis*.

The present study tests these competing hypotheses, using a quiz task with questions containing semantic illusions. While actual conversational situations are arguably way more complex than an experimental quiz situation, responding to questions is a very common action in social encounters. Using pre-recorded questions thus strikes a balance between exerting sufficient experimental

[1] The illusion effect is also boosted by additional phonological overlap between the target word and its replacement (Shafto and MacKay, 2000). Yet, the illusions presented in the present study were only semantically related to the target word, not phonologically.

[2] accessible at https://www.mpi.nl/dbmpi/sedb/sperco_form4.pl

control and tapping into the target speech production and comprehension processes. The presented questions differ in the point in time when response planning can begin. In one version of the question, response planning can begin already in overlap with the question, while in the other version, the answer to the question can only be known at the very end of the question. E.g., *Welche Tiere helfen dem Weihnachtsmann beim Verteilen der Neujahrsgeschenke?* ("What animals help Santa Claus to distribute New Year's presents?"; early planning) vs. *Der Weihnachtsmann verteilt die Neujahrsgeschenke mit der Hilfe von welchen Tieren?* ("Santa Claus distributes New Year's presents with the help of what animals?"; late planning). The two hypotheses outlined above make different predictions about the detection of semantic illusions in these two versions of the question. If comprehension is prioritized in dialogue situations, detection rates should not depend on whether the response is already being planned in overlap with the critical word or not. If fast turn-timing is prioritized on the other hand, detection rates should be lower in the early planning condition than in the late planning condition.

The point in time when response planning can start is confounded with the questions' sentence structure as well as with the position of the critical word within the question. To test the influence of these confounding factors, a control experiment was tested that did not require participants to actually answer the question. If a difference in detection rates between the two conditions in the main experiment is due to planning vs. not planning in overlap, this difference should not show in the control experiment, which does not involve response planning. If, on the other hand, differences in detection rate are due to differences in the form of the question itself, they should replicate in the control experiment.

2 Method

2.1 Participants

For the Main Experiment, 24 participants (age between 18 and 40 years) were recruited via Prolific and were paid to take part in the experiment online using their own computers. Another set of 24 participants was recruited for the Control Experiment.

2.2 Materials and Design

60 questions were composed, 30 of which were critical questions containing a semantic illusion. Of each question, two versions were composed, manipulating the point in time when the answer to the question can be known so that response planning can begin (planning: early / late; see example in (1)).

(1) (a) Early question: *Von welchem Tier wurde Rotkäppchen gefressen, | als sie ihre **Tante** besuchte?* ("What animal ate Little Red Riding Hood | when she visited her **aunt**?")

(b) Late question: *Als Rotkäppchen ihre **Tante** besuchte, wurde sie von welchem Tier gefressen |?* (When Little Red Riding Hood visited her **aunt**, what animal ate her |?)

In the early question in (1a), subjects can begin to plan their verbal response to the question already in the middle of the question (marked by the | symbol in (1)), whereas planning the response in the late question only becomes possible at the end of the question. The critical word containing the semantic illusion (printed in bold in (1)) is therefore located in a later part of the question in the early version, where response planning can be expected to be already ongoing, and in the earlier part of the question in the late version, where response planning cannot have started, yet.

Additionally, 30 filler questions were created. Filler questions were also composed in two versions, allowing for either early or late planning, but they did not contain any illusion (e.g., early planning: *Welcher Mann, der das Unternehmen Apple gründete, | war ein fortschrittlicher Boss?* (Which man, who founded the company Apple, | was a progressive boss?); late planning: *Welcher Mann, der ein fortschrittlicher Boss war, gründete das Unternehmen Apple |?* (Which man, who was a progressive boss, founded the company Apple |?)).

All questions were recorded in a female voice. They had a mean duration of 6 seconds (including 200 ms of initial silence), with a standard deviation of 1.18 seconds. While early and late questions did not differ greatly in length (early questions: mean (sd) = 5.84 s (1.0 s); late questions: mean (sd) = 6.16 s (1.33 s)), filler questions were slightly longer than critical questions (critical: mean (sd) = 5.48 s (0.88 s); fillers: mean (sd) = 6.53 s (1.21 s)).

Two balanced experimental lists were composed, so that questions were presented in only one of the conditions to each subject. The order of presentation of items within the list was random for every participant.

2.3 Procedure

2.3.1 Main Experiment

In the Main Experiment, participants were instructed to use their headphones to listen to the questions during the quiz part of the experiment and to respond verbally to the questions as fast and accurately as possible. They were made aware that not all the questions they would hear during the experiment were correct, using the example question *"Who assassinated US President Clinton?"* and the correct answer *"Nobody. Clinton has not been assassinated."*. They were further instructed to pay special attention to the fact that some questions would be incorrect and to answer them accurately. Lastly, they were instructed to say *I don't know* in response to any question that they did not know the answer to.

Each trial began with a fixation cross in the center of the screen for one second, followed by the auditory presentation of the question, which participants had to answer verbally. Participants' responses were recorded using their PC microphones and participants were instructed to press the space bar after they gave their response. One second after they pressed the space bar, the next trial started. Before the experiment, participants did four practice trials to get to know the procedure.

The quiz part was followed by a post-test questionnaire testing participants' knowledge of the correct versions of the 30 critical questions used in the quiz. Participants read questions asking about the critical information in each of the critical questions of the quiz (e.g., *"Wen besuchte Rotkäppchen, als sie vom Wolf gefressen wurde?"* (Whom did Little Red Riding Hood visit when she was eaten by the wolf?)) and typed their response in a text box. Questions appeared one at a time, replacing each other each time participants pressed 'enter' to confirm their response. The whole experiment took about 20 minutes.

2.3.2 Control Experiment

In the first part of the Control Experiment, participants were instructed to not respond to the questions they heard, but to judge whether the question was correct or erroneous and indicate their choice by clicking on the respective radio button. As a third alternative, participants could indicate that they did not know whether the question was correct or not. One second after participants confirmed their response by pressing the space bar, the next trial started. As well as in the Main Experiment, participants were made aware that not all the questions they would hear during the experiment were correct, using the same example. Participants did four practice trials prior to the experiment. The post-test questionnaire was the same as in the Main Experiment.

3 Results

3.1 Response Latencies

Response latencies were annotated manually in Audacity with respect to question offset and response onset. Of the total of 1440 trial recordings, 5 did not contain any response and were thus discarded. The remaining responses had a mean latency of 1650 ms (sd = 1220 ms; see **Figure 1**). Response latencies were fitted with a Bayesian mixed effects regression model with the R package *brms* (Bürkner, 2017; R Core Team, 2021), with Condition (early planning / late planning) and Type (critical / filler) plus their interaction as fixed effects and as random effects by subject and by item (see Appendix). Both factors were dummy coded, with early planning and critical condition as reference levels. Additionally, the centred duration of the questions in seconds was added as a control variable to the fixed effects structure of the model, since turn duration has been shown to affect turn transition times (Barthel et al., 2016; Magyari, 2015; Roberts et al., 2015), an effect that replicated here (β = -171.54 ms, CI = [-277.65 ms; -72.27 ms]). The prior for the Intercept was set to be normally distributed, with a mean of 1600 ms and a standard deviation of 2500 ms. Priors for the coefficients were vaguely informative as they were set to be normally distributed, with a mean of 0 and a standard deviation of 500 ms.

The model detected both a decisive main effect of Condition (β = -282.16 ms; CI = [-398.36 ms; -166.92 ms], BF_{10} = 5999)[3] as well as a strong main effect of Type (β = -211.55 ms; CI = [-422.47 ms; 6.98 ms]; BF_{10} = 16.86), indicating that early questions were responded to faster than late questions and that filler questions were responded to faster

[3]For a guideline to the interpretation of Bayes factors, see Andraszewicz et al. (2015).

than critical questions. However, the model also attested a decisive interaction effect of Condition × Type (β = 496.19 ms; CI = [251.73 ms; 744.07 ms]; BF_{10} = 5999). To investigate the origin of this interaction effect, hypothesis specific tests were conducted using the *hypothesis* function built into *brms*, which revealed that Condition had a decisive effect on response latencies only in filler trials (β = 530.26 ms; CI = [368.85 ms; 690.32 ms]; BF_{10} > 6000), but no effect in critical trials (β = 34.07 ms; CI = [-116.85 ms; 184.98 ms], BF_{10} = 1.88), indicating that responses were given faster in early than in late questions in filler trials but equally fast in critical trials. However, an additional model that was run only on the subset of critical trials in which participants accepted the illusion (see Appendix) revealed that in these trials Condition did have an effect on response latencies (β = 126.01 ms; CI = [-97.65 ms; 346.11 ms], BF_{10} = 4.88), indicating that response latencies in late questions were longer than in early questions when participants gave the expected response. While going in the same direction, this effect was smaller and statistically weaker than in filler trials, possibly because critical questions were slightly more difficult than filler questions and because this test relied on a lower number of observations (N_{early}= 118; N_{late} = 90).

3.2 Semantic Illusions

For an analysis of the proportions of semantic illusions that were detected or not, only critical trials were analysed. 25 trials with unintelligible or nonsense responses, 143 trials in which participants responded that they did not know the answer to the question, and 111 trials where participants revealed in the post-test questionnaire that they did not have the necessary factual knowledge to spot that the original quiz question contained an illusion were discarded, leaving 439 responses for analyses. Responses were coded as accepting the illusion when the expected answer was given, or as rejecting the illusion when the illusion was spotted. Descriptively, illusions were accepted in 84 out of 226 trials (37%) in the early planning condition and in 58 out of 213 trials (27%) in the late planning condition. The probability of the illusion being accepted was fitted with a Bayesian mixed effects regression model (family = bernoulli) with Condition as a dummy-coded fixed effect and as a random effect by subject and by item, with 'rejected' as the reference level (see Appendix). Condition was found to have a strong effect on the proportion of accepted illusions (β = -0.49; CI = [-0.96; -0.03]; BF_{10} = 22.67; with the best estimate for the Intercept at β = -0.67), showing strong evidence for the probability of illusions to be accepted to be higher in the early planning condition than in the late planning condition (see **Figure 2**).

For the Control Experiment, also only critical trials were analysed.[4] In 112 trials, participants indicated that they did not know whether the question was correct or not, and in 94 of the remaining trials, the post-test revealed that participants did not have the necessary factual knowledge to spot the original illusion, leaving 484 trials for analyses. Descriptively, illusions passed in 24 out of 239 trials (1.0%) in the early condition and in 22 out of 245 trials (0.9%) in the late condition. A model parallel to the Main Experiment model was fitted (see Appendix), which revealed that Condition had a very weak effect on the probability of the illusion being accepted (β = -0.4; CI = [-1.33; 0.49]; BF_{10} = 3.77, with the best estimate for the intercept at β = -2.75), showing merely anecdotal evidence for the probability of illusions to differ between the early and late planning conditions.

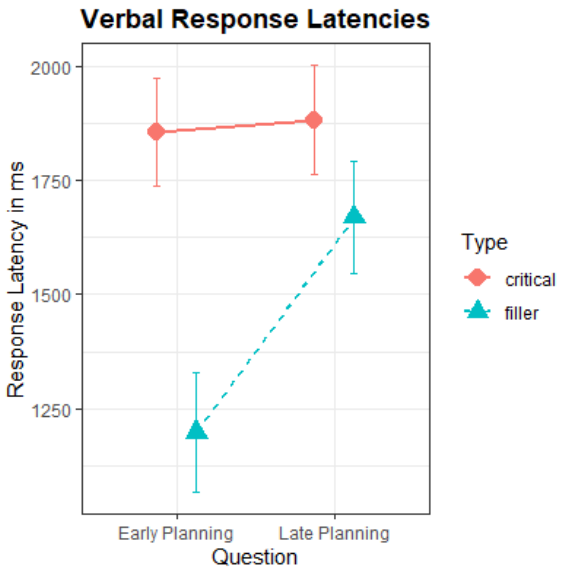


Figure 1: Mean response latencies by condition in critical questions (containing a semantic illusion) and filler questions (not containing an illusion). Error bars indicate 95% confidence intervals.

[4]Data from one participant were discarded because task instructions were ignored in the post-test questionnaire and most responses in the main part were 'I don't know' responses.

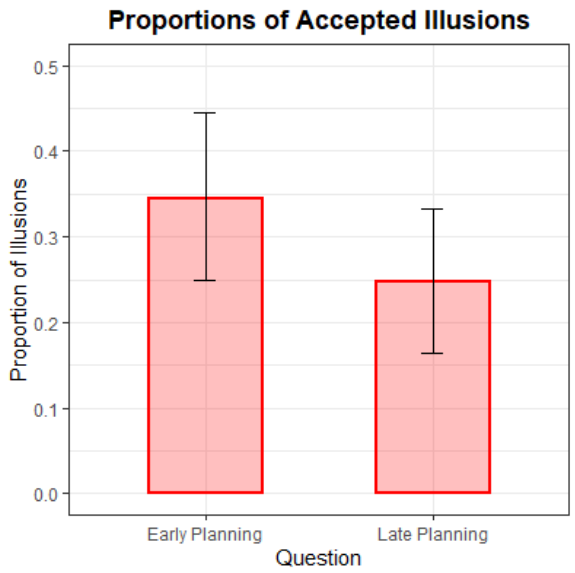


Figure 2: Fitted proportions of illusions that were accepted, i.e., not spotted, by participants in the Main Experiment. Error bars indicate 75% credible intervals. See section 3.2 for model description.

4 Discussion

Conversation is a well-practiced but cognitively demanding dual-task situation, where processes of language comprehension and speech planning can interfere with one another, either due to cross talk between the representations that are relevant for each of the tasks or due to limited resources that need to be shared between the tasks. This study tested two competing hypotheses about the default allocation of processing capacities in moments of increased cognitive load in a dialogic task. In conversational turn taking, cognitive load is especially high in next speakers when they are concurrently planning their upcoming turn and listening to the incoming turn. The *comprehension-prioritized hypothesis* states that when processing capacities are temporarily limited, processing of the incoming speech would be prioritized over response planning because the rate of incoming information to be processed is defined by the speech rate of the incoming turn and is thus not under the control of the listener/next speaker. If the incoming speech is not processed thoroughly at the rate it is coming in, part of the signal would be lost, which might be an undesirable characteristic of any adapted processing strategy. The *turn-timing-prioritized hypothesis*, on the other hand, states that well-timed responses are central in order to convey the intended messages in a conversational situation. And in order to be able to deliver the next turn quickly in response to the incoming turn, response planning needs to start and progress rapidly in overlap with the incoming turn. When this dual-task situation leads to increased processing load, response planning would be prioritized so as to not jeopardize seamless turn-timing. These two hypotheses were tested in a question-answer paradigm with questions containing semantic illusions, such as "How many animals of each kind did *Moses* take on the Ark?"

Previous studies showed that next speakers readily put themselves in the described dual-task situation when they are in conversation, starting to plan their next turn as soon as they can anticipate the message of the incoming turn, even though planning in overlap leads to increased processing load at turn-transitions (Barthel, 2020; Barthel and Sauppe, 2019; Bögels, 2020; Levinson and Torreira, 2015). The response latency results in the present study replicate these previous findings. Here, participants verbally responded to quiz questions. For half of the questions, the response could be planned already in the middle of the question, for the other half of the questions, the response could only be planned at the end of the question. Questions whose responses could be planned already in overlap with the question were responded to faster than questions whose response could only be known at the question's end, showing that participants started to plan their response already in overlap with the incoming question when this was possible, and thereby achieved shorter response latencies. This attested effect of question format on response latencies needs to be qualified, however. While the effect was observed to be strong in questions that did not contain a semantic illusion, the effect was not attested in questions that did contain a semantic illusion. However, post-hoc analyses did reveal the effect in questions containing a semantic illusion, but only when the illusion was not detected, i.e., when the question was answered as would be expected without the illusion. This pattern of results indicates that also in questions containing an illusion, participants started to plan their response as early as possible but had to abandon the planning process when they detected a mismatch between their anticipation of the upcoming input and the actual continuation of the question. In these cases, participants had to begin planning from scratch, this time to reject the illusion, and therefore did

not show any observable gain from early response planning.

Overall, about one third of the questions containing an illusion were accepted and answered as would be expected without the illusion. This acceptance rate was affected by whether planning in overlap with the illusion word was possible or not. Participants failed to detect the illusion more often when they were planning in overlap than when they were not concurrently planning, showing that speech planning in overlap is detrimental to semantic input processing. While the early and late planning questions differed in their format and in the position of the critical word containing the illusion, these differences were not driving the effect. This possibility can be excluded on the basis of the results of the control experiment, in which response planning was not necessary. When participants rated the questions for correctness instead of answering them verbally, the position effect disappeared.[5] This pattern of results supports the *turn-timing-prioritized hypothesis*, which predicted that in phases of high processing load in dialogue situations, dynamic progress in response planning would be prioritized over deep processing of the input, so that the response is ready for articulation shortly after the incoming turn comes to an end. The results do not support the *comprehension-prioritized hypothesis*, as comprehension was found to be less accurate when planning was executed in overlap with the question. Instead, the results indicate that in these phases, participants processed the input more shallowly and based their response planning on their anticipations of the question continuations (Ferreira and Patson, 2007; Ferreira et al., 2002; Song and Schwarz, 2008; Van Oostendorp and De Mul, 1990; van Oostendorp and Kok, 1990). Shallow input processing can be assumed to occur most prominently in situations when processing load increases, which are to be most frequent before turn-transitions (Barthel and Sauppe, 2019).

Prioritizing response planning can indeed be argued to be an efficient strategy in dialogue, even if it might be at the expense of comprehension accuracy. Listeners in conversation have been found to generate predictions about the incoming turn in order to be able to start planning their response early on the basis of their predictions (Corps et al., 2018; Magyari et al., 2014; Gisladottir et al., 2015, 2018). This early planning enables them to take their next turn quickly after the incoming turn ends. The fact that most turn-transitions are fast makes conversation efficient with respect to the utilisation of the available time, and on top of that, it is the basis for turn-timing to be interpreted as meaningful when transitions are slow (Henetz, 2017; Roberts and Francis, 2013; Roberts et al., 2011). In the majority of cases, predictions about the message of the end of the incoming turn are probably correct, as turn endings are often predictable (Magyari and de Ruiter, 2012). In these cases, relying on the predictions is certainly an efficient strategy. In cases where the upcoming input does not match the predictions, two reasons for the mismatch come to mind. Either the input was 'wrong', i.e., not as intended by the current speaker, e.g., when they erroneously replaced *Noah* with *Moses* (Fromkin, 1971; Meringer, 1908; Levelt, 1989), in which case the prediction was actually 'right' and the conversation can continue smoothly even if the error passes unnoticed. Or the prediction was wrong, which would lead to misunderstanding if the mismatch passes unnoticed. These latter, problematic cases can be considered to be rare enough in natural conversation for the turn taking system to be efficient, and if they do arise, they are commonly detected and dealt with by the interactants immediately in the next turn with the help of repair sequences (Dingemanse et al., 2015; Schegloff, 1992). Prediction and planning strategies can be argued to be readily built upon this safety-net that comes with conversational repair, as repair mechanisms are general purpose tools that are used for any form of misunderstanding, e.g., in problems in acoustic understanding or reference matching, and are not specific to fixing the consequences of prediction errors.

It remains difficult to judge the relevance of prediction for the probability of an illusion to pass unnoticed. Given that interlocutors predict the end of an incoming turn in order to prepare their response (e.g., Corps et al., 2019), the difference in illusion rates could be due to a higher predictability of the target word in the early planning condition than in the late planning condition. This line of thought would assume that comprehension is more shallow when predicting the input. This is indeed

[5] While response planning was not prohibited with certainty in the control experiment, the absence of an effect of question type indicates that subjects did not engage in response preparation but rather focused on comprehending the question in the control task.

possible and would underpin the assumption that input prediction as a conversational strategy is efficient because more resources are available for planning earlier before the end of the incoming turn. It is thus conceivable that a higher rate of illusions might not be due to the planning itself, but due to prediction of the target word or concept that was replaced by the illusion word. The absence of an effect of question type in the control experiment (where no response to the question was given) could be argued to refute this idea, since the questions were the same as in the main experiment and the target words were therefore equally predictable. However, due to the different tasks, not only response planning but also input prediction might have been reduced in the control experiment, which could have eliminated the effect of question type. In the context of the tasks of the present study, the two sides of the medal might be too closely coupled to tease apart their contributions. Arguably though, subjects might have engaged less in input prediction in the control experiment, *because* there was no need for response preparation. Consequently, comprehension of the input suffered when planning the next turn as compared to when not planning the next turn, possibly mediated by input prediction.

Contrary to the assumption that language input is processed more shallowly when predictions are maintained, it would also be sensible to argue that word substitutions should be *more* obvious when a prediction to hear a different word has already been generated. In this line of thought, not shallow comprehension but rather processing ease at the time of encounter should follow from prediction, which is corroborated by findings that comprehension is less effortful in high predictability sentences (e.g., Obleser and Kotz, 2011). In that case, a higher rate of illusions would be indicative of shallow comprehension due to concurrent response planning rather than due to prediction. Under these considerations, it should be easier to detect a word substitution when the input is predictable, so that lower illusion rates would be expected in the questions with word substitutions at their ends (i.e., in the early planning questions). The fact that the opposite pattern of results was found thus speaks for the interpretation that concurrent response planning rather than prediction was responsible for the differential illusion effect. Future research will be needed to conclusively disentangle the relative contributions of these two confounded factors.

One final side-note needs to be added about the comparability of the present study with previous studies investigating semantic illusions. Seeing that planning in overlap increases the rate of semantic illusions brings up the question what relevance this effect might have had in previous studies. This question is difficult to answer conclusively, firstly because the point in time when response planning was possible was not included as a control variable in previous studies, and secondly because results on that question are not directly comparable between studies, since, to the best of our knowledge, all previous studies presented the critical questions in print, whereas the present study is the first to illustrate the occurrence of semantic illusions with auditorily presented speech. What can be said with certainty, however, is that semantic illusions do not depend on speech planning in overlap. In the classic experiment by Erickson and Mattson (1981), two of the four critical questions contained the illusion word in a position after the question can be known and two contained the illusion word before the question can be known. In their study, both types of questions were reported to elicit semantic illusions. Moreover, also the present study found evidence that concurrent response planning is certainly not a prerequisite for semantic illusions to arise. Still, semantic processing of the input was found to be less effective during speech planning, and future studies should take this factor into account, either by controlling for the position of the illusion with respect to the point where planning can begin, by balancing their materials, and/or by statistically controlling for the influence of the factor post-hoc.

5 Conclusion

Semantic illusions have been found to be stronger when speech planning is executed while comprehending the illusory input than without concurrent speech planning. Hence, semantic processing of language input in dialogic situations can be assumed to be more shallow during speech planning, even when the planned content is contingent upon the content of the incoming speech. The effect of concurrent planning on semantic processing is possibly due to limited processing resources operating on related linguistic representations, so that next speakers need to strike an efficient balance between comprehension and planning before turn

transitions. Nonetheless, as it is a prerequisite for seamless turn-timing, speech planning in overlap with comprehension is a communicatively effective strategy, as it is a cornerstone of the turn taking system that forestalls abundant long gaps and allows turn-timing to be interpreted as meaningful by interlocutors. In sum, planning the next turn in overlap with the incoming turn does not seem to be efficient from a processing perspective, as comprehension accuracy suffers from concurrent speech planning. Still, prioritizing planning the next turn under high processing load before turn transitions could be a very effective strategy for communication, and the present experiment provides evidence that planning is prioritized over accurate comprehension in periods when these processes compete for cognitive resources.

Acknowledgments

This study was funded by the Berlin University Alliance and conducted within the X-Student Research Group *Language Comprehension in Dialogue*, led by the author. This paper has been improved by three valuable anonymous reviews. The author thanks Hannah Ida Hullmeine for recording the experimental stimuli.

References

Abdel Rahman, R. and Melinger, A. (2019). Semantic processing during language production: an update of the swinging lexical network. *Language, Cognition and Neuroscience*, 34(9):1176–1192.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., and Wagenmakers, E.-J. (2015). An Introduction to Bayesian Hypothesis Testing for Management Research. *Journal of Management*, 41(2):521–543.

Barthel, M. (2020). *Speech Planning in Dialogue - Psycholinguistic Studies of the Timing of Turn Taking*. PhD Thesis, Radboud University Nijmegen, Nijmegen.

Barthel, M. and Levinson, S. C. (2020). Next speakers plan word forms in overlap with the incoming turn: evidence from gaze-contingent switch task performance. *Language, Cognition and Neuroscience*, 35(9):1183–1202.

Barthel, M., Meyer, A. S., and Levinson, S. C. (2017). Next Speakers Plan Their Turn Early and Speak after Turn-Final "Go-Signals". *Frontiers in Psychology*, 8:393.

Barthel, M. and Sauppe, S. (2019). Speech planning at turn transitions in dialog is associated with increased processing load. *Cognitive Science*, 43(7):e12768.

Barthel, M., Sauppe, S., Levinson, S. C., and Meyer, A. S. (2016). The Timing of Utterance Planning in Task-Oriented Dialogue: Evidence from a Novel List-Completion Paradigm. *Frontiers in Psychology*, 7:1858.

Boiteau, T. W., Malone, P. S., Peters, S. A., and Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General*, 143(1):295–311.

Bond, Z. S. (1999). *Slips of the ear: errors in the perception of casual conversation*. Academic Press, San Diego, Calif.

Buchsbaum, B. R., Hickok, G., and Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science*, 25(5):663–678.

Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, 203:104347.

Bögels, S., Magyari, L., and Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, 5(12881):1–11.

Bürki, A., Elbuy, S., Madec, S., and Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language*, 114:104125.

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1).

Corps, R. E., Crossley, A., Gambi, C., and Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, 175:77–95.

Corps, R. E., Pickering, M. J., and Gambi, C. (2019). Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience*, 34(5):615–627.

Daliri, A. and Max, L. (2016). Modulation of Auditory Responses to Speech vs. Nonspeech Stimuli during Speech Movement Planning. *Frontiers in Human Neuroscience*, 10.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., and Enfield, N. J. (2015). Universal Principles in the Repair of Communication Problems. *PLOS ONE*, 10(9):e0136100.

Erickson, T. D. and Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5):540–551.

Fargier, R. and Laganaro, M. (2016). Neurophysiological Modulations of Non-Verbal and Verbal Dual-Tasks Interference during Word Planning. *PLOS ONE*, 11(12):e0168358.

Fargier, R. and Laganaro, M. (2019). Interference in speaking while hearing and vice versa. *Scientific Reports*, 9(1):5375.

Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, 11(1):11–15.

Ferreira, F. and Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, 1(1-2):71–83.

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, 47(1):28.

Gisladottir, R. S., Bögels, S., and Levinson, S. C. (2018). Oscillatory Brain Responses Reflect Anticipation during Comprehension of Speech Acts in Spoken Dialog. *Frontiers in Human Neuroscience*, 12.

Gisladottir, R. S., Chwilla, D. J., and Levinson, S. C. (2015). Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances. *PLOS ONE*, 10(3):1–24.

Hagoort, P. and Indefrey, P. (2014). The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, 37(1):347–362.

He, J., Meyer, A. S., and Brehm, L. (2021). Concurrent listening affects speech planning and fluency: the roles of representational similarity and capacity limitation. *Language, Cognition and Neuroscience*, pages 1–23.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Henetz, T. (2017). *Don't hesitate! The length of interturn gaps influences observers' interactional attributions*. PhD Thesis, Stanford University, Stanford.

Indefrey, P. and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.

Jescheniak, J. D., Matushanskaya, A., Mädebach, A., and Müller, M. M. (2014). Semantic interference from distractor pictures in single-picture naming: evidence for competitive lexical selection. *Psychonomic Bulletin & Review*, 21(5):1294–1300.

Kempen, G., Olsthoorn, N., and Sprenger, S. (2012). Grammatical workspace sharing during language production and language comprehension: Evidence from grammatical multitasking. *Language and Cognitive Processes*, 27(3):345–380.

Kendrick, K. H. and Torreira, F. (2014). The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes*, 52(4):1–35.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language*, 66(1):143–162.

La Heij, W., Dirkx, J., and Kramer, P. (1990). Categorical interference and associative priming in picture naming. *British Journal of Psychology*, 81(4):511–525.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press, London.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., and Pechmann, T. (1991). The Time Course of Lexical Access in Speech Production: A Study of Picture Naming. *Psychological Review*, 98(1):122–142.

Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):10–26.

MacKay, D. G. (1987). *The Organization of Perception and Action*. Cognitive Science Series. Springer New York, New York, NY.

Magyari, L. (2015). *Timing Turns in Conversation: A temporal preparation account*. PhD Thesis, Radboud University Nijmegen, Niljmegen.

Magyari, L., Bastiaansen, M. C. M., de Ruiter, J. P., and Levinson, S. C. (2014). Early Anticipation Lies behind the Speed of Response in Conversation. *Journal of Cognitive Neuroscience*, 26(11):2530–2539.

Magyari, L. and de Ruiter, J. P. (2012). Prediction of Turn-Ends Based on Anticipation of Upcoming Words. *Frontiers in Psychology*, 3(376):1–9.

Menenti, L., Gierhan, S. M. E., Segaert, K., and Hagoort, P. (2011). Shared Language: Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, 22(9):1173–1182.

Meringer, R. (1908). *Aus dem Leben der Sprache*. B. Behr's, Berlin.

Meyer, A. S. (1996). Lexical Access in Phrase and Sentence Production: Results from Picture–Word Interference Experiments. *Journal of Memory and Language*, 35(4):477–496.

Obleser, J. and Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, 55(2):713–723.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Roberts, F. and Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Roberts, F., Margutti, P., and Takano, S. (2011). Judgments Concerning the Valence of Inter-Turn Silence Across Speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5):331–354.

Roberts, S. G., Torreira, F., and Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6.

Roelofs, A., Özdemir, R., and Levelt, W. J. M. (2007). Influences of spoken word planning on speech recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):900–913.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

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

Schriefers, H., Meyer, A. S., and Levelt, W. (1990). Exploring the Time Course of Lexical Access in Language Production: Picture-Word Interference Studies. *Journal of Memory and Language*, 29:86–102.

Shafto, M. and MacKay, D. G. (2000). The Moses, mega-Moses, and Armstrong illusions: integrating language comprehension and semantic memory. *Psychological Science*, 11(5):372–378.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., and Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, 111(43):E4687–E4696.

Song, H. and Schwarz, N. (2008). Fluency and the Detection of Misleading Questions: Low Processing Fluency Attenuates the Moses Illusion. *Social Cognition*, 26(6):791–799.

Speckmann, F. and Unkelbach, C. (2020). Moses, money, and multiple-choice: The Moses illusion in a multiple-choice format with high incentives. *Memory & Cognition*, 49(4):843–862.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Van Oostendorp, H. and De Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica*, 74(1):35–46.

van Oostendorp, H. and Kok, I. (1990). Failing to notice errors in sentences. *Language and Cognitive Processes*, 5(2):105–113.

A Appendix - Bayesian regression models

Group-Level Effects:				
∼itemID (Number of levels: 60)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	439.04	61.03	326.68	567
sd(condition=late)	248.78	97.22	42.79	427.24
cor(Intercept,condition=late)	-0.54	0.27	-0.91	0.11
∼subjectID (Number of levels: 24)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	551.03	93.93	397.13	764.51
sd(condition=late)	136.81	87.58	5.97	323.47
cor(Intercept,condition=late)	-0.04	0.44	-0.84	0.86
Population-Level Effects:	Estimate	Est.Error	l-95% CrI	u-95% CrI
Intercept	1743.50	151.24	1450.64	2034.21
condition=late	34.07	92.59	-148.27	215.21
type=filler	-459.65	146.55	-745.62	-172.12
questionDuration_centered	-171.54	52.52	-277.65	-72.27
condition=late:type=filler	496.19	125.70	251.73	744.07
Residual Error:	Estimate	Est.Error	l-95% CrI	u-95% CrI
sigma	998.81	19.91	961.37	1038.92

Table 1: Model output of main reaction times model. Family = gaussian. Link = identity. Formula = responseLatency_inms ~ 1 + condition * type + questionDuration_c + (1 + condition | subjectID) + (1 + condition | itemID). Number of observations = 1435. Samples = 3 chains, each with iter = 3000; warmup = 1000; thin = 1. Factor reference levels: condition = early; type = critical.

Group-Level Effects:				
∼itemID (Number of levels: 30)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	461.46	124.87	224.50	719.84
∼subjectID (Number of levels: 24)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	414.23	106.40	230.71	650.00
Population-Level Effects:	Estimate	Est.Error	l-95% CrI	u-95% CrI
Intercept	1528.37	161.52	1208.68	1844.13
condition=late	126.01	134.40	-141.33	388.87
questionDuration_centered	-136.46	131.12	-397.47	123.19
Residual Error:	Estimate	Est.Error	l-95% CrI	u-95% CrI
sigma	912.35	53.02	815.92	1022.51

Table 2: Model output of reaction times model on subset of accepted illusions. Family = gaussian. Link = identity. Formula = responseLatency_inms ~ 1 + condition + questionDuration_c + (1 | subjectID) + (1 | itemID). Number of observations = 208. Samples = 3 chains, each with iter = 3000; warmup = 1000; thin = 1. Factor reference level: condition = early

Group-Level Effects:				
∼itemID (Number of levels: 29)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	0.81	0.24	0.39	1.33
sd(condition=late)	0.37	0.29	0.02	1.06
cor(Intercept,condition=late)	-0.11	0.55	-0.95	0.92
∼subjectID (Number of levels: 24)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	1.48	0.32	0.95	2.19
sd(condition=late)	0.31	0.25	0.01	0.93
cor(Intercept,condition=late)	0.01	0.57	-0.94	0.95
Population-Level Effects:	Estimate	Est.Error	l-95% CrI	u-95% CrI
Intercept	-0.67	0.39	-1.46	0.09
condition=late	-0.49	0.28	-1.06	0.06

Table 3: Model output of model on the rate accepted illusions in the Main Experiment. Family = bernoulli. Link = logit. Formula = responseLatency ~ condition + (1 + condition | subjectID) + (1 + condition | itemID). Number of observations = 439. Samples = 3 chains, each with iter = 6000; warmup = 2000; thin = 1. Factor reference level: condition = early

Group-Level Effects:				
∼itemID (Number of levels: 29)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	1.27	0.42	0.57	2.20
sd(condition=late)	0.37	0.29	0.02	0.52
cor(Intercept,condition=late)	-0.16	0.56	-0.97	0.92
∼subjectID (Number of levels: 24)	Estimate	Est.Error	l-95% CrI	u-95% CrI
sd(Intercept)	0.55	0.36	0.02	1.37
sd(condition=late)	1.00	0.56	0.07	2.20
cor(Intercept,condition=late)	-0.10	0.55	-0.94	0.91
Population-Level Effects:	Estimate	Est.Error	l-95% CrI	u-95% CrI
Intercept	-2.75	0.45	-3.77	-1.98
condition=late	-0.40	0.56	-1.55	0.66

Table 4: Model output of model on the rate accepted illusions in the Control Experiment. Family = bernoulli. Link = logit. Formula = responseLatency ˜ condition + (1 + condition | subjectID) + (1 + condition | itemID). Number of observations = 439. Samples = 3 chains, each with iter = 6000; warmup = 2000; thin = 1. Factor reference level: condition = early

Context Is Key: Annotating Situated Dialogue Relations in Multi-floor Dialogue

Claire Bonial[1], Mitchell Abrams[2], Anthony L. Baker[1], Taylor Hudson[3], Stephanie M. Lukin[1], David Traum[4], and Clare R. Voss[1]

[1]U.S. Army Research Laboratory, Adelphi, MD 20783
[2]Institute for Human and Machine Cognition, Pensacola, FL 32502
[3]Oak Ridge Associated Universities, Oak Ridge, TN 37831
[4]USC Institute for Creative Technologies, Playa Vista, CA 90094
claire.n.bonial.civ@mail.mil

Abstract

In order to account for the features of *situated* dialogue, we extend a multi-party, multi-floor dialogue annotation schema so that it uniquely marks turns with language that must be grounded to the conversational or situational context. We then annotate a dataset of 168 human-robot dialogues using our extended, situated relation schema. Despite the addition of nuanced dialogue relations that reflect the kind of context referenced in the language, our inter-annotator agreement rates remain similar to those of the original annotation schema. Crucially, our updates separate data that can be used to train dialogue systems in essentially any context from those utterances in the data that are only appropriate in a particular situated environment.

1 Introduction

In order to account for the features of situated dialogue, we extend our multi-floor dialogue annotation schema described in Traum et al. (2018) to better capture the nuances that arise when a human and a robot collaborate on a search-and-navigation task. Using the same data collection procedure—a "Wizard-of-Oz" experimental design (Riek, 2012), in which participants directed what they believed to be an autonomous robot to complete search-and-navigation tasks (Marge et al., 2016, 2017)—we collect 168 human-robot dialogues and subsequently annotate them with a novel situated dialogue relation schema we present in this paper. While the original dialogue annotation schema is effective for multi-floor dialogue, it does not provide any indicator in the annotation schema demonstrating where the language requires grounding in the conversational or physical context.

In this paper, we address this problematic gap in the original annotation schema (described in §2) by introducing eight new annotation categories that uniquely mark where a particular interpretation/execution of the input natural language instruction relies upon some knowledge of the context—physical context, conversational context, or the robot's own physical form and abilities, and the interplay of these factors (updates described in §3). We annotate 168 additional dialogues with the augmented annotation schema and provide a corpus analysis (§4.1) and inter-annotator agreement (IAA) analysis (§4.2), which demonstrates that IAA remains high despite introducing new and somewhat nuanced annotation categories. We thus contribute an annotation schema and corpus that is better suited to serve as training data for situated dialogue systems by identifying precisely where the language must be grounded within the conversational or physical context in order to be interpreted and executed correctly.

2 Background

2.1 Human-Robot Dialogue Data

The experimental design of the human-robot dialogue data collection breaks up the planned autonomous robot capabilities into dialogue and navigation components, with one human experimenter, or "wizard," standing in for each component (depicted in Figure 1). A participant, acting as the "Commander," issues verbal instructions to their remotely located robot partner. Their instructions are heard and responded to by the "Dialogue Manager"

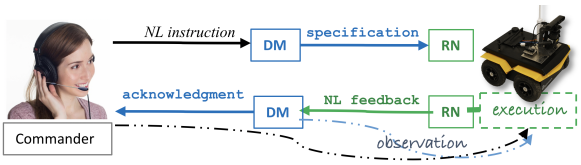


Figure 1: Natural Language (NL) and execution pairings for translate right (top) and acknowledgement (bottom). Dialogue turns for Commander in black, dialogue manager (DM) blue, robot navigator (RN) green.

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

(DM), whose role is to pass on a simplified version of the instructions, or "specification", via text message to the "Robot Navigator" (RN). The RN then joysticks the robot to execute the instruction, and this motion can be observed by the Commander and DM on a dynamically updating 2D LIDAR[1] map. The RN indicates completion or any problems in spoken natural language, and that status is acknowledged/described by the DM for the Commander. This data is therefore *multi-floor dialogue,* where communications between the Commander and DM is considered one conversational floor, while communications between the DM and the RN is considered another conversational floor. Note also that the information available to interlocutors within each floor is distinct—The Commander is unfamiliar with the remote environment and can *only* understand this environment based on the 2D LIDAR map that builds and updates as the robot enters a new space, while the DM and RN (i.e., the robot) are familiar with the space and are furnished with a map of the environment populated with unique names for each of the landmarks and spaces.

In previous work, we collected 60 human-robot dialogues following this protocol (Bonial et al., 2017; Marge et al., 2017). These dialogues were annotated using the schema presented in Traum et al. (2018), and used to train robot dialogue systems (Lukin et al., 2018; Gervits et al., 2021). In these efforts, input language from the Commander is associated with both feedback utterances from the DM to the Commander, as well as specifications for executing the input instructions in the form of the DM's "translation" of those instructions that are sent to the RN. Thus, on a high level, the dataset can be thought of as comprised of pairings of input natural language instructions with specifications for execution, and the complementary pairings of execution and generated natural language descriptions of what will be done, what is being done, or what has been done (see Figure 1).[2] Although effective for some natural language input, the annotation schema did not distinctly mark places where a particular pairing of natural language and execution was *only* valid in the conversational or physical context in which it arose. As a result, dialogue systems trained on this data could not handle, for example, input language that referenced a particular landmark in the current physical context, such as *the door ahead on the right*. Our modified annotation schema addresses this gap.

[1]Light Detecting and Ranging sensor

[2]Specifications are a controlled language, constrained to utterances included in the DM's wizard GUI described in Bonial et al. (2017), but are not a robotic planning language.

2.2 Dialogue Annotation

There are a variety of annotation schemas for dialogue available, including ISO standards for both dialogue acts (Bunt et al., 2012) and discourse relations (Prasad and Bunt, 2015). While these offer relations appropriate to dialogue, and perhaps even multi-party dialogue, they do not address the intricacies and challenges of *multi-floor dialogue*. Multi-floor dialogue is the focus of our original annotation schema, and is defined as "cases in which the high-level dialogue purposes are the same, and some content is shared, but other aspects of the information state, such as the participant structure and turn-taking expectations, are distinct" (Traum et al., 2018, p. 104).

This dialogue annotation schema was used to annotate the multi-floor, human-robot dialogue dataset described in Section 2.1. A dialogue excerpt from the dataset is given in Table 1. The annotation follows Grosz and Sidner (1986)'s *intentional structure* using the TRANSACTION UNIT (TU), which comprises an initial message from one speaker and all subsequent utterances across all floors that address the intention of that initial message. The internal structure of the TU is annotated using RELATIONS (rels) describing how a subsequent utterance relates to, or addresses, a previous utterance or its ANTECEDENT (ant). Relations are organized into a taxonomy of types where higher-order categories are distinguished based on whether they describe relations between utterances within or across conversational floors, and within or across speakers in a single floor. EXPANSIONS are relations between utterances of the same speaker and within the same conversational floor. RESPONSES are relations between utterances by different speakers within the same floor. TRANSLATIONS are relations between utterances in different conversational floors. Within each of these broad relation types, there is at least one but often two levels of relation subtypes. For example, TRANSLATIONS have two subtypes to characterize whether the information is being translated from the left floor to the right floor, TRANSLATION-**R**IGHT, or from the right floor to the left floor, TRANSLATION-**L**EFT (see Table 1). In contrast, RESPONSE has 17 subtypes, including ACKNOWLEDGMENT relations, which in turn has 8 subtypes, including **ACK**NOWLEDGMENT-DOING

	Left Floor		Right Floor		Annotations			
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**	**New Rel**
1	turn east ninety degrees				1			
2	and travel three feet				1	1	continue	
3		processing...			1	2*	processing	
4			turn left 90 degrees		1	1	translation-r	translation-r-situated
5			then...		1	4	link-next	
6			move forward 3 feet		1	2	translation-r	translation-r-default
7		turning...			1	1	ack-doing	
8		moving...			1	2	ack-doing	
9				done	1	6*	ack-done	
10		done			1	9	translation-l	

Table 1: Annotation exemplifying one TU with a situated translation (#4) and a default assumption that *travel* involves forward movement (#6), shown with original relations of Traum et al. (2018) and updated relations.

and ACKNOWLEDGMENT-DONE, which indicate that an instruction is being or has been carried out (see Table 1). For full details of all relation types, we refer the reader to Traum et al. (2018).

3 Situated Annotation Schema

We are addressing not only multi-party, multi-floor dialogue, but also *situated* dialogue that often draws upon the surrounding physical context, as well as the dialogue history and some assumptions relevant to the robot's own embodied form and capabilities. While our original annotation schema is uniquely suited to multi-floor dialogue, we have made several modifications to address the situated, contextual nature of the multi-floor dialogue found in the data.[3] These additions are summarized in Table 2, where we have also listed the directly relevant original annotation categories that we expanded upon. Our additions are made to two main relation types:

i. TRANSLATIONS from the left floor to the right floor, which allow us to pinpoint where and how certain translations draw upon the physical or conversational context (§3.1);

ii. ACKNOWLEDGMENTS of a preparatory action, not explicitly instructed, that is needed given the particular situated context or the particular capabilities and behaviors of the robot (§3.2).

[3]Note that although we see broad applicability of the annotation scheme, particularly the high-level types, with initial attempts to annotate other multi-floor dialogue corpora, such as (Martinovski et al., 2003), our approach has been to articulate only the low-level actions that appear in the analyzed data. Thus this research showcases the challenges of a specific domain, task, and robot (see Bonial et al. (2021) for related efforts extending dialogue annotations to a new task and domain).

3.1 Translation Across Floors

Translation across the conversational floors occurs when the speaker conveys the content from one conversational floor to an addressee in another conversational floor. In the data of interest here, TRANSLATIONS occur when the DM passes a message from the Commander in the left floor to the RN in the right floor, TRANSLATION-R (often instructions to be executed by the RN), or when the DM passes a message from the RN in the right floor to the Commander in the left floor, TRANSLATION-L (often feedback on the execution status of instructions). In our original annotation schema, TRANSLATION-L and TRANSLATION-R were the only two translation relations, along with a -PARTIAL flag that was used to indicate if the translation only addressed part of the original instruction. TRANSLATIONS from the left floor to the right are a critical aspect of training dialogue systems, as they provide the association between an unconstrained natural language instruction and a specification for execution by a robot (which has only a constrained behavior set). Essentially, TRANSLATIONS provide critical data for associating language and behavior.

However, we found that one cannot assume that a particular association is applicable in all physical and conversational contexts. In fact, a particular translation from the left floor to the right (TRANSLATION-R) is often valid only in the unique situational and conversational context where it was originally uttered. If such cases are not annotated distinctly from TRANSLATIONS that are valid in any context, this can lead to system responses that were learned in the training data but are not appro-

Relation	Definition
Translation-left	Provides the same content from speaker in right floor to addressee in left floor.
Translation-right	Provides the same content from speaker in left floor to addressee in right floor.
Translation-right-Direct	Uses the same or synonymous words, where the translation is applicable in any physical or conversational context.
Translation-right-Contextual	Draws upon situational or conversational context, but precisely what contextual information is being used is unclear, underspecified, or there are two or more factors.
Translation-right-Landmark	Refers to a unique landmark name known only to members of the right floor.
Translation-right-Situated	Relevant and/or synonymous to the original instruction in the current physical context but does not refer to a unique landmark.
Translation-right-History	All or part of the translation is only relevant given the dialogue history, in which it was established that a certain instruction should be interpreted in a particular way.
Translation-right-Default	Supplements information by relying on some default assumption related to a robot behavior or capability.
Translation-left-Partial	Only translates part of the command of an utterance or sequence.
Translation-right...Partial	Any of the above Translation-r subtypes that only translates part of the command of an utterance or sequence.
Acknowledgment-Will-comply	Acknowledgment of a command and a promise to do it in the future.
Acknowledgment-Doing	Acknowledgment that the speaker understands the command and that its execution is underway.
Acknowledgment-Done	Acknowledgment that a command or prior planned act has been completed successfully.
Acknowledgment-Will-comply Preparation	Acknowledgment of commitment to the preparation step consistent with compliance with the previous command, but not a promise/commitment of full compliance to the complete command (in contrast to will-comply)
Acknowledgment-Doing Preparation	Acknowledgment that the speaker understands the command and a preparation step required for compliance with the command is underway.

Table 2: Summary of added subcategories (shown in grey) and relevant categories from Traum et al. (2018).

priate in the test or use context. To address this issue, we introduce six new TRANSLATION-R subtypes to uniquely distinguish DIRECT TRANSLATIONS, which relate input language and execution specifications that can be consistently linked in any context, from those TRANSLATIONS that are only valid given particular aspects of the situational or conversational context.

3.1.1 Direct Translation

DIRECT TRANSLATIONS convey the content or intent of the speaker in one floor to the addressee in another floor, using the same or synonymous wording, without adding or subtracting content. These translations therefore relate specifications for execution with input language where this relation is consistently applicable—`Turn right 90 degrees` is *always* a valid expression of the specification for executing the instructions *Rotate right 90 degrees* or *Pivot 90 degrees to the right*, etc.[4] Accordingly, when used as training data, a strong association between input language and a particular execution is appropriate, regardless of context. Another example, translating *Take a photo* as `send image` is found in Table 3.

3.1.2 Landmark

In some cases, the instructions given refer to a particular object or landmark in the environment. Because of the nature of the experimental design where the DM and RN experimenters have complete information of the environment, all salient objects, rooms, hallways, and doorways were pre-coded with a unique identifier name. Thus, when the Commander mentions a particular landmark with a general reference (e.g., *Move to the doorway ahead on the right*), the TRANSLATION-R execution specification includes the specific name for that landmark (e.g., `Move through Kitchen-hall doorway`). Like other situational TRANSLATIONS, the reference used in the instructions cannot be consistently paired with a particular referent since *the doorway ahead on the right* will change depending upon the position of the robot. In Table 3, the translation `move into Conf Room` involves a LANDMARK TRANSLATION as the original destination reference of *through the doorway directly in front of you* is shifted to the named landmark `Conf Room`. Having an annotation category specifically for landmark mentions paves the way for experimentation incorporating a grounding system that will associate linguistic references to their referent in the environment.

3.1.3 Situated

In other cases, the instructions leverage spatial references to the environment as opposed to particular landmarks, where the execution specification

[4]Natural language instructions are italicized, while DM specifications are shown in Courier font.

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	go through the doorway directly in front of you				1		
2	and take a photo				1	1	continue
3		processing...			1	2*	processing
4			move into Conf Room		1	1	translation-r-landmark
5			then...		1	4	link-next
6			send image		1	2	translation-r-direct
7		moving...			1	1	ack-doing
8				uh done and sent	1	6*	ack-done
9		done, sent			1	8	translation-l

Table 3: Dialogue exchange exemplifying a LANDMARK TRANSLATION, referring to the unique identifier of the room that is referenced in the movement instruction (#1), and a DIRECT TRANSLATION of the second piece of the instructions (#2) that has distinct wording, but is applicable in any context.

for these instructions use a spatial reference that is only synonymous to the original in the current situated context. For example, in Table 1, the Commander instructs the robot *Turn east ninety degrees* where this is translated to `turn left 90 degrees`, which is only a valid specification for execution in that particular situated context—the robot's current heading is such that left and East are the same. Thus, again, the input language and execution specification cannot be consistently linked in all contexts. Although SITUATED TRANSLATIONS are conceptually a superset that includes LANDMARK TRANSLATIONS, we mark these distinctly as we expect grounding the references of SITUATED TRANSLATIONS to their referents will be more complex; they leverage abstract spatial language and regions as opposed to physical objects with clearer boundaries.

3.1.4 History

In some cases, an expectation for a certain behavior or certain manner of interpreting instructions may be set in the dialogue history, and then referenced later in the specification for execution. For example, several Commanders requested that the robot take a picture of what is in front of it after each movement instruction, to avoid repeating such requests as part of each instruction. As a result, a movement instruction such as *back up five feet* is linked with the translation `back up 5 feet...send image`, despite the fact that the direct antecedent instructions did not mention sending a picture. Instead, this portion of the specification for execution stems from the globally applicable request established in the dialogue history.[5] Other examples of HISTORY include anaphora (e.g., "take a picture of it"), deixis (g., e.g. "do that again"), or ellipsis (e.g., "two more feet"), where the translation includes the full content, part coming from previous TUs. Annotating these cases uniquely from other kinds of situational TRANSLATIONS again prevents the assumption that all cases of, for example, *back up five feet* be associated with an execution specification involving sending a picture, but also paves the way for incorporating higher-level instructions that apply throughout a dialogue by identifying where such instructions are deployed in the specification.

3.1.5 Default

DEFAULT TRANSLATION is applied when the input instruction does not make explicit some information which is instead inferred using a default assumption, generally regarding the robot's behaviors and capabilities. For example, in Table 1, the instruction *travel three feet* is linked with the translation `move forward three feet`, as it is assumed that the robot's default travel behavior would be a forward movement. Such assumptions are changeable based on the task, physical environment, and type of robot, thus the unique annotation category allows for identification of language/execution pairs that are only valid given a certain set of default assumptions.

[5] Instructions drawing upon utterances within the same TU are not annotated as HISTORY. E.g., an open-ended instruction *Move forward*, with a clarification—`How far?` *Three feet*—would have the DIRECT TRANSLATION `Move forward three feet`; the antecedents are the original instruction and clarification.

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	take a picture of the wall on your left				1		
2		processing...			1	1	processing
4			move to left wall		1	1	translation-r-contextual-partial
5			send image		1	4	continue
6		moving...			1	1	ack-doing-prep
7				done and sent	1	5*	ack-done
8		done, sent			1	7	translation-l

Table 4: Dialogue exchange where the translation of the instruction (#1) with the initial movement (#4) is motivated by underspecified and unknown aspects of the situated context, combined with default assumptions regarding where the robot needs to be to take an appropriate picture.

3.1.6 Contextual Translation, Underspecified

The back-off category CONTEXTUAL TRANSLATION is applied in cases where the kind of context used is underspecified such that it is not clear to the annotator what context, precisely, is being drawn upon, or more than one kind of contextual information is drawn upon within the same translation. In Table 4, the translation of the instruction *take a picture of the wall on your left* is translated with multiple steps, starting with the instruction to `move to left wall`, which is motivated by both the current position and orientation of the robot, as well as some default assumptions about the robot's abilities and requirements for taking a picture from a sensible vantage point. Since two types of context are used (the current situated context and default assumptions), CONTEXTUAL type is used.

3.2 Preparatory Actions

ACKNOWLEDGMENTS (Acks) and feedback to the participant are essential for establishing and maintaining common ground as well as trust and transparency in the system. As TRANSLATION-R relates pairs of input language to a specification of execution, ACKS relate the robot behavior/execution to a natural language description of that behavior that provides feedback and insight into what the robot intends to do, is doing, or has done. Reflecting the importance of this in dialogue, our original annotation schema included 8 subtypes of ACKS, capturing not only completion status, but also the level of confidence of the speaker in what was understood and the commitment to complete the instructed task (e.g., ACKNOWLEDGE-UNDERSTAND, ACKNOWLEDGE-UNSURE, ACKNOWLEDGE-TRY).

ACKS that do not match up with the Commander's intended instructions can be perceived as red flags that some miscommunication has taken place. This is beneficial when there is true misunderstanding, which can then be repaired. However, we found other cases where the robot acknowledged an action required to prepare for execution of the main instructed action. The preparatory action was not explicitly requested by the commander, and the commander might not understand the connection, so the DM utterance may be perceived as irrelevant by the Commander, and therefore signal misunderstanding. In fact, the robot has understood the instructions and is simply executing them in a way that reflects additional preparatory steps needed given its abilities. For example, an instruction *Take a picture of what's behind you* requires that the robot used for data collection first turn around 180 degrees before taking a picture, as its camera is a static, front-facing camera. ACKS that the robot will turn or is turning around, however, might not be perceived as appropriate acknowledgments of the original instruction, and therefore may actually undermine trust that the system has understood and is executing the instructions. To capture the fact that such ACKS are distinctly providing feedback on preparatory actions, we introduced two ACK relations described below. In the future, we will explore the potential value of making these acknowledgments explicitly mention the preparatory nature of the action (e.g., `I am turning in preparation to take a picture...`) to prevent the perception of a mismatch between the Commander's intention and the action being carried out.

3.2.1 Will Comply - Preparation

The first added subtype of acknowledgments, WILL COMPLY - PREP is used to mark acknowledg-

ments that reflect the speaker's commitment to do a preparatory step consistent with compliance with the previous command. While the relation type WILL COMPLY is acknowledgment of the speaker's commitment to comply with the command, this is not the case for WILL COMPLY - PREP, as there may be some intervening problem or clarification needed for full compliance, so the latter is a commitment restricted to the preparatory step. For example, the instruction *Go five feet north* requires the preparatory step of the robot turning to face North when its current heading is not in that direction. Note that just acknowledging the preparatory step `I will turn to face North` could be perceived as a mis-hearing or misunderstanding of the original instruction. Furthermore, this preparatory step is only needed in a physical context where the robot is not already facing North. Thus, like the situational TRANSLATIONS, this association of the natural language description of the execution of this instruction is only valid in a particular physical context. Marking these cases distinctly again allows us to pinpoint communications that rely upon context to be appropriate.

3.2.2 Doing - Preparation

The second added subtype of acknowledgments DOING - PREP is used to mark acknowledgments that a preparatory step consistent with compliance with the previous command is underway. For example, in Table 4, the instruction *Take a picture of the wall on your left* requires that the robot first move to the left wall (line 4) in order to take an appropriate picture. Like the acknowledgment WILL COMPLY - PREP, the feedback `moving` (line 6) could be perceived as evidence of a misunderstanding, since the original instruction does not mention any motion at all, and is instead focused on taking a picture. Again, this feedback is also only an appropriate natural language description of the execution specification given the specific physical context that requires the preparatory action.

4 Corpus & Annotation

We apply the updated dialogue annotation schema to a total of 168 dialogues collected from 56 Commanders, using the data collection procedure described in Section 2, however, using a virtual environment and robot.[6] Each Commander participates in three trials, corresponding to a different search-and-navigation task, which each lasts about 20 minutes. The spoken input of the Commander and spoken feedback of the RN were transcribed and time-aligned with the text chat messages of the DM. The aligned streams are compiled into a spreadsheet with rows and columns corresponding to the examples shown here in Tables 1, 3 and 4.

[6]This data can be released via a data-sharing agreement.

Relation	#	%
Translation-r	8556	
Direct	6017	70
Direct-partial	123	1
Contextual	163	2
Contextual-partial	47	<1
Landmark	766	9
Landmark-partial	67	<1
Situated	708	8
Situated-partial	201	2
History	200	2
History-partial	4	<1
Default	251	3
Default-partial	9	<1
Updated Ack Types	5573	
Will-comply	2092	38
Doing	3379	61
Will-comply-prep	27	<1
Doing-prep	75	1

Table 5: Frequencies and % of updated relations.

Annotations are added to the spreadsheet by one of a pool of undergraduate and graduate-level annotators with backgrounds in linguistics or computer science. All annotations are then validated by one of the senior project members.

4.1 Corpus Analysis

Across the entire corpus of 168 dialogues, there are 40,873 relations, and the most prevalent general relation types are ACKNOWLEDGEMENTS, making up 36.4% of corpus relations, and TRANSLATIONS, making up 36.5% of relations, with TRANSLATION-R comprising 20.9% of the corpus and TRANSLATION-L comprising 15.6%. Thus, the general relation types we update have a large impact on the corpus.

The frequencies of the extended relation types (and directly relevant original relations) are summarized in Table 5. DIRECT TRANSLATIONS, which do not draw upon any contextual information, are the majority (70%) of TRANSLATION-R. Thus, TRANSLATIONS that do draw on contextual information make up the remaining third of the corpus TRANSLATION-R relations, with LANDMARK and SITUATED TRANSLATIONS accounting for the largest percentages of 9% and 8%, respectively.

The updates to the ACK relations have a smaller impact on the corpus, as the new types make up only about 1.8% of the ACKS considered in this paper. However, we note that these complex cases

where the new PREP ACKS apply may still be prevalent enough to be problematic in training data for a dialogue system if they are not marked distinctly, and they can now be confidently separated out from potential noise or errors in the data where the wrong acknowledgment is mistakenly given, or the input instructions are genuinely misunderstood.

4.2 Inter-Annotator Agreement

Following the same procedure as Traum et al. (2018), we compute IAA on the three markables in the annotation schema: antecedents, relations, and transaction units (TUs). Three expert coders annotated a subset of 3 dialogues (a total of 896 utterances) using our extended schema. Results appear in Table 6, which also shows the reported IAA from the unmodified schema. Note that in the unmodified schema, two rounds of IAA were conducted, the first round on 3 dialogues of 482 utterances using 5 coders, and the second round on a single dialogue of 314 utterances using 6 coders. We compare this range of IAA from the four trials of the unmodified schema, to the range of IAA for the three trials annotated with the new schema.

Markable Type	Agreement		Distance Metric
	Unmodified Schema	Modified Schema	
Antecedents	0.72–0.82	0.79-**0.94**	Nominal[a]
Relation Types	0.77–0.89	0.83-**0.93**	Nominal[a]
Transaction Units	0.48–**0.93**	0.65-0.85	MASI[b]

[a]Krippendorff (1980) [b]Passonneau (2006)

Table 6: IAA of the original, unmodified schema of Traum et al. (2018) and our modified schema.

Our modified schema yields comparable or higher IAA than the original schema for antecedents (maximum 0.94) and relation types (maximum 0.93). Our TU IAA (maximum 0.85) is higher than the range of TU IAA reported for the first round of annotations with the unmodified schema (0.48-0.70), but the final round of TU annotation from in the unmodified schema achieves the highest agreement rate of 0.93. Note that our modified schema adapts the same coding for antecedents and TUs. Thus, although one might expect that adding annotation categories would lead to lower IAA, the addition of our new subtype relations did not produce significantly lower agreement scores, demonstrating that the new annotation categories are clearly identifiable.

5 Related Work

Speech and dialogue acts have been used as part of the meaning representation of task-oriented dialogue systems since the 1970s (Bruce, 1975; Cohen and Perrault, 1979; Allen and Perrault, 1980). For a summary of some of the earlier work in this area, see Traum (1999). Although the refinement and extension of Austin's (1962) hypothesized speech acts by Searle (1969) remains a canonical work on this topic, there have since been a number of widely used speech act taxonomies that differ from or augment this work, including an ISO standard (Bunt et al., 2012). Nevertheless, these taxonomies often have to be fine-tuned to the domain of interest to be fully useful.

With the aim of developing dialogue systems, Narayan-Chen et al. (2019) propose a dialogue act schema that is somewhat more limited than Traum et al. (2018), in order to annotate dialogue focused on a collaborative building task in the Minecraft gaming environment. Bonn et al. (2020) further annotate the Minecraft corpus with Abstract Meaning Representation (AMR) (Banarescu et al., 2013) that has been updated with more detailed spatial relations. Bonial et al. (2020) also propose an annotation schema that combines both illocutionary force and propositional content into an augmented version of AMR and use this to annotate a sample of the same human-robot dialogue dataset described in Traum et al. (2018). We plan to explore the contrasts and complementarity of these annotation schemas that have been used to annotate task-oriented dialogue.

6 Conclusions & Future Work

We extend the annotation schema presented in Traum et al. (2018) so that it now uniquely marks where the language requires grounding in the physical or conversational context. While much more work is needed to provide a schema capable of training a system on how it should relate language to context, our extensions take the first critical step towards such exploration, while also separating out the training data that is largely applicable in any context. We demonstrate that the new categories introduced, which all mark up distinct features of situated language, are clearly discernible to human annotators through IAA that remains high. We are optimistic that these extensions will improve performance of dialogue systems trained on this data, which we are currently implementing.

References

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178.

John Langshaw Austin. 1962. *How to Do Things with Words*. Harvard University Press and Oxford University Press.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Claire Bonial, Matthew Marge, Ron Artstein, Ashley Foots, Felix Gervits, Cory J. Hayes, Cassidy Henry, Susan G. Hill, Anton Leuski, Stephanie M. Lukin, Pooja Moolchandani, Kimberly A. Pollard, David Traum, and Clare R. Voss. 2017. Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue. In *AAAI Fall Symposium*.

Claire N Bonial, Mitchell Abrams, David Traum, and Clare R Voss. 2021. Builder, we have done it: Evaluating & extending dialogue-AMR NLU pipeline for two collaborative domains. *Proceedings of the the 14th International Conference on Computational Semantics*.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Bertram C. Bruce. 1975. Generation as a social action. In *Theoretical Issues in Natural Language Processing*, pages 64–67.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Belis-Popescu, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Philip R Cohen and C Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive science*, 3(3):177–212.

Felix Gervits, Anton Leuski, Claire Bonial, Carla Gordon, and David Traum. 2021. A classification-based approach to automating human-robot dialogue. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 115–127. Springer Singapore.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA.

Stephanie M. Lukin, Felix Gervits, Cory J. Hayes, Anton Leuski, Pooja Moolchandani, John G. Rogers, III, Carlos Sanchez Amaro, Matthew Marge, Clare R. Voss, and David Traum. 2018. ScoutBot: A Dialogue System for Collaborative Navigation. In *Proceedings of ACL 2018, System Demonstrations*, pages 93–98, Melbourne, Australia.

Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. 2016. Applying the Wizard-of-Oz technique to multimodal human-robot dialogue. In *RO-MAN 2016: IEEE International Symposium on Robot and Human Interactive Communication*.

Matthew Marge, Claire Bonial, Ashley Foots, Cory Hayes, Cassidy Henry, Kimberly Pollard, Ron Artstein, Clare Voss, and David Traum. 2017. Exploring variation of natural human commands to a robot in a collaborative navigation task. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 58–66.

Bilyana Martinovski, David Traum, Susan Robinson, and Saurabh Garg. 2003. Functions and patterns of speaker and addressee identifications in distributed complex organizational tasks over radio. In *Diabruck: seventh workshop on semantics and pragmatics of dialogue*. Citeseer.

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

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. of LREC*.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of iso 24617-8. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 80–92.

Laurel Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 1(1).

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

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 104–111, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Traum. 1999. Speech acts for dialogue agents. In Anand Rao and Michael Wooldridge, editors, *Foundations of Rational Agency*, pages 169–201. Kluwer.

Speaker Intimacy in Chat-Talks: Analysis and Recognition based on Verbal and Non-Verbal Information

Yuya Chiba[1], **Yoshihiro Yamazaki**[2], **Akinori Ito**[2]
[1]NTT Communication Science Laboratories, Japan
[2]Graduate School of Engineering, Tohoku University, Japan
{yuuya.chiba.ax@hco.ntt.co,yoshihiro.yamazaki.t2@dc.tohoku.ac,
aito@spcom.ecei.tohoku.ac}.jp

Abstract

Conversations based on mutual intimacy are critical for maintaining positive relationships. Conversational AIs, which are widely spread in society, are assumed to be continuously used by users in daily life. To establish long-term relationships with users, AI systems have to handle dialogues based on an awareness of user intimacy. In this study, we experimentally examined a method to estimate a speaker's intimacy to a dialogue partner in chat-talks. We used a multimodal human-human conversation corpus of 71 Japanese participants. The corpus contains metadata related to subjective intimacy score of speakers. First, we identified the effective features to estimate the speaker's intimacy by comparing the statistical parameters of the features. Then, we proposed a model to estimate the speaker's intimacy by observing the several utterances.

1 Introduction

Conversational AIs, represented by the smart speakers, are widely used in daily life. Such systems are assumed to be continuously used by users, and strategies for maintaining and developing long-term relationships with users is becoming more important. However, current dialogue systems cannot take strategy to maintain a friendly relationship with users. In this situation, the system sometimes discourages the users by responding to them disinterestedly even they talk in a friendly manner.

To establish a long-term relationship, key roles include a sense of closeness and intimacy resulting from social conversations (Bickmore et al., 2005; Cassell and Bickmore, 2003). In human-human conversations, participants express intimacy with dialogue partners by such behaviors as speaking style, facial expressions, and posture (Hornstein, 1985; Planalp, 1993). Therefore, sociable conversational agents are required to manage verbal and non-verbal behavior to build a friendly relationship with users. There is also a lot of discussion about building relationships between humans. For example, the social penetration theory (Altman and Taylor, 1973) and Knapp's relationship model (Knapp et al., 2014) explained the development of relationships as a mutual process that gradually progresses.

In human-machine conversation, a few studies have developed systems that convey intimacy to users, and effectively improve user impressions. For example, Bickmore et al. (2005) built a relational agent that introduced "immediacy" behavior (Argyle, 1988) that supports multiple interactions with users over an extended period of time. Kageyama et al. (2018) evaluated dialogue systems by changing speech styles. Similarly, Kanda et al. (2009) developed for a shopping mall guide robot that changes its behavior. Kim et al. (2013) confirmed that systems that greet the users by name are perceived as friendly. However, these above studies assumed that relationships develop in association with a particular number of accumulated interactions and that they are unilaterally replicated. To achieve dialogue management that reflects a development of relationship, a system should show intimacy to users, and simultaneously recognize intimacy from them.

In this paper, we propose a method to estimate the level of intimacy of speakers to achieve systems that engage in conversations based on mutual intimacy. We target chat-talks because such conversations play an important role in establish interpersonal relationships (Rich, 1979). In human-human dialogues, dialogue behaviors based on intimacy have been investigated by analysis based on annotation. However, it remains unclear whether such information can be extracted as features from audio-visual signals. Therefore, we first identify the features that are useful for

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

estimating speaker intimacy by comparing statistical parameters of them. Then, we constructed a speaker intimacy estimation model using multimodal information. Our proposed model discriminates among tree levels of speaker intimacy by observing several utterances.

2 Related Studies

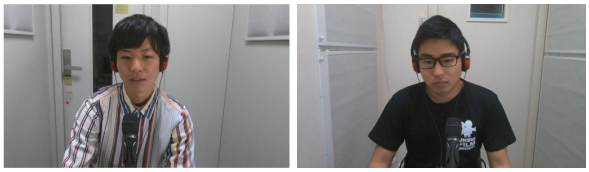
2.1 Dialogue Behavior based on Intimacy

Behaviors related to interpersonal relationships has been discussed from various perspectives. According to social penetration theory (Altman and Taylor, 1973), self-disclosure, which intentionally reveals personal information, becomes more frequent and deeper as relationships develop. Hornstein (1985) and Yamazaki et al. (2020) reported that the choice of speech intention is affected by relationships. Hall (1963) explained that the attitudes of participants change based on their interpersonal relationships. Mutual imitation is also considered to be an expression of friendship and preference. The entrainment of acoustic and prosodic features is correlated with a rapport between speakers (Lubold and Pon-Barry, 2014). The chameleon effect (Chartrand and Bargh, 1999), which is the mimicry of facial expressions and posture, is a similar phenomenon.

In addition, some studies have analyzed dialogue behaviors by focusing on such relationship stages, as friend, acquaintance, and confidant. For example, Hornstein (1985) concluded that friends use more implicit openings, raise topics, and express more responsiveness to each another by asking questions. The floor time distribution or the number of interruptions (Planalp, 1993) and various activities (Rands and Levinger, 1979) also changes according to a step of the relationships. In terms of rapport, Grahe and Bernieri (1999) reported that participants are likely to sustain longer eye contact, smile more, and lean more toward each another when building rapport.

The analysis described by these studies is based on self-reports or human annotation. It is not clear that such information can be extracted from audio-visual signals as effective features for intimacy estimation. In this paper, we investigated effectiveness of multimodal features by comparing the statistical parameters among levels of speaker's intimacy.

High intimacy



Low intimacy

Figure 1: Examples of high and low intimacy dialogue in SMOC.

2.2 Interpersonal Relationship Recognition

The estimation of interpersonal relationships has been examined by several media. Zhang et al. (2018) predicted interpersonal relationships between people in images based on facial expressions. Chu et al. (2015) proposed an immediacy prediction model using posture-based features. User profiles and sentence lengths are effective cues in SNS and e-mail interactions (Nishihara and Sunayama, 2009; Xiong et al., 2016). In human-robot conversations, Kanda and Ishiguro (2004) estimated interpersonal relationships between the participants based on interaction time. Although these studies focused on interpersonal relationship, they did not use conversational information. We are building an estimation model that is useful in various conversational situations by incorporating the aforementioned conversational information.

The most relevant research to our study is Soleymani et al. (2019), which estimated the intimacy levels of verbal self-disclosure in interview dialogues using multimodal information. In contrast, we focus on speaker intimacy in chat-talks, and do not limit the target to self-disclosure. In addition, our model is designed to incorporate such interaction between speakers as entrainment and synchrony.

3 Spontaneous Multimodal One-on-One Chat-Talk Corpus

3.1 Overview of Corpus

We used a Spontaneous Multimodal One-on-one Chat-talk (SMOC) corpus (Yamazaki et al., 2020) for the experiments. The target corpus contains the

Table 1: Summary of experimental data of this paper.

No. dialogues	345
No. recordings	690
No. pairs	69
No. speakers	71 (females: 19, males: 52)
No. utterances	23,379

audio and video of one-on-one dialogues between Japanese participants. The dialogues were conducted through by video communication between sound-proof chambers at close distance without a time lag. The audio data have no crosstalk and the video was recorded from the front of the speaker. Figure 1 shows examples of dialogue scenes between speakers with high and low intimacy.

3.2 Recording Conditions

Two participants were paired up and engaged in chat-talks. The dialogues were conducted by both acquainted and unacquainted pairs. The participants engaged in the dialogues to build a relationship with their partners. Each participant talked about five topics with two different speakers. One of the examples of topics is "My favorite foods and beverages, and the ones I don't like." Each topic lasted about 20 minutes. The speech was recorded by microphones (AT4055), and the facial expressions and gestures were recorded by video cameras (GoPro HERO7 Black) in front of the speakers. The captured video and audio data were send to the display and headphones in another sound-proof chamber through video connection for multimodal communication. The speakers talked with the partner while looking each other through the monitor. The audio data were stored with 16 kHz sampling and 16-bit quantization. The video data were recorded at 1920×1080 resolution and in a 59.94-fps MP4 format. The dialogue data have transcriptions with time-information of the beginning and ending of the utterance determined by phoneme-alignment.

We used 345 dialogues of 69 pairs recorded at an early stage of corpus construction. The total number of the dialogue recordings was 690 (345 dialogues × 2 participants). We summarized the number of the data in Table 1. The data were split into utterances based on time information.

3.3 Labels of Subjective Intimacy

The corpus has the metadata about speaker's intimacy with his/her dialogue partner. Before the conversation, each participant was asked the following questions: 1) Do you know your dialogue partner?, 2) How long have you known him/her, and 3) How close do you feel to your dialogue partner? The second and third questions were only answered by the participants who answered "yes" to the question 1). For the third question, the participants rated intimacy on a 5-grade scale, from one (not at all) to five (very much).

In this paper, we used the answer of the third question for the labels of subjective intimacy to his/her dialogue partner. The intimacy score of the participants of unacquainted pairs was set to 0. The number of dialogue recordings of score 0 was 280. Among the acquainted pairs, the numbers of the dialogue recordings rated three, four, and five were 100, 130, and 180, respectively. No participant rated less than two.

4 Analysis of Multimodal Features based on Intimacy

In this section, we analyzed the SMOC corpus based on subjective intimacy to identify the effective features for estimation. We extracted linguistic, acoustic, and visual features.

4.1 Word Frequency Distribution

First, we compared the word frequency distribution between different subjective intimacy scores. The utterances were segmented using MeCab[1] (Kudo, 2006), which is a Japanese morphological analyzer, with the NEologd dictionary[2]. We constructed Bag-of-Words (BoW) vectors for each score, and visualized the distance between them by multidimensional scaling. The result is shown in Figure 2. The figure shows the distribution of words of each score were roughly separated into three clusters: a group of scores 4 and 5, score 3, and score 0.

One reason why word frequency distribution is different between the groups is the influence of "honorifics." The target corpus's language, Japanese, has a clear honorific mechanism. The speech style changes based on the relative social position or closeness of the social distance to the dialog partners. In Japanese, the honorifics is often expressed by the auxiliary verb of the ending of the utterance. For example, the verb *taberu* (to eat) can be transformed to *tabe-masu* to express the honorifics. Here, we focused on *desu* and

[1] http://mecab.sourceforge.jp

[2] https://github.com/neologd/mecab-ipadic-neologd

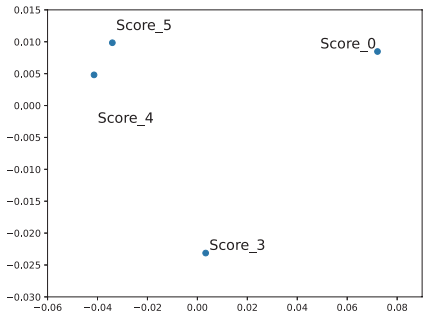


Figure 2: Visualization of distance between BOW vectors of subjective intimacy scores based on multidimensional scaling.

Table 2: Difference of average frequency of DAs relating "Question" between intimacy levels ($^*p < 0.05$).

Category	Comparison	Diff.	t	p-value
Information	Low − Mid.	0.20	1.09	0.82
	Low − High	−0.30	−2.32	0.06
	Mid. − High	−0.50	−2.77	0.02*
Fact	Low−Mid.	0.54	2.36	0.06
	Low−High	0.46	2.80	0.02*
	Mid.−High	−0.09	−0.38	1.00
Experience	Low−Mid.	0.10	2.59	0.03*
	Low−High	0.12	4.07	< 0.00*
	Mid.−High	0.01	0.29	1.00
Habit	Low−Mid.	0.08	2.65	0.03*
	Low−High	0.12	5.29	< 0.00*
	Mid.−High	0.03	1.11	0.81
Desire	Low−Mid.	0.01	0.43	1.00
	Low−High	0.05	2.52	0.04*
	Mid.−High	0.03	1.37	0.51

masu, which are among the most basic auxiliary verbs, to express honorifics in Japanese. When we investigated the word usage rate of the target corpus, the order of the use of these auxiliary verbs decreased monotonically as the intimacy scores increased. Such a lexical change does not necessarily exist only in Japanese. Some languages change how to address someone to convey attitudes toward the addressee (i.e., T-forms and V-forms (Brown and Gilman, 1960)). Therefore, lexical features are effective features for estimation even in other languages.

Based on the above analysis, we separated the dialogue behavior of the participants into three classes: score 0 as "low intimacy," score 3 as "middle intimacy," and scores 4 and 5 as "high intimacy."

4.2 Dialogue Acts

Next, we compared the average frequency of the dialogue acts (DA) among levels of intimacy. DAs were extracted using Richindexer[3]. The kind of the DAs is the same with Meguro et al. (2010). First, we conducted a one-way layout ANOVA that factored the level of intimacy. Then, we conducted multiple comparison tests with Bonferroni correction for each DA that showed a significant difference by ANOVA. In this paper, we focus on DAs related "Question" due to space limitation. We summarized the results for sub-categories of "Question" in Table 2.

The table showed that such questions as the facts, experience, and habits frequently appeared in the group of low intimacy. It is adequate that these questions tend to appear at the early stage of relationships when the participants are getting to know each other. The trends of the differences varied by sub-categories, although the total number of questions decreased. These results coincide with a conventional study (Yamazaki et al., 2020), although they partially contradict another report (Hornstein, 1985). Hornstein (1985) concluded that friends were responsive to the partner by asking more questions. One possible reason of the difference is cultural differences. Although verification of the cultural difference is not a purpose of this paper, we plan to compare the DAs using other language corpora in future studies.

4.3 Entrainment of Prosody

In the analysis of acoustic features, we focused on the interaction between speakers. We extracted the prosodic features, the speaking rate, and the switching pauses from the utterances and calculated the entrainment. For the prosodic features, the maximum and the mean of the log F0 and intensity were selected based on a previous study (Kawahara et al., 2015). Here, Levitan and Hirschberg (2011) proposed a quantification method for three kinds of entrainment: proximity, convergence, and synchrony. We focused on the proximity of the acoustic features. Turn-level proximity evaluates how close the utterance's acoustic feature is to that of the preceding interlocutor's utterance. Concretely, the proximity is expressed by the absolute difference of the average of the feature between adjacent utterances. In the analysis, we compared the average proximity over dialogue among intimacy levels.

First, we conducted a one-way layout ANOVA that factored the level of the intimacy and then

[3]https://www.rd.ntt/e/research/MD0057.html

Table 3: Difference of average proximity of prosodic features: A positive difference indicates that the right group has the large entrainment ($^*p < 0.05$).

Features	Comparison	Diff.	t	p-value
Speaking rate	Low−Mid.	1.50	1.65	0.30
	Low−High	1.78	2.77	0.02^*
	Mid.−High	0.28	0.32	1.00
Switching pause	Low−Mid.	0.03	0.78	1.00
	Low−High	0.04	4.38	$< 0.00^*$
	Mid.−High	0.00	2.39	0.05
max. f0	Low−Mid.	−0.01	−1.83	0.21
	Low−High	−0.01	−2.51	0.04^*
	Mid.−High	0.00	−0.05	1.00
mean f0	Low−Mid.	−0.01	−1.48	0.42
	Low−High	−0.01	−3.66	$< 0.00^*$
	Mid.−High	−0.01	1.12	0.79

a multiple comparison test with Bonferroni correction. Table 3 shows the results of the multiple comparison tests for the features that obtained significant differences by ANOVA. The speaking rate and switching pause were significantly different between low and high intimacies, indicating that speakers who feel more intimacy to the interlocutor tend to synchronize their speaking rates and switching pauses. In terms of log F0, the entrainment was larger in the group of the low-level intimacy against our expectations. The global features calculated from the entire utterance may be too coarse to capture the entrainment. In future studies, we will calculate the entrainment using the features obtained from the beginning and ending segments of the utterance as same with the previous study (Kawahara et al., 2015).

4.4 Facial Expression Synchrony

For visual cues, the features expressed during the dialogue partner talking to are also important. Thus, we focused on the synchrony of facial expressions. Facial Action Units (AU) were extracted using OpenFace (Baltrušaitis et al.). Levitan and Hirschberg (2011) quantified the synchrony by the correlation coefficient between the features of interlocutors. Here, the intimacy scores of the target corpus differ from speaker by speaker. The correlation coefficient cannot be calculated by a three-level classification because the scores may be different between speakers of the same dialogue. Therefore, we compared the unacquainted group (U) and acquainted group (A) (i.e., score 0 and others).

Table 4 shows the results of a Welch's t test between the two groups when we compared the average synchrony over dialogue. AU02 (Outer Brow Raiser), AU06 (Cheek Raiser), and AU12 (Lip Corner Puller) were significantly larger in the acquainted group. In particular, the AU06 and AU12 features increase when the speaker expresses a smile. Therefore, it is indicated that the smile tends to co-occur in acquainted groups. In contrast, no significant differences were observed in other AUs since the other facial expressions are less likely to appear in the target dialogue.

Table 4: Difference of average synchrony of action units: U and A represent unacquainted and acquainted pairs. A negative difference indicates that the acquainted pairs have the large entrainment ($^*p < 0.05$).

Action Unit	$U - A$	t	p-value
AU01	-1.493×10^{-2}	−1.319	0.188
AU02	-3.386×10^{-2}	−2.501	0.013^*
AU04	0.578×10^{-2}	0.496	0.621
AU05	-0.217×10^{-2}	−0.181	0.857
AU06	-9.453×10^{-2}	−5.884	$< 0.000^*$
AU07	-2.107×10^{-2}	−1.560	0.120
AU09	0.058×10^{-2}	0.036	0.971
AU10	-1.275×10^{-2}	−0.830	0.407
AU12	-8.044×10^{-2}	−4.888	$< 0.000^*$
AU14	-2.044×10^{-2}	−1.192	0.234
AU15	0.383×10^{-2}	0.336	0.737
AU17	-1.279×10^{-2}	−0.994	0.321
AU20	-0.415×10^{-2}	−0.440	0.660
AU23	-2.133×10^{-2}	−1.870	0.062
AU25	2.197×10^{-2}	1.353	0.177
AU26	-0.163×10^{-2}	−0.126	0.900
AU28	1.382×10^{-2}	1.424	0.158
AU45	0.779×10^{-2}	0.680	0.497

4.5 Gaze Activity

Finally, we investigated the gaze actions. For gaze, it is reported that the participants with high rapport are likely to sustain eye contact longer Grahe and Bernieri (1999). Therefore, we focused on gaze variations. Let the gaze angle at time t be $\mathbf{g}_t = (x_t, y_t)$, the gaze variation is represented as:

$$\Delta \mathbf{g} = \frac{1}{T-1} \sum_{t=2}^{T} ||\mathbf{g}_t - \mathbf{g}_{t-1}||. \quad (1)$$

x_t and y_t are the gaze angles of the horizontal and vertical axes obtained using OpenFace. T is the number of frames of each dialogue. We extracted the gaze variation from every utterance, and compared the average gaze variation over dialogue.

First, we conducted a one-way layout ANOVA that factored the level of intimacy and obtained a significant difference ($p < 0.001$). Then, a multiple comparison test with Bonferroni correction was conducted. Table 5 shows the results. Significant differences were observed between "Low"

Table 5: Difference of average gaze variance: A positive difference indicates that the right group has a small variation ($^*p < 0.05$).

Comparison	Diff.	t	p-value
Low−Mid.	-0.010×10^{-3}	0.008	1.000
Low−High	1.097×10^{-3}	3.854	$< 0.001^*$
Mid.−High	1.107×10^{-3}	2.881	0.012^*

and "High," and "Mid." and "High." The gaze variation significantly decreased in the high intimacy group. This result suggests that the gaze feature is effective to estimate the speaker's intimacy.

5 Intimacy-Level Estimation Network

In the following sections, we examined an intimacy recognition method that reflects our analysis. Figure 3 shows our proposed network. As shown in the analysis, such interactions between speakers as the entrainment seem important to intimacy estimation. Therefore, the proposed intimacy recognition model takes continuous utterances as input. Here, $l_{t,n}$ and $a_{t,n}$ are the linguistic and acoustic features at time n of the t-th utterance. In addition, $v^i_{t,n}$ is the visual feature of speaker $i \in (s, p)$. s and p represent the speaker and the dialogue partner of respective utterances. N^l_t, N^a_t, and N^v_t are the length of the linguistic, acoustic, and visual feature sequences. y_t is the prediction result.

First, the network extracts the verbal and non-verbal features every utterance and encode them to the representation vectors. We employed the multi-stream attention-based BLSTM (Chiba et al., 2020) as an utterance-level encoder. In this method, the feature sequence of the respective modality is input to the individual attention-based BLSTM (Mirsamadi et al., 2017). Then, the concatenation of the representation vectors of each modality is sent to the linear layer. From these processes, the multi-stream BLSTM fuses the utterance-level multimodal information. For the visual features, it is important to represent the correlation of features (i.e., synchrony) between both speakers. Therefore, we feed the visual features of both speakers to the network. The fully-connected layer is connected after the input layers of the speaker's and the partner's visual features for dimensional reduction.

In addition, our analysis showed that the entrainment between the preceding and current utterances is important for the acoustic features. Thus, we used BLSTM for the succeeding layers of the utterance encoder to capture the relationship between utterances. The context BLSTM takes the representation vectors of continuous T utterances as input, and its output is input to the single fully-connected layer to obtain the prediction result.

5.1 Feature Extraction

The network takes the word sequence as input for the linguistic feature. Each word was converted to a 300-dimensional embedded vector using FastText (Joulin et al., 2016). As acoustic features, the eGeMAPS (Eyben et al., 2015) were extracted with 10-ms frame-shift and 20-ms frame-width. The eGeMAPS features include not only such prosodic features as pitch and loudness but also spectral features. We used the 46-dimensional features, including the Δ features. On the other hand, we used OpenFace (Baltrušaitis et al.) to extract the visual features. We used the features relating to AU, gaze direction, and face direction for the experiments. As same with the acoustic features, the Δ features were calculated for the visual features. The number of dimensions of the visual features was 80. For the visual features, the features of both speakers at the target utterance segment were extracted.

Analysis suggests that the speaking rate or gaze variance were effective for recognition. Therefore, we employed segmental features for the acoustic and visual features to enhance such information. Segmental features were obtained by calculating the statistics (e.g., mean, variance, and range) of the above features every 200 ms. Such statistics are useful to explain speaking rate and gaze variance at the local segment. We employed 12 kinds of statistics same with the Schuller et al. (2009). The number of dimensions of the definitive features were 552 for the acoustic features ($46 \times 12 = 552$) and 960 for the visual features per speaker ($80 \times 12 = 960$).

6 Experiments

6.1 Setup of Experimental Data

As in the case of analysis, we used the SMOC corpus for the recognition experiment as well. We separated the utterances of 345 dialogues into training, development, and test sets so that any of two sets do not share the same speaker. The training, development, and evaluation data were 16,314, 3,465, and 3,600 utterances, respectively.

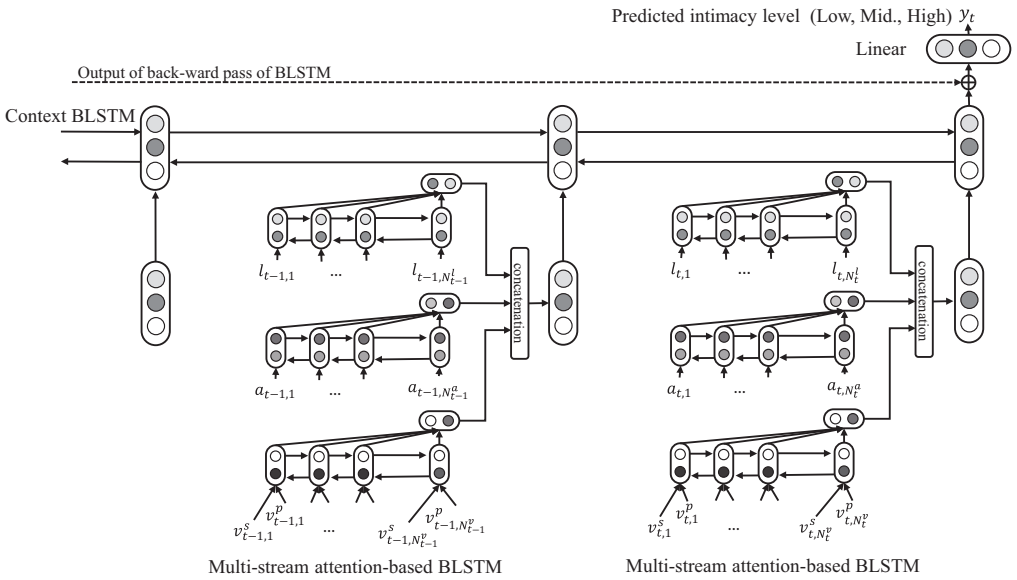


Figure 3: Network architecture for intimacy recognition: $l_{t,n}$ and $a_{t,n}$ are linguistic and acoustic features at frame n of t-th utterance. $v^i_{t,n}$ is the visual features of participant $i \in (s, p)$. s and p represent the speaker and the partner, respectively. N^l_t, N^a_t, and N^v_t are sequence length of linguistic, acoustic, and visual features. y_t is prediction result. $\oplus$ shows the summation.

The intimacy labels of the SMOC corpus were appended to the dialogues. We assigned the same label as the original dialogue to the utterances.

6.2 Conditions of Training Network

We conducted three-class discrimination among low, middle, and high levels of intimacy. The experimental data have a bias toward the distribution of intimacy levels, and we employed weighted cross-entropy loss to train the network. The losses in each class were multiplied by a weight that is proportional to the inverse of the sample size.

The numbers of hidden units were common among the multi-stream BLSTM, the context BLSTM, and fully-connected layers. We investigated the classification performance while changing the number of nodes of the hidden layers to 16, 32, 64, and 128. We used the condition that yielded the best accuracy for the validation set for the definitive evaluation. The number of layers of each BLSTM was 1. We connected the dropout layers after the output of each stream and the context BLSTM. The dropout rate was set to 0.3. The optimization method was Adam with a learning rate of 0.0005. The mini-batch size was 32 and the maximum number of epochs was 100. In the following sections, we show the recognition results for the test set.

7 Experimental Results of Intimacy Recognition

First, we evaluated the effectiveness of the multimodal features. In this experiments, we used continuous four utterances for the classification (i.e., $T = 4$). Table 6 shows the recognition results. A, V, and L denote the acoustic, visual, and linguistic features, respectively. Rec., Pre., and F1. represent the recall, the precision, and the F1-score. Chance shows the results when all test samples are classified to high-level intimacy, which is the most frequent class.

As shown in the table, the results of the proposed models surpassed the chance-level results. The results indicated that the model was adequately trained to estimate the level of intimacy from verbal and non-verbal cues. Comparison of the single modality showed that a higher F1-score was obtained with linguistic information. This result suggests that the utterance styles and the choices of the DA were captured using linguistic features. The combination of audio, visual, and linguistic features improved the performance, and we obtained an F1-score of 0.594. Therefore, the non-verbal information employed in this study was an effective feature to enhance verbal information.

However, the performance of the acoustic and visual information alone did not surpass the lin-

Table 6: Intimacy Recognition Results: A, V, and L denote acoustic, visual, and linguistic features, respectively. Rec., Pre., and F1. represent the recall, precision, and F1-score. Bold fonts are the best performance between modalities. Chance shows results when all test samples are classified as high-level intimacy, which is the most frequent label.

	Low			Middle			High			Macro Average		
Modality	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
A	0.313	0.739	0.439	0.399	0.354	0.375	0.565	0.257	0.353	0.425	0.450	0.389
V	0.262	0.554	0.356	0.165	0.032	0.053	0.568	0.495	0.529	0.332	0.360	0.313
L	**0.857**	0.770	**0.811**	0.272	0.196	0.228	0.652	0.762	0.703	**0.594**	0.576	0.581
A+V	0.275	0.709	0.397	0.499	0.338	0.403	0.587	0.266	0.367	0.454	0.438	0.389
A+L	0.627	**0.835**	0.716	0.469	0.314	0.376	0.671	0.672	0.672	0.589	0.607	0.588
V+L	0.759	0.791	0.775	0.258	0.115	0.159	0.652	**0.801**	**0.719**	0.557	0.569	0.551
A+V+L	0.567	0.811	0.667	**0.506**	**0.432**	**0.466**	**0.693**	0.608	0.648	0.589	**0.617**	**0.594**
Chance	–	–	–	–	–	–	–	–	–	0.177	0.333	0.231

Table 7: Influence of length of dialogue context: Table shows F1-score of classification.

No. utterances N	Low	Mid.	High	Average
2	0.539	0.370	0.556	0.488
4	0.668	0.466	0.648	0.594
6	**0.736**	**0.467**	**0.730**	**0.645**
8	0.716	0.416	0.715	0.616

guistic feature. In particular, the visual information had the lowest estimation performance. This result reflected that the features extracted from OpenFace are insufficient to comprehensively represent non-verbal behavior. Posture and gesture are cues that predict rapport (Grahe and Bernieri, 1999), and we will examine the effectiveness of them for intimacy-level estimation in future studies.

Next, we examined the influence of context length. Table 7 shows F1-scores when changing the length of the dialogue context. As shown in the table, performance improved with a longer dialogue context, and we obtained the best performance at $N = 6$. It is confirmed that our proposed model can estimate the speaker's intimacy to some extent by observing three utterance interchanges. Since the labels were originally assigned to each dialogue, it is considered to be appropriate that the long dialogue context is effective to estimate the speaker's intimacy. On the other hand, performance decreased when the number of interchanges exceeded three (i.e., $T = 8$). One reason for this result is the dialogue data is insufficient. In particular, the data size of the middle-level intimacy was relatively small, and the F1-score did not improve even the network observes the longer context. Therefore, the dialogue data of acquainted pairs that are not close friends should be collected in future studies.

8 Summary and Future Studies

In this paper, we examined the recognition method of speaker intimacy in chat-talks. First, we identified the effective verbal and non-verbal features to estimate subjective intimacy-levels. Then, we developed an intimacy-level estimation model that reflected the analysis results. Our proposed model discriminated user intimacy among low, middle, and high levels. From experiments, we obtained the best F1-score of 0.645 when using the acoustic, visual, and linguistic features. However, some remaining issues must be solved to apply our proposed method to actual dialogue systems.

First, the data used in this study are human-human dialogues, and the behavior of participants might be different in human-machine dialogues. One possible solution is model adaptation. Our proposed network can be adapted to the human-machine dialogues by fine-tuning. In near future, we plan to collect human-machine dialogues based on the wizard-of-Oz basis. Besides, there is a class imbalance problem. In the target dataset, the data size of middle-level intimacy is relatively small, and the performance of this class did not improve. Therefore, collection of dialogue between acquaintance speakers is needed to improve the overall performance of the base model.

In addition, it is crucial that how the system behaves to recognized user intimacy to achieve a dialogue system based on mutual intimacy. Therefore, we next plan to examine a dialogue generation method combining the intimacy-estimation network with a recent response generation model (e.g., (Smith et al., 2020)).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20K19903.

References

Irwin Altman and Dalmas Taylor. 1973. *Social penetration: The development of interpersonal relationships*, volume 212. Holt, Rinehart & Winston.

Michael Argyle. 1988. *Bodily Communication*. New York: Methuen & Co, Ltd.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 1–10.

Timothy Bickmore, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *Proc. CHI*, pages 1212–1215.

Roger Brown and Albert Gilman. 1960. *The pronouns of power and solidarity*. Bobbs-Merrill.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132.

Tanya Chartrand and John Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *J. Pers. Soc. Psychol.*, 76(6):893–910.

Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition. In *Proc. INTERSPEECH*, pages 3301–3305.

Xiao Chu, Wanli Ouyang, Wei Yang, and Xiaogang Wang. 2015. Multi-task recurrent neural network for immediacy prediction. In *Proc. ICCV*, pages 3352–3360.

Florian Eyben, Klaus Scherer, Björn Schuller, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Jon Grahe and Frank Bernieri. 1999. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior*, 23(4):253–269.

Edward Hall. 1963. A system for the notation of proxemic behavior 1. *American Anthropologist*, 65(5):1003–1026.

Gail Hornstein. 1985. Intimacy in conversational style as a function of the degree of closeness between members of a dyad. *J. Pers. Soc. Psychol.*, 49(3):671–681.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText. zip: Compressing text classification models. *arXiv preprint:1612.03651*, pages 1–13.

Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito. 2018. Improving user impression in spoken dialog system with gradual speech form control. In *Proc. SIGDIAL*, pages 235–240.

Takayuki Kanda and Hiroshi Ishiguro. 2004. Friendship estimation model for social robots to understand human relationships. In *Proc. ROMAN*, pages 539–544.

Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An affective guide robot in a shopping mall. In *Proc. HRI*, pages 173–180.

Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. 2015. Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In *Proc. APSIPA-ASC*, pages 392–395.

Yunkyung Kim, Sonya Kwak, and Myung-Suk Kim. 2013. Am I acceptable to you? Effect of a robot's verbal language forms on people's social distance from robots. *Comput. Human Behav.*, 29(3):1091–1101.

Mark Knapp, Anita Vangelisti, and John Caughlin. 2014. *Interpersonal communication and human relationships*. Pearson.

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. INTERSPEECH*, pages 3081–3084.

Nichola Lubold and Heather Pon-Barry. 2014. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proc. the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12.

Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling listening-oriented dialogue using partially observable Markov decision processes. In *Proc. COLING*, pages 761–769.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proc. ICASSP*, pages 2227–2231.

Yoko Nishihara and Wataru Sunayama. 2009. Estimation of friendship and hierarchy from conversation records. *Inf. Sci.*, 179(11):1592–1598.

Sally Planalp. 1993. Friends' and acquaintances' conversations II: Coded differences. *J. Soc. Pers. Relat.*, 10(3):339–354.

Marylyn Rands and George Levinger. 1979. Implicit theories of relationship: An intergenerational study. *J. Pers. Soc. Psychol*, 37(5):645–661.

Elaine Rich. 1979. User modeling via stereotypes. *Cognitive Science*, 3(4):329–354.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Proc. INTERSPEECH*, pages 312–315.

Eric Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*, pages 1–10.

Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal analysis and estimation of intimate self-disclosure. In *Proc. ICMI*, pages 59–68.

Liyan Xiong, Yin Lei, Weichun Huang, Xiaohui Huang, and Maosheng Zhong. 2016. An estimation model for social relationship strength based on users' profiles, co-occurrence and interaction activities. *Neurocomputing*, 214:927–934.

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In *Proc. LREC*, pages 443–448.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569.

The red cup on the left: reference, coreference and attention in visual dialogue

Simon Dobnik[1,2] and **Vera Silfversparre**[1]
[1]Department of Philosophy, Linguistics and Theory of Science
[2]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
simon.dobnik@gu.se, gussilfve@student.gu.se

Abstract

We examine how conversational partners refer, co-refer and direct attention in conversations over a visual scene. Using an extension of the CoNLL annotation scheme for coreference for the visual domain we annotate the Swedish part of the Cups corpus. The annotation consists of identifying noun phrases and assigning them IDs of entities in the visual scene. We perform quantitative and qualitative linguistic analyses of the annotated data which point towards interesting observations of how conversational participants direct attention: it is likely that entities are co-referred to within the same conversational game, for spatial descriptions there is a preference for lateral dimensions over front and back and more attention is directed towards entities that are visually ambiguous or those that are part of the task. Overall, we demonstrate that referential attention is driven by both visual and conceptual, task-related information.

1 Introduction

Visual dialogue takes place in some visual context either physical or virtual. Conversational participants discuss the visual scene but they also relate it to their beliefs, desires and intentions as defined by the task they are engaged in. An important challenge for building visual dialogue systems is to model how such perceptual information interacts with higher level conceptual aspects of their information state in order to generate and interpret referring expressions such as "the red cup on the left" in the setting of a collaborative dialogue (Clark and Wilkes-Gibbs, 1986; Byron, 2003). In this paper we examine referring expressions in visual dialogue, in particular the mechanisms behind how speakers and hearers generate and interpret them in a highly visually and linguistically ambiguous environment. Conversational partners must rely on mechanisms of *attention* that assigns salience to contextual information from the visual, linguistic and task-related domains which affects how referring expressions are generated and interpreted (Kelleher et al., 2005). Literature on attention (Lavie et al., 2004) distinguishes between perceptual selection, a process that selects relevant visual features, and cognitive control, a process that selects the relevant conceptual information, which compete for the same cognitive resources. Since conversational participants are engaged in a collaborative task joint attention will be aimed at.

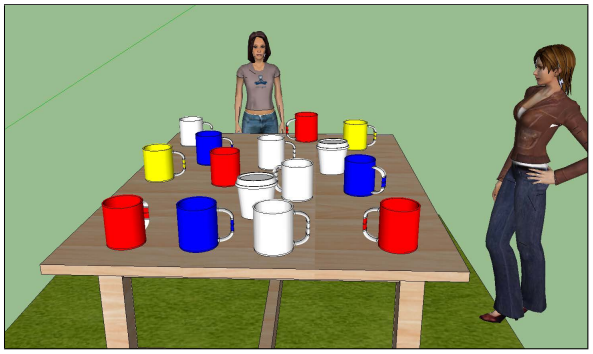
Reference and coreference resolution has been studied both in the domain of the textual documents (Sukthanker et al., 2020), in the domain of situated dialogue (Kelleher et al., 2005; Rolih, 2018; Smith et al., 2011) or in the domain of vision and language (Kottur et al., 2018; Yu et al., 2019). In the domain of textual coreference, one of the most known resources is the section of the OntoNotes corpus annotated for coreference as a part of the CoNLL-2011 shared task (Pradhan et al., 2011). Another well-known resource is the ARRAU corpus (Poesio et al., 2018; Uryupina et al., 2020). In the domain of visual dialogue the SCARE corpus (Stoia et al., 2008) contains spoken dialogues in a virtual reality maze environment with buttons, cabinets and doors. We follow the tradition of coreference annotation in the textual domain, by starting with the CoNLL 2011/2012 scheme and extending it to the domain of the visual dialogue of the Swedish part of the Cups corpus (Dobnik et al., 2015, 2020). The corpus is different from other corpora used in research on referring in that it comes with a single visual scene with a known ground truth representation of entities from which different views can be generated and over which participants can engage in long dialogues. This makes it an ideal candidate for studying coreference. Appendix A.1 shows some examples discussed here. Based on this annotation we address the following questions:

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

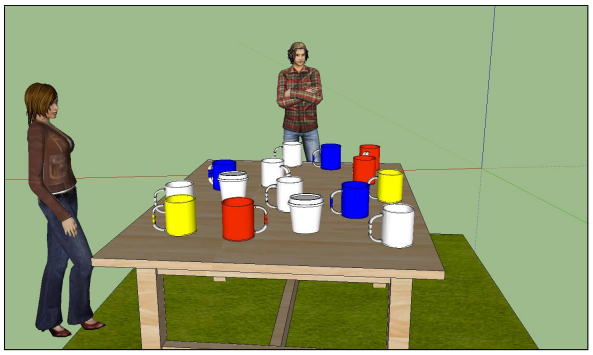
Q1: How do the interlocutors refer and co-refer to entities in the visual scene?

Q2: What are the issues with the referent annotation when starting with an annotation scheme developed for the textual domain and how can they be addressed?

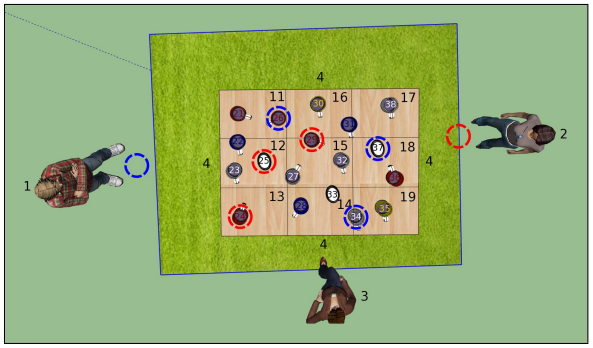
Q3: How is the attention (estimated from the reference of descriptions) distributed over the visual scene?



(a) The view of P1



(b) The view of P2



(c) Ground truth view of the scene

Figure 1: The scene as seen by P1 (a) and P2 (b). (c) shows a top-down view of the scene with all objects included and their object IDs. Objects marked with coloured circles cannot be seen by a participant marked with the same colour. P3 is a passive observer Katie.

2 Data and annotation

The Cups corpus (Dobnik et al., 2015, 2020) was created to examine collaborative dialogue over a visual scene and therefore resembles the Map Task (Anderson et al., 1991). It was previously used to study selection of reference frames, dialogue games (Storckenfeldt, 2018) and coreference (Dobnik and Loáiciga, 2019). A virtual scene containing a table with cups of different types and colours, two active conversational participants at the opposite sides of the table and a passive observer has been created in 3d-modelling software as shown in Figure 1. A static view of the scene was created for each participant. In addition, some objects were removed from the view of each participant but these were kept in the view of the other participant. The same views were used for all participant pairs. The data collection was done in a lab environment. Participants are instructed to interact over a textual computer interface in order to find and make a note of the missing cups which defined their collaborative task. To encourage spontaneous longer dialogue the task was not restricted in time. The nature of the task prevented participants from communicating through intonation, prosody, eye-gaze and body gestures. Table 1 summarises the current size of the corpus. We refer to the corpus as (sv.P05.100) which stands for the 100th turn of the P05 dyad of the Swedish sub-corpus.

Corpus	Dialogue	Turns	Native speakers of
English	en.P01	157	Swedish
	en.P02	441	English
Swedish	sv.P01	118	Swedish
	sv.P02	114	Swedish
	sv.P04	75	Swedish
	sv.P05	163	Swedish
	sv.P06	248	Swedish
	sv.P07	308	Swedish

Table 1: The Cups corpus. Here we annotate and analyse the Swedish (sv) part.

For the purposes of this study we annotated the Swedish sub-corpus (sv) with the CoNLL 2011/2012 annotation scheme (Pradhan et al., 2011) used for textual data but in contrast to OntoNotes (Pradhan et al., 2011) we annotate all noun phrases, as in the ARRAU corpus (Poesio et al., 2018). Note, however, that OntoNotes also contains annotation of coreference for verbs and temporal expressions. The annotation was done by the second author and then interesting and challenging examples were discussed with the first au-

Dlg	P01	P02	P04	P05	P06	P07	Total
RFs	197	360	278	395	463	571	2264

Table 2: The number of referring expressions in the Swedish part of the Cups corpus per dialogue.

thor. Based on this discussion, annotations were adjusted and notes were made for the annotation manual. The annotation file is automatically tokenised and then the annotation consists of two parts. First, noun phrases are identified using the BIO tags (`B-NP`, `I-NP` and `O`). Then, co-reference chains are identified over noun phrases by assigning referent IDs to them, e.g. `(11,13` for the opening word of a noun phrase and `11,13)` to the closing word of the same noun phrase while no tag is assigned to the intermediate words. In the standard textual coreference annotation (OntoNotes and ARRAU), the IDs are incremented as new referents are introduced in the text. However, in our work we pre-identify referents as entities (participants, objects and regions) identifiable in the visual scene as shown in Figure 1c. In this respect our annotation resembles the annotation of the SCARE corpus of visual dialogue (Stoia et al., 2008) where IDs are also pre-assigned to entities in the visual environment but is different from it in that we extend the assignment of IDs in two ways. For NPs that cannot be assigned a referent special tags were introduced and additional numeric tags were assigned to entities in the scene that were not previously identified in Figure 1c (see Section 3.2, (sv.P06.64-67)). Overall, in our adaptation of the CoNLL annotation scheme all noun phrases are annotated, a particular entity always has the same ID and a single noun phrase can be assigned several IDs. Referring expressions with the same IDs are coreferential. An example annotation is shown in Appendix A.2. All annotations are available at https://github.com/sdobnik/cups-corpus.

3 Results

3.1 Reference and coreference to entities (Q1)

Table 2 shows the number of referring expressions used in individual dialogues and in total in the Swedish part of the Cups corpus. The counts vary across different dialogues: for example P07 contains approximately three times as many referring expressions compared to P01. These referring expressions are assigned 3,867 references to entities in total which is 1.71 times the number of referring

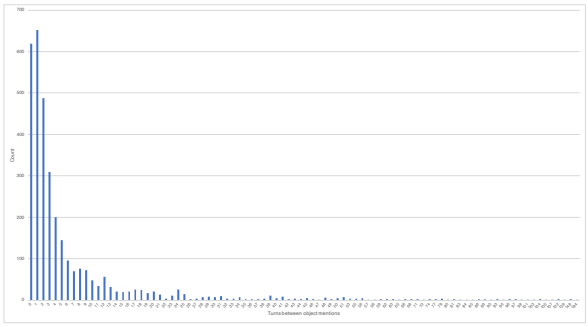


Figure 2: Latency of turns before an object is re-referred to. Latency 0 means that the objects is re-referred to in the same turn.

expressions. This indicates that very frequently an expression is referring to more than one entity. Out of 3,867 references to entities there are 3,515 references to entities with pre-defined IDs and 352 references to entities without IDs that were added dynamically by annotators. This shows that participants primarily refer to and attend to the entities that were pre-annotated in the scene which was done relative to the task and the goal of the conversation participants engage in. However, 352 (9.1%) of references to entities not pre-annotated indicates that the notion what is an entity (an object or a region) might not be straightforward. For example, the participants might refer to parts of the objects in order to disambiguate them, e.g. "sen en vit med lock, den står något närmare dig, sen en vit med handtag" (Then a white with a lid, it is somewhat closer to you, then a white with a handle.) (sv.P07.26-29). References to new objects are also frequently required when referring to regions as participants create regions and rows dynamically based on the topology of the scene and internal relations of objects rather than a global and equal geometric split of the scene.

Since conversational participants have access to the same visual scene throughout the dialogue objects or regions may be visually or linguistically attended more than once. To quantify how objects are re-referred to over the course of the dialogues we calculate the number of turns between two consecutive references to the same entity (≠mentions). Figure 2 shows that coreference ranges from 0 to 194 turns. It is very common for the object to be referred to in the same turn and then within 1 to 4 different turns. After that, the coreference to the same object decreases fast. For example, the same object is re-referred to over 20 turns less than 20 times, and 10 times over 30 turns. That coref-

erence is focused on a smaller number of turns indicates that participants collaboratively discuss and re-refer to objects until the ambiguity (both visual and conversational) is resolved (sv.P02.36-41, dialogue in Appendix A.1). The distribution of coreference also indicates that the objects might be re-referred to within the scope of the same conversational game. (Storckenfeldt, 2018, p.28) reports that the mean length of the annotated conversational games in this corpus range from 2.9 to 5.5 utterances which corresponds to the coreference figures reported here. Longer coreference could then be explained by the fact that entities are re-referred to in another conversational game. A possible reason to return to an entity is to use it as a landmark or a comparison when locating other entities. Once an entity is visually and linguistically grounded in the common ground it becomes a part of the shared knowledge and therefore a useful referential landmark (sv.P06.21-24). We expect that the usage of landmark entities also decays in time and landmarks that were more frequently referred to are preferred (Kelleher and Dobnik, 2020). As they are salient in the common ground and reference to them is not under discussion anymore, they only need to be referred to once as landmarks. This also explains a drop in frequencies after 4 turns.

3.2 Reference, coreference and visual dialogue (Q2)

In this section we examine questions related to annotating reference and co-reference in visual dialogue using the CoNLL 2011/2012 annotation scheme and suggest its required extensions.

The first question relates to the annotation of expression that are not referring to the corpus scenes and therefore cannot be assigned an object ID. Swedish also uses a demonstrative pronoun *det* as in "det finns" or "det är" which corresponds to English "there is" and "it is/they are" (sv.P06.127-129). Such pronouns are not referring and we annotate them as *expletives*. Conversational participants may refer to *entities outside the visual scene*, for example "in my picture" referring to a printed sheet of paper with a visual scene (sv.P01.61), or "byracka" referring to the other participant in a friendly derogatory way (sv.P06.4-6). Thirdly, there may be *non-referring expressions* that are used. These could be to abstract entities "in princip" (in principle, basically) or negated expressions "ingen lockmugg" (no cup with a lid) (sv.P04.51-56) but negated expressions are sometimes referential referring to an object that the other person previously referred to (sv.P04.52) or referring to an object of not being of that kind. Fourthly, *interrogative noun phrases* occuring in direct and indirect questions are also non-referential, e.g. "vad" (what) (sv.P05.145), "vilken farg" (what colour) (sv.P06.174-175). Table 3 shows the distributions of annotations of these categories in individual dialogues and in total. Expletive and non-referring expressions are most common but also note that there are considerable differences between different dialogues, e.g. there are 54 Non-R in P07 but only 4 in P04. This indicates that different conversational dyads might use different referring strategies.

Dialogue	Ext	Expl	Non-R	Wh-Q
P01	5	6	8	2
P02	6	21	13	6
P04	5	17	4	1
P05	2	25	27	7
P06	9	23	17	7
P07	13	29	54	6
Total	40	121	123	29

Table 3: The distribution of expressions not referring to objects IDs: external reference objects (Ext), expletive expressions (Expl), non-referring expressions (Non-R) and expressions used in wh-phrases (Wh-Q).

The second question relates to how to apply the existing annotation scheme on the data. Noun phrases can be complex containing embedded noun phrases of the form NP Relation NP, for example "de vita med handtag och utan lock" (the white one with handles and without lids) (sv.P04.40) and "en röd mugg på din vänsterkant" (a red cup to your left side) (sv.P05.57). Should "handles" and "your left side" also be annotated? Motivated by the research on spatial relations where two NPs are distinguished as Target and Landmark we decided to annotate each NP separately, provided that they are referring to distinctive objects and regions. However, sometimes this convention becomes hard to follow and this is related to whether the NPs are considered as referring to entities or properties of a single entity, e.g. "en röd mugg med lite rött på handtaget" (a red cup with some red on the handle) (sv.P04.3) where "some red on the handle" was annotated as as a single entity rather than two distinctive entities. This convention is different from (Stoia et al., 2008) in the SCARE corpus where embedded noun phrases such as "this cabinet on the right" are annotated as belonging to the umbrella

noun phrase, possibly because here the annotation is limited by fixed pre-defined entities.

Note that participants see a slightly different scene where some cups are missing from their view which means that it may happen that they associate a certain description with different objects/cups. In other words, there may be a miscommunication of reference but which is normally resolved through clarification in dialogue once participants discover there are inconsistencies in their information states. Errors in the way the objects are described or expressions interpreted might also happen (sv.P05.110-115). In such cases we annotated expressions as referring to objects relative to the information state of the utterance speaker, the object that they intend to refer to. In most cases this can be resolved from the visual context of the speaker but sometimes the annotators have to guess about the cognitive state of the speaker.

To answer the question how difficult it is to annotate coreference using this annotation scheme before starting the annotation of the Swedish dialogues we re-annotated the first 14 turns or 250 words of one of the English dialogues (en.P02) for which annotations already exist (although there the annotator might have used different strategies as described above). To measure the agreement on noun phrase identification we calculate the κ coefficient on the BIO tags (B-NP, I-NP and O) which results in $\kappa = 0.84$. Unfortunately, κ cannot be used for referent identification as each noun phrase might refer to one or several referents. To estimate agreement on referent identification we calculate a Sørensen–Dice coefficient that measures the overlap of the identified referents of each noun phrase. We then average all coefficients over all noun phrases. The average Sørensen–Dice coefficient is $\bar{DSC} = 0.70$. Overall, there is a good agreement on both annotation tasks.

3.3 Reference and attention (Q3)

Examining what referents are referred to in dialogue might tell us something about participants attention on the scene. Examining this might give us important preliminary insights about the strategies of reference resolution. Figure 3 shows global reference to entities over all dialogues.

There is a tendency that participants (1 and 2) refer to themselves or each other the most (not so much to the passive observer Katie, 3), followed by the objects (21–38) and then regions (11–19).

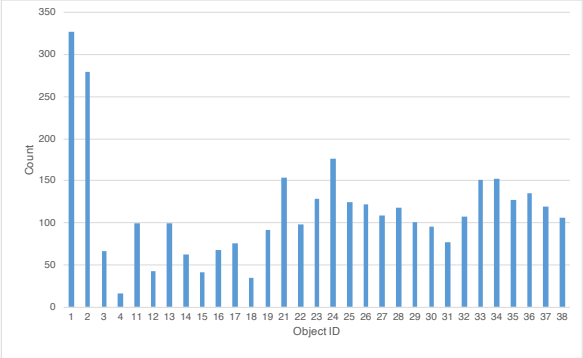


Figure 3: Reference to entities over all dialogues: 1-3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. See also Figure 1.

The participants' references to themselves reflect the collaborative nature of the task. Katie on the other hand is only used infrequently as a landmark to relate other objects to, for example: "På den sida där Katie inte står" (On the side where Katie is not standing) (sv.P02.48-49) or to set the spatial frame of reference or perspective on the scene "okej första raden fran katie pa hennes hogra sida. . . " (Okay, in the first row in front of Katie on her right side,. . .) (sv.P01.54-57).

Cups are more frequently referred to than regions which is expected as they are the objects of the task while regions refer to their locations. Note that 11, 13, 19 and 17 are the most attended regions. These represent corners of the table and are therefore good landmarks. Another reason why our pre-annotated and to participants invisible regions might be used less is that such geometric division is less natural for participants to refer to who dynamically create regions based on the object topology. For example, they do not say "mittenkvadraten närmast dig" (the central square closest to you) but "the second row closest to you" which does not match the geometric regions. We were aware of this when pre-annotating the grid and our hope was that the grid would provide some coarse granularity to annotate regions but in some cases it is hard to match the region referred to and the geometric region and in these cases labels for new dynamic regions were created. Sometimes it is hard to determine whether an expression is referring to a region, for example in elliptical noun phrases. We considered " din vänsterkant" (your left corner) (sv.P05.57.11-12), "till höger om den vita" (to the right of the white one) (sv.P05.39.10-14) as regions but "den står typ innanför den gula muggen" (It is roughly standing outside of the yellow cup)

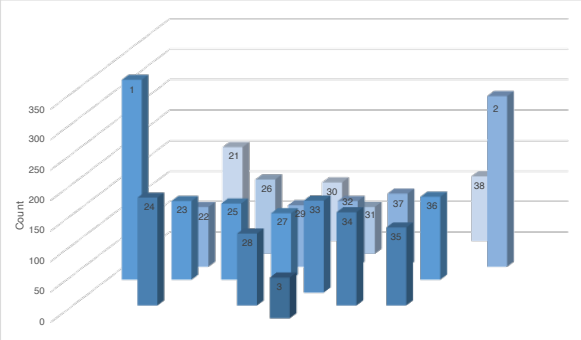


Figure 4: Attention over objects as measured by reference to them. The columns are arranged in the same spatial configuration as objects on the scene. Object 4, the table, is not shown.

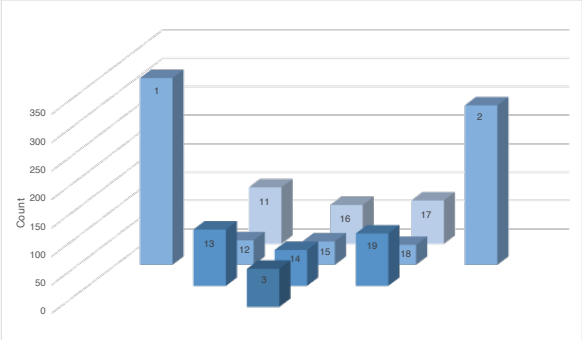


Figure 5: Attention over regions as measured by reference to them. The columns are arranged in the same spatial configuration as regions in the scene. For reference P1 (1), P2 (2) and Katie (3) are also included.

(sv.P04.12). This example demonstrates that the interpretation of these expressions as regions depends on the context.

Cups in the visually ambiguous configurations and cups that are missing from either participants view are referred to more often and therefore receive more attention. For example cup 24 which is hidden from P2 but also easily confused with cup 21 which is positioned in the opposite corner close to P1 (see Figure 1c). Moreover, cup 21 can also be confused with cup 26 which is close by and missing for P1. Similarly, there can be misunderstanding regarding cup 34 which is hidden from P1 but there is a similar cup 33 close by (sv.P06.24-26, dialogue in Appendix A.1) and cup 23 and 25 where the latter is hidden from P2.

Figure 4 shows the distribution of attention over the visual scene as measured by the reference to objects in dialogue. It can be seen that overall (with variations described previously) attention is more or less distributed over objects which can be explained by the nature of the task: the participants need to evaluate a consistency of each other's descriptions against the entire scene. There is a tendency that cups closer to Katie receive more attention than cups on the opposite side, the ranking being 24, 21, 34, 33... in descending order. Note that there is a similar ambiguity on both sides of the table: between 34 (not visible to P1) and 33 on Katie's side and 21, 26 (not visible for P1) and 29 on the opposite side. Therefore, there may be an effect of the presence of Katie on the grounds that she is an animate being and a good point of reference to relate other objects to (Lipp et al., 2004). However, she is not referred to specifically as she is not taking part in the task.

Figure 5 shows the distribution of attention over regions. Regions on the side of the table (13,14,19 and 11,16,17) attract more attention than regions in the middle (12,15,18). This indicates that participants prefer the lateral dimension over the front-back dimension when relating objects which coincides with observations from literature on spatial cognition: "från mitt håll står det en take-away bakom den vita muggen / snett vänster om" (From my perspective, there is a take-away behind the white cup. Diagonally to the left.) (sv.P05.37-44). Also, regions in the corners of the table (11,13,17,19) receive more attention than the middle regions on both sides (14,16), in fact these are also the most attended regions. This appears to be due to the fact that these corners are closest to participants who split the table in two halves: "mer på min sida än på din" (more on my side than on yours) (sv.P02.62-63). Note that closest to a participant does not mean closest to the speaker. Reference to participants is much higher than regions and so is reference to cups, presumably due to the nature of the task. Regions are mainly used as landmarks to describe the location of cups: "på kates vänstra sida innåt framför dig" (on Katie's left side in front of you) (sv.P06.58).

Comparing the attention over cups in Figure 4 with attention over regions in Figure 5 we can observe a low correspondence, e.g. attention on regions 13 and 11 might be associated with objects 24 and 21. However, when we sum the references to cups per regions, the cups that fall in the middle regions (12,15,18) are referred to more often than the cups that overlap with the lateral regions (11,16,17) and (13,14,19). Therefore, it could be that for the reference to the cups in the central regions the side

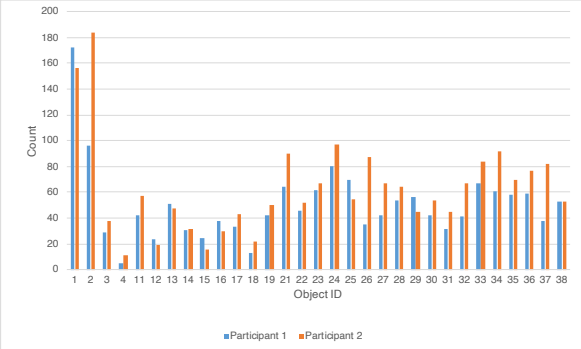


Figure 6: Reference to entities over all dialogues per participant: 1–3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. See also Figure 1.

regions are serving as landmarks, e.g. "Den står emellan den röda muggen på din vänstra sida och den gula muggen som står lite längre bort på din vänstra sida" (It is standing between the red cup on your left side and the yellow cup that is standing a bit further away on your left side.) (sv.P04.18).

Do conversational partners refer/attend to entities differently? Figure 6 shows reference to entities over all dialogues per participant. The reference to entities follows the same pattern for both participants which therefore also corresponds to the pattern in Figure 3. This shows that there is no preference for cups that would be closer and or more distant to a particular participant. Distance to an object does not seem to affect attention of that object. On the contrary, there is a tendency that objects that are hidden from the other participant (P1: 26, 34 and 37; P2: 24, 25 and 29) receive considerable attention. A likely explanation for this is that this is because conversational participants are engaged in a collaborative task that requires referring and subsequently co-referring to the same objects by the other partner until the task is completed (sv.P06.24-26, sv.P02.36-42, sv.P07.117-122, dialogues in Appendix A.1). Participant P1 refers to themselves more often than P2 and vice versa (sv.P06.220). (Dobnik et al., 2020) observe on the same dataset that the speaker's spatial perspective is used more often than hearer's. P2 refers to more objects than P1.

4 Discussion

We examined reference and coreference in visual dialogue. We argued that through patterns of reference in dialogue and the visual scene we can reconstruct patterns of attention that lead to production of these referring expressions. Due to the collaborative nature of dialogue these are also used by hearers to interpret referring expressions. Information how perceptual and discourse contexts interact in generation and interpretation of referring expressions is relevant for any computational application of vision and language as it allows us to resolve ambiguity that results through underspecification of referring expressions. As a starting point we took an established reference and coreference annotation scheme from the textual domain and adapted it to the domain of visual dialogue where linguistic expressions are also matched with referents grounded in the visual scene. This departs from the annotation strategies in the textual domain where discourse entities are introduced sequentially in text as they are referred to and then are subsequently re-referred to. However, in this domain discourse referents are already present once a participant sees and parses the scene: we indicated this by assigning participants, objects and regions fixed IDs. Additionally, we allow creation of dynamic entities and regions which are introduced in the same way as in the traditional textual co-reference annotation. Our notion of co-reference is also slightly different from the notion of co-reference in the text only domain. We do no specifically annotate coreference as a relation between referring expressions but this can be be inferred from the annotation scheme. We annotate a reference of referring expressions as a list of objects that an expression is referring to and hence if two referring expressions refer to the same objects then they are coreferential. Our annotation convention also allows us to compare referring expressions for partial (co)reference in case only some of the object IDs match. Additionally, object IDs could also be grouped and groups assigned IDs if coarse granularity of coreference would be required. For example, "(En 36) av (dem 21, 26, 36) står på (min sida 17, 18, 19), lite till (höger 18, 19) om (mitten 18)." (One of them stands on my side, a bit to the right of the middle) (sv.P02.19).

As conversation progresses, participants associate referring expressions with these entities based on how salient they are in the common ground; this is the reason why a description such as as "the red cup on the left" can be used successfully and the hearer can resolve its reference. We argue that the salience can be modelled as attention and the analysis of data in this paper is a first step towards computational modelling of reference and coref-

erence resolution in visual dialogue. Below we summarise our main findings.

Objects are most frequently co-referred to within the same conversational game. Our analysis of longer open dialogues shows that participants most frequently corefer to entities within 0 to 4 turns which coincides with the previous research on the length and structure of conversational games. These can also be nested. Conversational games depend on the collaboratory (sub)task that the participants are performing and once a task is complete and participants reach a mutual agreement they continue with a new task and a conversational game. Tasks are structured around a certain strategy which rarely considers the entire scene. Therefore, if a location of particular objects has been discussed, disambiguated and added to the common grounds of participants there is no need to discuss them again, unless they are used as salient landmarks for discussion of subsequent objects in new conversational games (sv.P02.73-82, dialogue in Appendix A.1). Structuring dialogue into sub-units explains why there are underspecified referring expressions since their scope can be resolved within the scope of these units.

The strategies to assign and resolve reference and co-reference are dynamic and creative. Although we have identified entities and regions in the visual scene within a certain level of granularity we frequently found cases where this was insufficient to fully capture the reference of the linguistic expressions. Firstly, not all noun phrases are referential, for example they can be expletives, referring to entities not present in the scene, non-referring (abstract and generics) or undetermined entities (wh-phrases and noun phrases used in questions). Here the challenge is that the same referring expression can be either referential or non-referential depending on the context in which it is used. For example, in "So maybe we could possibly go row by row, do you think? And say which cups are there? Or how should we work out where your unique cup are and vice versa?" does the speaker refer to specific alignments of cups, an abstract grid of rows or rows in general (en.P02.9)? Secondly, it is sometimes hard to decide what should be identified as a scene entity and what the granularity of regions should be, cf. our earlier example whether a sub-region of a handle constitutes a separate region (sv.P04.3). Thirdly, the same expression used by two conversational participants may be considered to refer to different entities by different conversational participants. These issues were resolved (i) by introducing four labels (expletive, external, non-referring and wh-questions) and annotation conventions (ii) by identifying referents on the basis of the information state of the speaker and (iii) by sometimes introducing new reference IDs to distinct parts of the visual scene dynamically. Even following these conventions, we sometimes had to make sub-optimal decisions in borderline cases. Overall, we were striving for regularity and consistency of the annotation scheme, so that it can be used in computational applications, as well as for informational richness and accuracy of semantic representation.

Reference of referring expressions points to spatial attentional patterns in the visual scene. For example, lateral regions are more attended than front-back regions while cups in the middle regions receive more attention. This is possibly because lateral dimensions serve as landmarks for describing target objects or cups or because front and back dimensions are referred to differently, relative to P1 and P2, e.g. "close to you" or as "left or right of Katie". Reference to lateral dimensions is frequently combined with the front and back dimension but this is described as a relation between objects rather than a reference to regions (and therefore may not be annotated): "sa till vanster; gul. sedan vit takeaway starx nedanför till höger…" (So to the left: yellow. Then a white take away just below to the right) (sv.P01.86-89). Note that spatial descriptions such as "vänster" (left) and "ovanför" (above) can either elliptically refer to regions or relations between objects. Distance of a participant to an object does not mean that a participant puts more attention to it. This is because participants collaboratively discuss all the objects in the scene, not just those that are close to them. Moreover, the data shows that certain cups receive attention when they are located in the areas where visual ambiguity is high, for example in regions where there are neighbouring similar cups or where there are cups hidden from a participant. Therefore attention to objects is driven by the task which is disambiguating the location of cups (sv.P05.64-88).

Overall, our work shows that conversational participants can communicate successfully in situations where both linguistic and visual information are underspecified. The underspecification is resolved from a variety of signals which do not

necessarily have a fixed meaning across all contexts. Strategies are chosen on the fly without a specific communicative signal which suggests that conversational participants need to reach agreements by processes of "virtual bargaining" (Misyak and Chater, 2014). This suggest that in visual co-reference resolution we should not look so much for patterns that can be directly extracted from the data as these might be context-specific but for communicative strategies that are available to participants and go beyond specific contexts. In (Loáiciga et al., 2021a,b) we compare reference and coreference in the Cups corpus with the Tell-me-more corpus (Ilinykh et al., 2019). The latter consists of shorter dialogues (normally one conversational game) over real images of environments that are different for each dialogue. The results indicate that the same strategies can be found in different contexts and tasks.

We examined the attention on the objects in the scene as a whole but it would also be important to examine how the attention changes when the dialogue unfolds. We expect that this would reveal interesting generalisations that would guide a computational system for co-reference resolution of individual referring expressions. A closer study of reference and segments of dialogues or dialogue games would also be in place as these may be natural boundaries of coreference. Are the same attentional patterns found within a dialogue and across the participant pairs? Is there a difference between English and Swedish dialogues? We have studied how referring expressions are mapped to objects but how are objects described by referring expressions (within conversational games) and how are referring expressions adapted when the same objects is re-referred to in the subsequent turns? Given the mechanisms of joint (visual and linguistic) attention how can linguistic forms be simplified and still be effective?

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. I am grateful to Sharid Loáiciga for important comments on various stages of this work. Three anonymous referees provided insightful and extremely useful comments of the original submitted version.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Donna K Byron. 2003. Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 39–47.

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

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in English and Swedish dialogue. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.

Simon Dobnik and Sharid Loáiciga. 2019. On visual coreference chains resolution. In *Proceedings of LondonLogue – Semdial 2019: The 23rd Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, London, UK. Queen Mary University of London.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

John D. Kelleher, Fintan J. Costello, and Josef van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167(1):62–102.

John D. Kelleher and Simon Dobnik. 2020. Referring to the recently seen: reference and perceptual memory in situated dialogue. In *CLASP Papers in Computational Linguistics: Dialogue and Perception – Extended papers from DaP-2018 Gothenburg*, volume 2, pages 41–50, Gothenburg, Sweden. University of Gothenburg, CLASP, Centre for Language and Studies in Probability and GUPEA.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3):339–354.

Ottmar V Lipp, Nazanin Derakshan, Allison M Waters, and Sandra Logies. 2004. Snakes and cats in the flower bed: fast detection is not specific to pictures of fear-relevant animals. *Emotion*, 4(3):233.

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

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

Jennifer B. Misyak and Nick Chater. 2014. Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130487.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems. Master's thesis in language technology, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, June 14.

Cameron Smith, Nigel Crook, Simon Dobnik, Daniel Charlton, Johan Boye, Stephen Pulman, Raul Santos de la Camara, Markku Turunen, David Benyon, Jay Bradley, Björn Gambäck, Preben Hansen, Oli Mival, Nick Webb, and Marc Cavazza. 2011. Interaction strategies for an affective conversational agent. *Presence: Teleoperators and Virtual Environments*, 20(5):395–411.

Laura Stoia, Darla Magdalena Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: a situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 650–653, Marrakech, Morocco. European Language Resources Association (ELRA).

Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue referring to spatial scenes. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Ylva Byrman.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. 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 5122–5131, Hong Kong, China. Association for Computational Linguistics.

A Appendices

A.1 Referring in dialogue

(sv.P04.46-49)

46 P1: mellan den blå och gula$_{28,35}$, framför Katie, ser jag en mugg$_{33}$ med lock och utan handtag
Between the blue and yellow in front of Katie, I see a cup with a lid and without a handle.
47 P2: Står den$_{33}$ lite längre bort från Katie (lite mer mot mitten) än den gula$_{35}$ och den blå$_{28}$?
Is it standing a bit further away from Katie (a bit more towards the middle) than the yellow and the blue?
48 P1: lite mot mitten inte exakt mellan den blåa och gula$_{28,35}$
A bit towards the middle, not exactly between the blue and yellow.
49 P2: OK, den muggen$_{33}$ kan jag se.
Ok, I can see that cup.

(sv.P06.24-26)

24 P2: lite till vänster om den står en vit$_{34}$
A bit to the left of it, there is a white.
25 P1: ja som$_{33}$ har en annan form$_{25,33,37}$ och till vänster i hörnet en röd
Yes, that has another shape and to the left, in the corner, a red.
26 P2: jag har ingen röd! och den vita34 har samma form$_{21,22,23,24,26,27,28,29,30,31,32,34,35,36,38}$
I have no red! And the white has the same shape.

(sv.P02.36-42)

36 P2: Jag ser ju två röda$_{21,26}$ i ditt vänstra hörn
I see two red in your left corner.
...
40 P1: Ser du två röda$_{21,26}$ bredvid varandra$_{21,26}$?
Do you see two red next to each other?
41 P2: Precis, en$_{21}$ är längst ner i hörnet och en$_{26}$ är precis nedanför den$_{21}$ (från mitt håll sett)
Exactly, one is in the bottom corner and one is just below it (as seen from my perspective)
42 P1: Så är det inte för mig. Jag har en röd mugg$_{21,24}$ i varje hörn från där jag står.
It is not like that for me. I have a red cup in each corner from where I stand.

(sv.P07.117-122)

117 P1: den raden$_{21,22,23,24}$ som är närmast mig som du precis beskrev
The row closest to me that you just described.
118 P2: mm
Mm.
119 P1: bakom den$_{21,22,23,24}$ i din riktning står en take away-mugg$_{25}$
Behind it in your direction, there is a take away cup.
120 P1: ”på en ensam ””rad””$_{25}$”
On a separate row.
121 P2: ok, så rad två$_{25}$ för dig är en ensam take away mugg$_{25}$?
Ok, so row two for you is a solo take away cup?
122 P1: snett till vänster bakom den vita muggen$_{23}$ mitt framför mig
Diagonally to the left behind the white cup just in front of me.

(sv.P02.73-82)

73 P2: Sen är det två vita$_{33,34}$ kvar. En$_{33}$ har lock$_{5133}$ och en$_{34}$ har inte det. Den$_{34}$ som inte har lock$_{5125,5133,5137}$ står längst ut av dem$_{33,34}$, i princip framför Katie.
Then, there are two white left. One has a lid and one does not. The one that does not have a lid is positioned farthest out of them, basically in front of Katie.
74 P2: Ser du den$_{34}$ som står precis framför henne?
Do you see the one just in front of her?
75 P1: Nej det som står framför henne är en blå mugg$_{28}$.
No, what is in front of her is a blue cup.
76 P2: Hmm, okej. Finns det inget bakom den muggen$_{28}$?
Hum, okay. Is there nothing behind that cup?
77 P1: Den blåa muggen$_{28}$?
The blue cup?
...
79 P2: Precis. Ser du något bakom den$_{28}$?
Exactly. Do you see anything behind it?
80 P1: Nope
Nope.
81 P2: Okej... då tror jag att du kan anteckna att det står en vit mugg$_{34}$ utan lock$_{5125,5133,5137}$ precis framför Katie.
Okay. . . Then I think you can mark a white cup without a lid just in front of Katie.
82 P1: Okej, antecknat.
Okay, noted.

A.2 Annotation

Reference and coreference annotation of (sv.P04.25-26). The columns represent dialogue ID, participant ID, turn number, word number in turn, word, noun phrase tag, coreference annotation and English translation. The latter is only used here and is not part of the corpus annotation.

```
P04 1 25 1 jag    B-NP (1)         I
P04 1 25 2 ser    O                see
P04 1 25 3 tre    B-NP (22,28,31   three
P04 1 25 4 blåa   I-NP  22,28,31) blue

P04 2 26 1 Jag    B-NP (2)         I
P04 2 26 2 kan    O                can
P04 2 26 3 också  O                also
P04 2 26 4 se     O                see
P04 2 26 5 3      B-NP (22,28,31   3
P04 2 26 6 blå    I-NP             blue
P04 2 26 7 muggar I-NP  22,28,31)  cups
P04 2 26 8 .      O                .
```

Justifiable reasons for everyone: Dialogical reasoning in patients with schizophrenia

Christine Howes
University of Gothenburg
christine.howes@gu.se

Ellen Breitholtz
University of Gothenburg
ellen.breitholtz@ling.gu.se

Mary Lavelle
City, University of London
mary.lavelle@city.ac.uk

Robin Cooper
University of Gothenburg
robin.cooper@ling.gu.se

Abstract

Patients with schizophrenia are known to have difficulties in reasoning, but previous work has not looked at how such deficits manifest in face-to-face interactions. Using a unique corpus of triadic interactions discussing a moral dilemma, half of which involve a patient with schizophrenia, we show that patients are more likely than their interlocutors and control groups to provide arguments which reject the constraints of the task. Patients are also more likely to be consistent in their reasoning across a dialogue than their interlocutors or controls. Our results suggest that patients do not have impaired reasoning abilities but rather reason on the basis of a different view of the task than non-patients.

1 Introduction

This paper investigates reasoning and argumentation in dialogues involving patients with schizophrenia. Patients perform poorly on social cognitive assessments designed to examine the mental operations underlying social interaction (Green et al., 2015). They show difficulty inferring what others are thinking (Brüne, 2005), and a bias towards jumping to conclusions when making decisions (Dudley and Over, 2003; Serrano-Guerrero et al., 2020). However, few studies have investigated how patients actually verbalise their reasoning during actual social encounters, or how patients' ability to reason influences their interactions more broadly.

To investigate this we examined the reasoning of patients, and their interacting partners, during triadic face to face social interactions. The reasoning in dialogues including a patient was compared to those seen in comparable control interactions. Crucially, for our purposes, the non-patients interacting with a patient were unaware of the patient's diagnosis, thus avoiding the potential changes to behaviour observed in interactions with patients with schizophrenia that can be attributed to stigma (Perry et al., 2011). Across all interactions, participants were asked to discuss a moral dilemma and reach an agreement. The dilemma stated that there were four passengers in a hot air balloon which was falling and one passenger needed to be sacrificed to save the remaining three. This task is known to elicit chains of reasoning, as shown in example (1), which we return to in section 3.2, below.

(1) Group 13, lines 32–39

3: But cancer research scientist. That's the type of research he must have lots of notes

1: But that's what I'm thinking I'm sure I'm sure there's notes there somewhere. That someone else can work on what he did.

3: And let's give the child a future cause she is a prodigy so th- that the cancer research scientist

1: Well she is a prodigy but having said that you know, like, what's she gonna do for anyone?

2 Background

Schizophrenia is a severe psychiatric disorder that affects millions of people worldwide. Patients are known to have difficulty with language (Covington et al., 2005; Stephane et al., 2014) and reasoning (Hooker et al., 2000; Zajenkowski et al., 2011; Contreras et al., 2016; McLean et al., 2017), and difficulty interacting with others is one of the most debilitating features of the disorder. However, the reasons for patients' social deficits are poorly understood and treatment options remain limited (Horan and Green, 2017).

A wealth of evidence suggests that patients have difficulty perceiving and interpreting social cues

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

from the world around them including interpreting others' emotions and inferring others' thoughts (Green et al., 2015; Brüne, 2005; Penn et al., 2008). Patients with schizophrenia have also been identified as having reasoning deficits, particularly biases of jumping to conclusions and evidence integration (Dudley and Over, 2003; McLean et al., 2017; Serrano-Guerrero et al., 2020). This means that patients are quicker to reach conclusions, possibly based on more limited evidence, and also more likely to stick with their initial conclusion even in the face of new evidence, suggesting that they are less flexible in their reasoning. Moreover, it has been hypothesised that reasoning impairments may underpin patients' social deficits (Corcoran and Frith, 2005).

However, these findings are derived from the results of pen and paper cognitive tasks, completed in isolation. They differ substantially from actual social interaction with others and it is unclear if patients' performance on such tasks reflects their social deficit as it presents during actual social interactions. Indeed, recent evidence suggests that patients' performance on such reasoning tasks reflects the cognitive demands of the task rather than patients' reasoning ability per se (Klein and Pinkham, 2018).

The few studies that have investigated patients' social interactions directly reveal that patients display atypical patterns of participation (Lavelle et al., 2014). Furthermore, the presence of a patient with schizophrenia in an interaction influences the non-verbal behaviour of their interacting partners, both in clinical contexts (Lavelle et al., 2015) and during first meetings with unfamiliar strangers, despite the diagnosis of the patient being undisclosed to their interacting partners (Lavelle et al., 2013, 2014). Studies indicate that this is also true in dialogue for disfluencies (Howes et al., 2017), and the relationship between self-repair and gesture (Howes et al., 2016).

This preliminary study aims to assess whether the results from offline cognitive tests which show reasoning deficits in patients generalise to face-to-face interactions with healthy participants. We are also interested in investigating whether healthy interlocutors reasoning behaviour is influenced by the presence of a patient (without this being explicitly known), as is the case for non-verbal behaviours (Lavelle et al., 2013, 2014) and turn-taking cues (Howes et al., 2017).

Specifically we investigate the following three questions:

1. Compared to participants in the control group interactions, do patients provide fewer reasons for saving/throwing the passengers?

2. Does the presence of the patient in an interaction lead to a different pattern of reasoning in patients' healthy participant partners?

3. Are patients less flexible/more consistent in their argumentation compared to healthy controls?

3 Method

3.1 Participants

The corpus, described in more detail in Lavelle et al. (2013), consists of 40 triadic conversations of approximately five minutes. There are 20 interactions involving one patient with a diagnosis of schizophrenia and two non-psychiatric controls who were unaware of the patient's diagnosis. The 20 control interactions each involved three healthy participants. Participants within each triad were unfamiliar to each other. For both the control group and the patient group one dialogue was not correctly recorded. The data available for analysis therefore consists of 19 patient interactions and 19 control interactions.

3.2 Task

The subjects discussed the balloon task – a moral dilemma which requires participants to reach agreement on which of four passengers should be thrown out of a hot air balloon that will otherwise crash, killing all the passengers, if one is not sacrificed. The four passengers are:

William Harris – the balloon pilot who is the only passenger with any balloon flying experience

Susanne Harris – William's wife, a primary school teacher who is 7 months pregnant with their second child

Dr Robert Lewis – a cancer research scientist, who believes he is on the brink of discovering a cure for most common types of cancer

Heather Sloan – a nine-year old musical child prodigy who is considered by many to be a "twenty-first century Mozart"

This task is known to elicit dialogues containing extended reasoning sequences, as illustrated in example 1. In this short extract, typical of the exchanges of reasons the task elicits, participant 3 provides a reason for not saving Dr Robert Lewis ("he must have lots of notes"), which participant 1 elaborates on ("I'm sure there's notes there somewhere. That someone else can work on what he did"). Subsequently participant 3 offers a reason to save Heather Sloan ("let's give the child a future cause she is a prodigy"), while participant 1 provides a possible reason not to save her ("Well she is a prodigy but having said that you know, like, what's she gonna do for anyone?").

3.3 Annotation

The dialogues were video recorded and motion captured. They were transcribed for the verbal content using ELAN (Wittenburg et al., 2006).

As the transcriptions were segmented based on the sound properties of the interaction, a single turn can be transcribed as multiple utterances, where there are within turn silences. For the purposes of our study, we call a turn a stretch of talk by a single speaker, regardless of how many sub-utterances it contains. As we treat a change of speaker to indicate a new turn, this means that some contributions which should in fact be counted as single turns may be broken up by intervening (even overlapping) material by another speaker, such as a backchannel (Yngve, 1970; Kjellmer, 2009). This pattern of interleaving of turns and utterances is a known issue in quantificational dialogue research (Purver et al., 2009) and decisions about what counts as a turn or an utterance may have consequences for comparisons to other work, though as we treat all groups the same here, it can just be considered as noise in the data.

These anonymised text transcripts were used as the basis for the annotation of reasons. Annotators were not aware which participants were patients or even which dialogues were the control dialogues and which contained a patient.

The annotation involved a two step process. First, each utterance was coded for whether it related to any of the passengers in the balloon, or all of them. The annotators were prompted by the question: "Who does the utterance relate to?" These were not mutually exclusive categories – the same utterance could relate to several of the people in the balloon, and could be marked as such. The 'relates to everyone category' was only used if the participants were described collectively. Two of the dialogues were annotated by two of the authors. Cohen's kappa was between 0.65 and 0.9 for each of the categories relating to a person in the balloon.

The second stage examined those utterances which had previously been marked as being about one of the passengers (or everyone) and answered the question "Does the utterance make an argument that is directly or indirectly a reason for saving or not saving one, or all, of the passengers?". Where the reasoning in the turn spanned several utterances annotators were instructed to only annotate the final utterance of the reasoning sequence. Examples are shown in Table 1. Cohen's kappa for the two dialogues annotated by multiple authors was between 0.60 and 0.88 for each of the "reasons for" categories, despite the relatively few cases in some of the categories.

4 Results

As can be seen from table 2, patients produce fewer reasons on average than both their partners and controls. However, this is mediated by the amount of speech that participants produce – when we normalise by the number of turns (as seen in Table 3) there is no significant difference in the total proportion of reasons given between the participant types. Although on average there is a smaller proportion of patients' turns which contain a reason, there is wide variation in the numbers.

4.1 Number of reasons

While it appears that patients produce fewer reasons per turn in a number of categories (as seen in Figure 1 and Table 3), the wide variability and small number of cases in the data means that these numbers are not statistically significant. However, even taking into account the low power of analyses, which are based on one value per individual, we do see significant differences in the Save Everyone and Don't Save Anyone categories.

Of the total number of reasons provided by patients, as compared to both their partners and the control group, a higher proportion are about saving (or not saving) everyone. Reason to save everyone: $\chi^2_2 = 13.81, p = 0.001$; Reason to not save anyone: $\chi^2_2 = 16.95, p < 0.001$. This means that although patients are providing a similar number of reasons per turn as both their partners and controls, more of the reasons they give are a rejection of the

Text	Reason for
And let's give the child a future cause she is a prodigy	Save_Prodigy
Well she is a prodigy but having said that you know, like, what's she gonna do for anyone?	Don't_Save_Prodigy
if the wife jumps over, it means that she will die and her her unborn baby will die, so I mean that's two people who'll die	Save_Woman
the pregnant one would probably be the heaviest	Don't_Save_Woman
if Tom goes I think that nobody can drive this balloon	Save_Pilot
I would throw out the pilot and get the pilot to teach them how to fly the hot air balloon	Don't_Save_Pilot
The doctor could save lives	Save_Doctor
he's coming to like discovering this new cure, but he's probably been working with others	Don't_Save_Doctor
Think about how many child prodigies that we could save with Robert Lewis's cancer treatment	Save_Doctor; Don't_Save_Prodigy
Nobody's gonna go they they can control the balloon	Save_Everyone
Everybody should go down with the ship	Don't_Save_Anyone

Table 1: Annotation examples

	Patient			Patients' Partner			Controls			Total		
	Mean	Count	s.d.	Mean	Count	s.d.	Mean	Count	s.d.	Mean	Count	s.d.
Save Reasons	6.37	121	4.65	9.42	358	6.11	10.96	625	7.80	9.68	1104	6.96
Don't Save Reasons	4.21	80	4.37	7.50	285	5.76	7.58	432	5.86	6.99	797	5.70
Total Reasons	10.58	201	7.99	16.92	643	10.67	18.54	1057	12.31	16.68	1901	11.42
Total Turns	50.37	957	32.49	62.00	2356	31.10	66.58	3795	38.25	62.35	7108	35.26

Table 2: Overview of reasons given (raw data)

constraints of the task (which specifically states that the participants should come to an agreement about which of the passengers to throw out).

4.2 Consistency

Visual inspection of the dynamics of the reasons given in each triad (see Figures 2 and 3 for examples) suggested that the patients were more consistent in their reasoning in the sense that they did not seem to be as likely to produce a reason for and against the same individual, and produce arguments for or against fewer of the four individuals in the balloon.

These impressions were confirmed. In order to assess consistency, we used a very simple binary measure (consistent/inconsistent). Each participant was classified as inconsistent if they provided at least one argument for and against throwing the same individual, and consistent otherwise. In our data, patients are more likely to be consistent in the reasons they provide with 10 out of 19 patients (53%) not providing a reason both for and against the same individual, compared to 9 out of 38 of their partners (24%, $\chi^1_2 = 4.78, p = 0.03$) and 10 of 57 controls (18% $\chi^1_2 = 9.05, p = 0.003$).

For the number of individuals arguments were provided for or against, 11/19 of the patients (58%) produced arguments for 2 or fewer individuals, compared to 13/38 of their partners (34%, $\chi^1_2 = 2.91, p = 0.09$ not statistically significant, but a trend in the expected direction) and 14/57 of the healthy controls (25%, $\chi^1_2 = 7.17, p = 0.007$).

4.3 Qualitative observations

In this section we will present some examples that illustrate the patterns identified in the quantitative data, i.e. that patients appear less inclined to stick to the rules set up in the hypothetical situation and rather treat it as a situation in which they are themselves present. The impact of patients' behaviour on their interacting partners will be presented and the potential rationale for patients' reasoning will be discussed.

4.3.1 Save everyone or no-one

In examples 2 and 3 below, the patient (in bold) refuses to accept the premise of the task and instead argues that everyone should jump from the balloon (example 2) or no-one should jump from

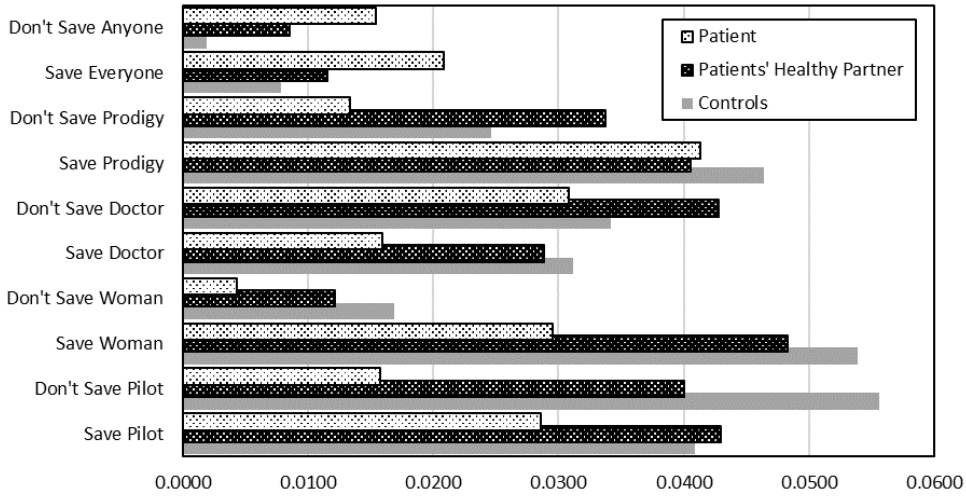


Figure 1: Reasons per turn by participant type

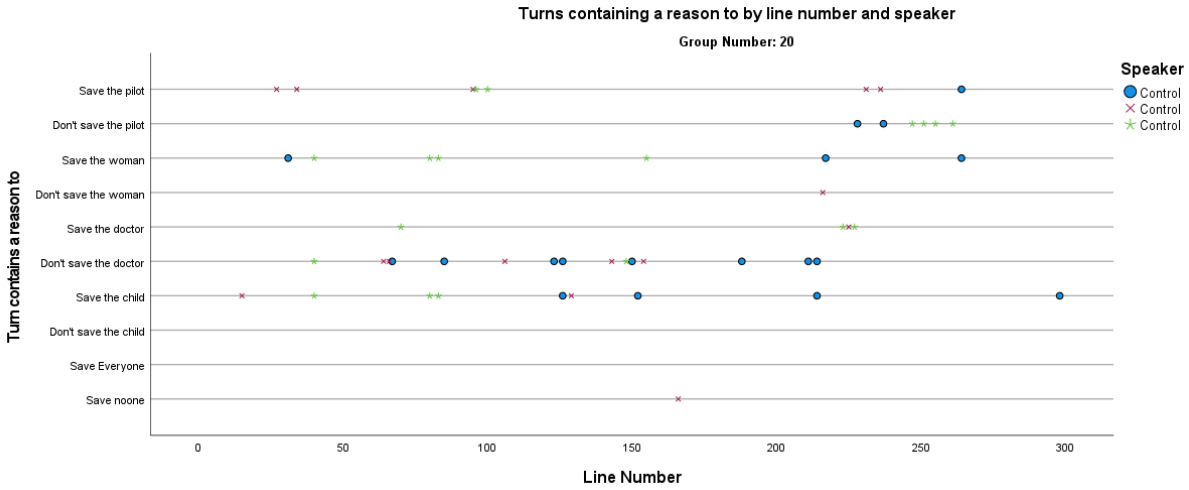


Figure 2: Single control dialogue showing the sequentiality of reasons given throughout the dialogue

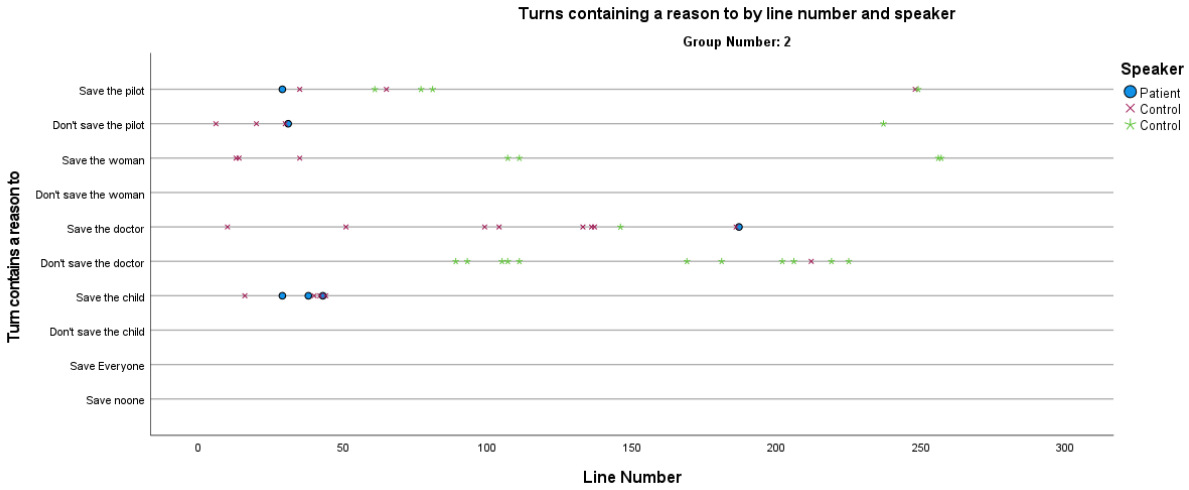


Figure 3: Single patient dialogue showing the sequentiality of reasons given throughout the dialogue

	Patient		Patients' Partner		Controls		Total	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Save Pilot	0.029	0.032	0.043	0.054	0.041	0.043	0.040	0.045
Don't Save Pilot	0.016	0.034	0.040	0.053	0.056	0.078	0.044	0.066
Save Woman	0.030	0.050	0.048	0.055	0.054	0.065	0.048	0.060
Don't Save Woman	0.004	0.013	0.012	0.036	0.017	0.035	0.013	0.033
Save Doctor	0.016	0.024	0.029	0.042	0.031	0.037	0.028	0.037
Don't Save Doctor	0.031	0.048	0.043	0.064	0.034	0.050	0.036	0.054
Save Prodigy	0.041	0.042	0.041	0.040	0.046	0.046	0.044	0.043
Don't Save Prodigy	0.013	0.029	0.034	0.056	0.025	0.035	0.026	0.043
Save Everyone	0.021	0.055	0.012	0.026	0.008	0.021	0.011	0.031
Don't Save Anyone	0.015	0.034	0.009	0.024	0.002	0.006	0.006	0.021
Total Save Someone	0.136	0.094	0.172	0.121	0.180	0.119	0.170	0.116
Total Don't Save Someone	0.080	0.070	0.137	0.098	0.133	0.121	0.126	0.108
Total Reasons	0.216	0.122	0.309	0.190	0.313	0.221	0.296	0.199

Table 3: Reasons given for saving or not saving each person by number of turns

the balloon (example 3), both scenarios resulting in everyone dying. From a utilitarian point of view, this is not very rational as no one is spared. On the other hand it might seem more *fair* that one person is not sacrificed.

(2) Group 3, lines 163–170

3: [you just have] to you have to accept everybody you have to [accept]
1: [everyone.]
1: yeah.
3: [<unclear/>]
2: [You think] we should all jump?
3: I think
3: well that er th- ah everybody should go down with the ship, yeah

In example 2, the patient's interacting partners do not appear to challenge the patient's deviation from the task rules. Participant 2 does request clarification of what the patient has said (*you think we should all jump?*) suggesting that the patient's argument requires further discussion, but they make no explicit reference to this being outside of the premise of the task. Furthermore, patients' partners use the term *we* should jump rather than *they* suggests that they may be adopting the patient's interpretation of the task being about themselves rather than abstract individuals. This pattern may indicate a compensation by the patient's interacting partners to the rigid stance of the patient.

By contrast, in example 3, the patient's partners become increasingly explicit about the fact that the patient has deviated from the premise of the task. This is demonstrated by participant 3 firstly asking the patient directly (*who do you think should go?*). Following the patient's response (*nobody's gonna go...*), participant 1 explicitly acknowledges the problem with it and restates the premise of the task. Eventually participant 1 presents the patient's position to them, although it is clear from the dialogue that they do not understand or share it (*but you're hoping on a miracle then*).

(3) Group 5, lines 88–107

3: [Well who do you think should go?] Who do you think should
2: Nobody's gonna go they they can control the balloon
2: knows the pilot
2: but he don't want to [<unclear/>]
1: [But one of] them has to go
1: one of [the four]=
3: [has to]
1: = has to go
2: [<unclear/>]
1: [Otherwise] they [all die.]
2: [I don't know.]
2: I don't know.
2: If you're gonna die, the pilot is there.
1: But that's the premise of the issue [that there a-]
2: [No I don't I] don't think they're gonna die.

1: <laughter/> [<laughter/>]

2: [I can <unclear/>] let's save them with that other people you know how they want to save themselves

1: [Right].

2: [<unclear/>] <unclear/>

1: But you're hoping on a miracle then.

4.3.2 Consistency of position

Examples (4)–(6) are extracts from the same dialogue demonstrating the consistency of one patient's arguments and rationale over the course of an interaction. Patients' pattern of consistency suggests a rigidity in their ability to consider multiple alternative view points, which aligns with findings from cognitive tests (García-Mieres et al., 2020). In the current example the patient (participant 2) states that the woman should be saved and presents their reason as (*it's two people, there's a baby there as well*). The patient states their argument to save the woman in response to an alternative view of one of their interacting partners (participant 3), and prefaces their reason with a moral value judgement (*it's just not right*).

(4) Group 4, lines 55–72

3: [Well] all the all his wife has got going for her

1: <laughter/> [<laughter/>]

3: [is that] his

3: she's his wife.

3: And she's

3: expecting.

2: but it's it's just not right

2: **it's two people.**

2: **there's a baby there as well**

1: yeah.

3: yeah [I know.]

1: [it is]

1: You're [killing two]=

2: [<unclear/>]

1: = lives [not just one]=

2: **[so there's two] [lives in there]**

3: [but I still] that that means it goes back to the weight as well innit?

3: she's a little bit extra [you know]

Following a short interlude where the possibility of throwing the prodigy is discussed, they return to the question of the pregnant woman, who participant 3 still advocates for throwing. While participant 1 offers a new reason for not throwing the woman (*if you throw the wife out the pilot won't be able to control the balloon or he might jump off*), the patient reiterates the same reason they had previously offered for keeping the woman (*there's a baby...*). Following a further exchange she reiterates it again (*you'd be killing two lives*).

(5) Group 4, lines 91–110

3: = I personally would say throw the wife out.

3: That's probably the

3: pilot be happiest then.

2: No child [can deserve that.]

3: [<laughter/>][<laughter/>]

1: [<laughter/>]

2: **There's a baby you want there, his [baby]**

1: [yeah] [it's a bit bad]

3: [yeah, but]

1: and like I said I think

1: [<unclear/>]

3: [There ain't a baby] there

1: I think if you throw the wife out though

1: I think the pilot [will]=

2: [mmm.]

1: = s-

1: won't be able to control the balloon [or he might]=

2: [mmm.]

1: = jump off

2: **and you'd be killing two lives too**

Between extract (5) and (6), participant 3 concludes that they have to keep the doctor and the only choice is between the child prodigy and the wife. At this point in the dialogue the healthy participants have presented and discussed multiple arguments for and against throwing the wife, while the patient continues to reiterate the same argument – that throwing the wife of the pilot involves sacrificing the unborn child.

In example (6) again, we see evidence of the patient deviating from the abstract and hypothetical nature of the task and discussing it as though they were involved. For example, when discussing throwing out the pregnant woman the patient states *I couldn't live with myself*. The patient justifies

their decision to keep the woman based on personal lived experience (*cause I'm a mother*).

(6) Group 4, lines 141–153

3: but the baby's not born yet.
3: <laughter/> [<laughter/>]
2: **[no but it's a] it's a life isn't it ?**
3: It is a life but
3: the baby's [not]
2: [couldn't] live live with myself.
2: Do you know what I mean
1: Right.
2: Cause I'm a mother <laughter/> [<unclear/>]
3: [I'm a fa]ther
2: [Because]
3: [<laughter/>] <laughter/>
2: Well you never carried a baby.

5 Discussion

Patients provide a similar number of reasons per turn as non-patients but a greater proportion of the patients' reasons involve rejecting the constraints of the task. This might have to do with the fact that patients have greater difficulty in seeing the task as a kind of abstract game as opposed to an imagined real-life situation in which decisions have to be made. Task based assessments have shown that schizophrenia patients have difficulty employing abstract thinking, which may account for this finding (Flavell, 1956; Oh et al., 2014).

Patients were shown to be more consistent than non-patients in that they produced fewer arguments both for and against throwing a particular person. This could indicate a lack of ability or willingness to weigh different arguments against each other. This pattern may stem from a cognitive rigidity, which has been identified in patients with schizophrenia using cognitive tests (García-Mieres et al., 2020). It may also be consistent with viewing the task as an imagined real-life situation rather than as an abstract game.

We plan to further investigate why this might be the case in future work, as our current analysis does not distinguish between several possible causes. For example, it may be that patients are simply more defensive or less engaged with the task (which would be consistent with their on average shorter dialogues) or it might be that that they find it hard to reason counterfactually, or that something about their experience as patients makes them take a different moral stance.

Our qualitative analysis suggests that patients' partners may choose to manage patients' deviation from the task in two ways: i) to adapt to the patient's position and discuss the task as the patient has interpreted it; ii) to explicitly tell the patient that they have deviated from the task rules. This may have implications for the success of these interactions and how they are experienced by the interacting partners.

In general our results suggest that patients do not have an impaired reasoning ability but rather reason on the basis of a different view of the task than that of the non-patients. Furthermore our qualitative results suggest that the patients offer fewer different types of reasons for the same conclusions (e.g. *do not throw the pregnant woman*) compared to healthy participants and controls. In order to investigate this hypothesis further, we are currently annotating the types of reasons given by patients, patients' partners and controls to find out how these differ between the different groups and the impact this has on the interaction success.

Acknowledgments

This work was supported by the Dialogical Reasoning in Patients with Schizophrenia (DRiPS) project funded by Riksbankens jubileumsfond (P16-0805:1). Howes, Breitholtz and Cooper were additionally supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP).

References

Martin Brüne. 2005. "Theory of mind" in schizophrenia: A review of the literature. *Schizophrenia Bulletin*, 31(1):21–42.

Fernando Contreras, Auria Albacete, Pere Castellví, Agnès Caño, Bessy Benejam, and José Manuel Menchón. 2016. Counterfactual reasoning deficits in schizophrenia patients. *PloS One*, 11(2):e0148440.

Rhiannon Corcoran and Christopher D Frith. 2005. Thematic reasoning and theory of mind. Accounting for social inference difficulties in schizophrenia. *Evolutionary Psychology*, 3(1):1–19.

Michael A Covington, Congzhou He, Cati Brown, Lorina Naci, Jonathan T McClain, Bess Sirmon Fjordbak, James Semple, and John Brown. 2005.

Schizophrenia and the structure of language: The linguist's view. *Schizophrenia research*, 77(1):85–98.

Robert E. J. Dudley and David E. Over. 2003. People with delusions jump to conclusions: a theoretical account of research findings on the reasoning of people with delusions. *Clinical Psychology & Psychotherapy*, 10(5):263–274.

John H Flavell. 1956. Abstract thinking and social behavior in schizophrenia. *The Journal of Abnormal and Social Psychology*, 52(2):208.

Helena García-Mieres, Judith Usall, Guillem Feixas, and Susana Ochoa. 2020. Placing cognitive rigidity in interpersonal context in psychosis: Relationship with low cognitive reserve and high self-certainty. *Frontiers in Psychiatry*, 11.

Michael F Green, William P Horan, and Junghee Lee. 2015. Social cognition in schizophrenia. *Nature Reviews Neuroscience*, 16(10):620.

Christine Hooker, Neal J Roese, and Sohee Park. 2000. Impoverished counterfactual thinking is associated with schizophrenia. *Psychiatry*, 63(4):326–335.

William P Horan and Michael F Green. 2017. Treatment of social cognition in schizophrenia: Current status and future directions. *Schizophrenia Research*, 203:3–11.

Christine Howes, Mary Lavelle, Patrick G. T. Healey, Julian Hough, and Rose McCabe. 2016. Helping hands? Gesture and self-repair in schizophrenia. In *Resources and processing of linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (RaPID) at LREC*.

Christine Howes, Mary Lavelle, Patrick G. T. Healey, Julian Hough, and Rose McCabe. 2017. Disfluencies in dialogues with patients with schizophrenia. In *Proceedings of the 39th annual meeting of the Cognitive Science Society*, London, UK.

Göran Kjellmer. 2009. Where do we backchannel? On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics*, 14(1):81–112.

Hans S Klein and Amy E Pinkham. 2018. Examining reasoning biases in schizophrenia using a modified "jumping to conclusions" probabilistic reasoning task. *Psychiatry research*, 270:180–186.

Mary Lavelle, S Dimic, C Wildgrube, R McCabe, and S Priebe. 2015. Non-verbal communication in meetings of psychiatrists and patients with schizophrenia. *Acta Psychiatrica Scandinavica*, 131(3):197–205.

Mary Lavelle, Patrick GT Healey, and Rosemarie McCabe. 2013. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, 39(5):1150–1158.

Mary Lavelle, Patrick GT Healey, and Rosemarie McCabe. 2014. Participation during first social encounters in schizophrenia. *PloS one*, 9(1):e77506.

Benjamin F McLean, Julie K Mattiske, and Ryan P Balzan. 2017. Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: A detailed meta-analysis. *Schizophrenia Bulletin*, 43(2):344–354.

Jooyoung Oh, Ji-Won Chun, Jung Suk Lee, and Jae-Jin Kim. 2014. Relationship between abstract thinking and eye gaze pattern in patients with schizophrenia. *Behavioral and Brain Functions*, 10(1):1–8.

David L Penn, Lawrence J Sanna, and David L Roberts. 2008. Social cognition in schizophrenia: An overview. *Schizophrenia Bulletin*, 34(3):408–411.

Yael Perry, Julie D Henry, Nisha Sethi, and Jessica R Grisham. 2011. The pain persists: How social exclusion affects individuals with schizophrenia. *British Journal of Clinical Psychology*, 50(4):339–349.

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*, pages 262–271, London, UK.

Estrella Serrano-Guerrero, Miguel Ruiz-Veguilla, Agustín Martín-Rodríguez, and Juan F Rodríguez-Testal. 2020. Inflexibility of beliefs and jumping to conclusions in active schizophrenia. *Psychiatry research*, 284:112776.

Massoud Stephane, Michael Kuskowski, and Jeanette Gundel. 2014. Abnormal dynamics of language in schizophrenia. *Psychiatry Research*, 216(3):320–324.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Victor H. Yngve. 1970. On getting a word in edgewise. In *Papers from the 6th regional meeting of the Chicago Linguistic Society*, pages 567–578.

Marcin Zajenkowski, Rafał Styła, and Jakub Szymanik. 2011. A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders*, 44:595–600.

Don't you think that a rhetorical question can convey an argument?

Denis Ioussef, Ellen Breitholtz and Christine Howes
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
d.ioussef@hotmail.com; ellen.breitholtz@ling.gu.se;
christine.howes@gu.se

Abstract

Rhetorical questions have been addressed from many different linguistic perspectives, however, their interactional role has been hitherto underexplored. We here present an exploratory study of rhetorical questions in a corpus of dialogues discussing a moral dilemma from an interactional perspective, using the notions of enthymemes and topoi. Results show that rhetorical questions are used to introduce enthymematic arguments and to facilitate linking together parts of arguments over several utterances.

1 Introduction

Rhetorical questions (RQs) have been addressed from different perspectives in linguistics, from Discourse Analysis and Speech Act Theory to transformational approaches, yet the interactional aspect of RQs has been neglected. We investigate the roles RQs play in the context of dialogue from an interactional approach, in particular as a device for delivering arguments.

1.1 Speech Acts and RQs

In Speech Act Theory RQs are regarded as indirect speech acts. By asking a question without expecting an answer, a speaker breaks the sincerity condition for questions and gives rise to a conversational implicature, typically conveying a statement (among other functions in discourse an imperative, a piece of advice, a criticism, a threat to face or an argument, etc.) (Grice, 1975; Searle, 1975; Brown and Levinson, 1987). On this view, the RQ makes its answer obvious, through context or by syntactic markers, to the exclusion of other possible answers to the literal meaning of the question. The statement that answers an RQ (and is implied by it) in the majority of cases exhibits the opposite polarity, that is the answer is settled in the negative in the mind of the hearer (Egg, 2007; Pope, 1972; Ilie, 1994; Han, 2002, 1998). In order to explain the mechanism of delivering arguments through RQs, one needs a way to derive the statement or imperative to explicate the illocutionary force of the RQ – the so called polarity shift or reversal being one of the main tools for glossing the implied statement the RQ carries, and a useful clue to evaluate the felicitousness of an RQ in a given context.

1.2 Discourse context and RQs

Cerović (2016) investigates the use of rhetorical questions in the institutional setting of a police interrogation, where a suspect uses RQs to challenge allegations posed by the detectives, and to demonstrate epistemic primacy regarding the crime, vis-a-vis the detectives interrogating him – the suspect asks "What do I know?" to assert "I know nothing" about his knowledge of the crime. The use of RQs instead of responding directly, poses a challenge to authority in a setting where it is detectives who are supposed to be "asking the questions" (Cerović, 2016). In accordance with this, Frank (1990) argues for the primary role of RQs being persuasive devices attenuating the social cost of face-threatening acts, by "strengthening assertions and mitigating potential threats to face" (Frank, 1990, p.738). Since RQs often convey sarcasm and are otherwise ambiguous regarding interpretation, subjective and easily misunderstood, the intent of a speaker isn't always clear-cut. Relying on SAT alone, a researcher would mistakenly classify as RQs cases where only context can cue such a reading. An assessment needs to be made not only of the speaker's intent but of the contextual environment and of the hearers response, made possible by Discourse Analysis (Frank, 1990).

Ilie's take is that RQs are not a special category of questions that needs not or can not be answered, but rather primarily pragmatic units that "are nei-

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

ther answerless, nor unanswerable questions, and that they display varying degrees of validity as argumentative acts" (Ilie, 1994, ii). Ilie distinguishes five identifying features as criteria for RQs. These are taken to be cognition oriented -– an RQ evokes a cognitive process in the mind of the addressee that mirrors the process in their own mind and arrives at the same conclusion, inducing the addressee to reconsider their own held assumptions. These five criteria are (p.45-46): (i) the discrepancy between the interrogative form of the rhetorical question and its communicative function as a statement, (ii) the polarity shift between the rhetorical question and its implied statement, (iii) the implicitness and the exclusiveness of the answer to the rhetorical question, (iv) the speaker's commitment to the implicit answer, and (v) the multifunctionality of rhetorical questions.

1.3 Rhetorical questions as enthymematic arguments

In argumentation as it occurs in natural dialogue participants often rely on common sense rather than strictly logical deduction in order to interpret the arguments made. Many arguments in dialogue are enthymematic – that is, the arguments presented lack some premises which would be required in a fully logical chain of reasoning. Instead, enthymematic arguments (*enthymemes*) rely on notions or warrants in the minds of the listeners. These are often referred to as *topoi* (Aristotle, ca. 340 B.C.E./2007; Ducrot, 1988; Anscombre, 1995; Breitholtz, 2020).

When we interact we expect certain topoi to be common ground, or to be accommodated during the course of the interaction. Different topoi can underpin one and the same enthymeme, which can lead to misunderstanding, disagreement, or agreement on completely different grounds.

When presented with an inference in a conversation the participants need to find among their rhetorical resources an applicable general principle that would make sense of it, that is to both interpret and to validate it. Ilie (1994) calls enthymemes those RQs that function as whole arguments, that is, imply conditional statements. She distinguishes three types of RQs that are enthymemes according to what kind of inference they correspond to, modus ponens, modus tollens and disjunctive syllogism. However, in our analysis RQs function as different aspects of enthymematic inferences. RQs can serve as replies or as something to be replied to, thus expressing the premise or conclusion of an argument (except those that reply to the question itself, i.e that verbalise the implied statement, as they only serve to strengthen the expressed standpoint).

Consider the example below of an RQ that delivers a statement of the speaker's opinion, (the immediate context is whether throwing out a child from an air balloon could prevent it from crashing):

(1) She's nine years old, she's so light anyway – is she really gonna make a difference? (GP12, 38)

In this example (1) the structure of the argumentation can be described as this enthymeme:

(2) $\dfrac{\text{she's nine years old} \quad \text{she's too light}}{\text{throwing her out won't make a difference}}$

Topos: if x isn't heavy enough, throwing x out of the balloon won't help it fly. The proposition that "she's too light" is itself the conclusion of, "she's nine years old":

(3) $\dfrac{\text{she's nine years old}}{\text{she's too light}}$

The wavy line represents defeasibility of the argument – that while there may be a good reason for the conclusion to follow from the premise, it may with additional information be invalidated. For example, one could imagine the child being so heavy that she would constitute an exception to the generalisation that nine year olds do not weigh much – she may not be a typical nine year old and the general rule may not hold true in her case. In other words, the topos may be accepted as valid but not its application in a particular context. The speaker appears to be aware of the possibility that the topos may not necessarily be accommodated as relevant to the situation, making the generalisation more specific later in the dialogue: "I just think the child is too too light anyway I mean even if the child was morbidly obese" (GP12, 211).

We argue that the employment of an RQ to form an argument strengthens the argumentative force through a presumption that the topos warranting it should already be acceptable to the addressee. This is due to the role of RQs in cognition itself, not simply to social tension and the risk of threatening face in the possibility of a challenging response to the RQ in case of eventual disagreement. The speaker, by using an RQ, presumes the notion behind it to be acceptable to the addressee, thus expressing their

own commitment to the implied answer and the expectation of the addressee to do likewise. Casting it as a question to be answered negatively sets off such a process in the mind of the hearer, whether they end up agreeing with the conclusions of this reasoning process or not. We propose that this process can be described as the successful elicitation of a topos that would warrant the enthymeme expressed by an RQ, and the RQ form itself accentuates, or makes salient in discourse the expectation that a topos is already available, and should be acceptable to the other conversation participants.

Schlöder et al. (2016) analyse why-questions in dialogue from a rhetorical perspective, drawing on Ginzburg's (2012) account of Question under Discussion (QUD). A why-question elicits a reason for the question under discussion: when someone utters a proposition p, the answer to "why p?" is an enthymeme q "therefore p", and the answer presupposes that there is a topos that warrants that enthymeme. A reason is factive, when what is asked about is why p holds, and meta-discursive when inquiring about the reason for the act of saying p.

There is a special case where the antecedent of a why-question is a conditional statement. Here the why-question elicits a reason for the stated enthymeme, to explicate the topos that underpins it. A why-question can be posed again to elicit a reason for the one already given, again be questioned, and so on, as there may always be "a topos in the context that the interlocutors do not explicate, but implicitly accommodate" (Schlöder et al., 2016, p.4). So, enthymemes can be nested: a reason provided for one inference is itself an unstated premise in a superordinate enthymeme, as it itself presupposes the application of another topos. Nesting of enthymemes may be useful in the examination of the role of RQs in arguments. The duality of RQs as interrogatives and statements makes it possible for a speaker to answer their own RQ, or to reply with an RQ to their own statements, which allows them to provide backing for the proposition implicit in the RQ, or use RQs to a reject a proposition they made. In the following example, 1 poses two RQs, that can be glossed as the inference "She has no special quality. (So) there is no reason we want to keep her"

(4)

1 So then we have the pregnant woman, so it's two people in one.
2 yeah.
1 Wh- what's her special quality? Why do we want to keep her at all?
2 Well, if you threw her out, maybe the pilot might well go mad, through losing his wife and his child.
3 But if you threw her out, maybe the pilot might jump out as well.
2 Yeah.
3 Hence, then you'd have two spaces left in the balloon. So you wouldn't have to throw anyone. (GP08, 68-78)

By asking a general question after the fact of stating something that could serve as its answer, 1 implicitly denies that their preceding statement is relevant as an answer.

1.4 Research questions

We report a preliminary study to explore the following questions:

1. Can RQs express enthymematic arguments, or parts of them?

2. Can their use make the warranting topoi likelier to be accommodated by participants, or make the topos that would warrant the enthymematic argument more acceptable?

3. Is the expression of enthymemes through RQs (as well as the structural correspondence between enthymemes and RQs, and their argumentative power) linked to an RQ having the illocutionary force of the statement it implies?

2 Method

2.1 Participants

The corpus, from Lavelle et al. (2013), consists of 40 triadic conversations of approximately five minutes. There are 20 interactions involving one patient with a diagnosis of schizophrenia and two non-psychiatric controls who were unaware of the patient's diagnosis. The 20 control interactions, each involved three healthy participants. Participants within each triad were unfamiliar to each other. This preliminary study focuses on the transcripts from 4 dialogues; 2 including a patient and 2 controls.

2.2 Task

The subjects discussed the balloon task – a moral dilemma which requires participants to reach agreement on which of four passengers should be thrown out of a hot air balloon that will otherwise crash, killing all the passengers, if one is not sacrificed. The choice is between Dr Robert Lewis – a cancer research scientist, who believes he is on the brink of discovering a cure for cancer; William Harris – the balloon pilot who is the only passenger with any balloon flying experience; Susanne Harris – William's wife, a primary school teacher who is 7 months pregnant with their second child; Heather Sloan – a nine-year old musical child prodigy who is considered by many to be a "twenty-first century Mozart". This task is known to elicit dialogues containing extended reasoning sequences.

2.3 Annotation

In order to capture as many borderline cases as possible, the criteria we use for what questions are regarded as rhetorical are simply those questions that, taking context into consideration, do not expect informational answers (as far as can be deemed likely from a non-participants' point of view), including cases where there is a probability of a non-rhetorical reading, or where a rhetorical question is responded to as an ordinary one. The reason for this is that in many cases the likelihood of a rhetorical contra informational reading varies, and since we worked with transcripts only, this likelihood can not be determined without prosodic and nonverbal cues.

3 Results

In these 4 dialogues we identified 19 RQs. 6 of these were regular Y/N RQs, 9 were regular Wh-RQs, and 4 were irregular RQs. We will discuss these types in turn, with examples, below.

3.1 Yes/No RQs

Is she really gonna make a difference?

(5) (GP12, 36-39)

1: Are we all agreed that the kid's not going?
2: erm.
1: She's nine years old, she's so light anyway **– is she really gonna make a difference?**
3: Well I'm not throwing a kid out, I just couldn't cope with it.

The RQ "is she really gonna make a difference?" has the illocutionary force of a statement with a negative polarity "She isn't really gonna make a difference", and it expects only negative answers like an ellipsis of the implied statement: "She is not". The rhetorical reading is motivated by the premise of the implied statement: "She's nine years old, she's so light anyway" provided by the speaker, by the modal adverb really, and anyway connecting the premise to the conclusion implicit in the RQ. The entire argument can be glossed as "She isn't really gonna make a difference [if thrown out], [because] she's nine years old, she's so light anyway", and the chain of reasoning can be represented as the enthymeme in (2), repeated here as(6).

(6) she's nine years old she's too light
throwing her out won't make a difference

Later in the dialogue, when asked for a reason to throw out the child, 1 repeats the argument that the child is too light, even in the case of morbid obesity, and covers the possibility of the child being an exception to a general notion of nine year olds being lightweight. The tag question in the last line is a yes/no-RQ implying a statement of the opposite polarity of the tag (and identical to the statement part of the tag question).

(7) (GP12, 204-213)

1: No no if that kid was a trouble maker
3: *laughter* No *laughter*
1: I would throw them out
3: *laughter* No I i- it's just ethically I I ca- I can't make that choice.
2: Why?
1: I just think the child is too too light anyway I mean, even if the child was morbidly obese.
3: *laughter*
1: **They're not gonna be as heavy as a sandbag, are they?** So.

"They're not gonna be as heavy as a sandbag" evokes a topos more specific than the previous one, defining the range of being heavy enough as at least the equal weight of a sandbag (8)

(8) x is not as heavy as a sandbag
throwing x out won't make a difference

Don't you think that p?

(9) (GP12, 96-112)

2: Yeah but the big question is if you throw the pilot out is what to expect, are you expected to be able to land the thing safely.

3: mmm.

2: Because if not then it's pointless throwing the pilot out. Because you kill everybody then.

3: Yes. But there is a chance

1: Don't you think that if she's been married to him she might have a little bit of piloting?

3: Yeah, exactly.

1: She might have been on a hot air balloon more than once.

3: Yeah.

1: So she might sort of know the general idea of how to land one.

The RQ implies the conditional statement: if "she's been married to him" then "she might have a little bit of piloting", evoking the topos in (10)

(10) x is married to a pilot ⇝ x has experience of piloting

In this example the speaker makes an argument relevant to the discourse through the use of an RQ to introduce an enthymeme, and further explicates their reasoning (by drawing on implicit topoi: i) that pilot's wives come along on flights sometimes; ii) that going on flights gives one piloting experience and iii) that piloting experience generally includes ability to land the aircraft.

It can be further noted that the introductory expression "don't you think that . . . " turns a statement into an RQ (whose implication can be derived by removing the introductory expression, in a similar way to a sentence final question tag). The RQ can be glossed as "Surely you think that if she's been married to him she might have a little bit of piloting". The gloss can explain the persuasive power of the RQ – why it expects (and in this case, receives) an affirmation for an answer. The introductory "don't you think that . . . " lays bare an emblematic property of RQs to make it likelier that the addressee will mirror the speaker's thinking process and agree with them.

3.2 Why RQs

Who needs a pilot?

(11) (GP12, 113-121)

1: But the scenario still says it's gonna crash. There's nothing, they can't do anything to land it. It's gonna crash. It's got to the point where they've actually thrown the food out, thrown the sandbags. Fully prepared that it's gonna crash, there's no way to land it.

2: mmm

1: So it's gonna crash, **who needs a pilot?**

3: mmm

We can gloss the RQ as "No one needs a pilot", and the whole utterance as "If the balloon is gonna crash, then no one needs a pilot". This evokes a topos like (12) stating that if a balloon is doomed to crash and it has passengers, then no one who is a passenger needs a pilot.

(12) x is a passenger of a balloon doomed to crash ⇝ x doesn't need a pilot

Imagine the RQ "do they really need a pilot?" instead of the one above. It is still drawing on the same topos as "who needs a pilot?", but it would be more dependent on it being assumed by other participants.

What's her special quality? With two wh-RQs in succession, responding to their own statement about the pregnant woman, 1 is conveying the idea that there does not exist a special quality about her, and so there exists no reason to keep her:

(13) (GP08, 68-78)

1: So then we have the pregnant woman, so it's two people in one.

2: yeah.

1: Wh- what's her special quality? Why do we want to keep her at all?

2: Well, if you threw her out, maybe the pilot might well go mad, through losing his wife and his child.

3: But if you threw her out, maybe the pilot might jump out as well.

2: Yeah.

3: Hence, then you'd have two spaces left in the balloon. So you wouldn't have to throw anyone.

The rhetorical reading is due to the presence of the NPI "at all" in the second of the RQs. We can see that they together make an inference when glossed as statements of non-existence: "She has no special quality. (So) there is no reason we want to keep her at all". 1 is drawing upon a notion relevant to the situation described in the balloon task, that a special quality needs to be found for an individual that should be saved.

The locutionary act of asking for an instantiation of this topos seemingly contrasts with 1's previous turn where they ascribed the pregnant woman a quality of counting as two (or, her death being equal to two). This explains why 2 and 3 in the following turns choose to give the RQs informational answers: to provide a reason for why being pregnant/counting as two lives counts as a special quality. Since the RQ implies a null set, the quality that 1 mentioned preceding it isn't found among answers to the inquiry of what her special quality may be, that would motivate saving her. It appears that the RQ allows 1 to reject that being pregnant and counting for two is applicable as a reason to be saved.

Who listens to classical music? The following exchange 3 is arguing for throwing out the child musical prodigy:

(14) (GP10, 54-58)

3: I think they should dash the child
1: *laughter*
3: It's just a child
1: The prodigy, nooo
3: **Who listens to classical music?**

3 expresses the standpoint that the child should be thrown out, because she is "just" a child, to which 1 objects when referring to the child as prodigy, as a reason to not throw her out. 3 follows up with the RQ "Who listens to classical music?" implicitly stating a hyperbolic "No one listens to classical music", to reject the notion that being a musical prodigy is a quality worth saving her for, as she is a prodigy in classical music. In other words, 3 draws on another, more specific topos than the one that warrants 1's protest. Let's assume 1 finds musical prodigies worth saving in general, as they make great music (15).

(15) $\dfrac{\text{x makes great music}}{\text{x should be saved}}$

Then, 3's RQ triggers the availability of the topos in (16), that if no one likes classical music, and someone is making classical music, they aren't making great music. In other words, the RQ invokes another topos as a reason for why (15) is unfounded.

(16) $\dfrac{\text{no one likes classical music} \qquad \text{x makes classical music}}{\text{x does not make great music}}$

How difficult is it to fly the balloon? A common theme in many arguments in the dialogues is that balloons are easy to fly, since operating its propane valve seems like a binary operation – either open or close it.

(17) (GP08 145-149)

1: **How difficult is it to fly the balloon?**
3: He could train the Mozart.
1: It's just going up and down.

The argument 1 makes is that a task that consists of only two modes of action is not in the upper range for what is complicated, and evokes a topos that delimits the range for what is to be considered a difficult task (analogous to the pragmatic scales that Rohde (2006) describes as being made salient in the context by the RQ). This can be glossed as evoking the topos that if something doesn't have many options then it is not very difficult.

It can be said then, that the topos drawn on can be treated as a generalisation of the contextually relevant property of elements in the relevant range of expected answers to the RQ. The same can be observed in the excerpt below.

(18) (GP10, 34-47)

2: **How hard is it to, um, navigate a balloon?**
3: *laughter* I don't know *laughter*
2: *laughter* Yeah *laughter*
1: *laughter* Exactly, that's what I was thinking, yeah *laughter*
2: You let hot air in and when when you wanna go you let hot air out.
1: Yeah, it is common sense I suppose.

Below we see 1 employing three RQs drawing on the same idea of a balloon not being difficult to fly, providing additional grounds to throw out the pilot – that flying the balloon can be easily taught.

(19) (GP08, 167-172)

1: **The thing is, how easy, or difficult it is to actually teach how to fly a balloon?** I mean it's just two things really. **What does a pilot do? It's not like flying a Boeing 727 is it?**

2: Well, yeah it is just two things like.

3: But if the balloon was sinking anyway, you wouldn't wanna train anyone, you'd just wanna jump out.

The first RQ, glossed as "It is not difficult to teach how to fly a balloon" implies the consequent of a similar topos to that evoked in the previous two examples, with the subsequent comment ("it's just two things") making up the antecedent.

The second RQ expects an empty set as the answer, that is the absence of what is difficult. "There is nothing [difficult] a [balloon] pilot does".

The implied statement of the third RQ is derived from simply eliminating the question tag "It's not like flying a Boeing 727." To make sense of this argument being presented here, the additional topos that "flying a Boeing 727 is difficult" needs to also be accommodated.

Of the three RQs only the first expects answers on the low end of a scale of difficulty. The second two imply cut-off thresholds for the scale, that the difficulty of flying a balloon can not exceed. The second one, a wh-RQ, implies that the whatever a balloon pilot does it is not difficult, or rather, its difficulty is so negligible it is below what can considered as such. The third, a tag yes/no-RQ, characterises the high end of the scale for piloting aircraft, by placing a passenger plane at that end of the scale.

3.3 Irregular cases

Is he gonna be kind of generous about it or is he gonna sell the cure?

(20) (GP12, 22-27)

1: There's always another doctor out there who is I'm I'm almost curing cancer but he hasn't really.

3: **And is he gonna be kind of generous about it or is he gonna sell the cure?**

2: Sell the drug to make tons of money

3: Yeah exactly it's

2: But but yeah those things apart I I I still think he's probably the most important person in in the balloon as he has the power the power to save lives all round the world from then on.

3: Or the power to make money.

"Is he gonna be kind of generous about it or is he gonna sell the cure?" is an exception in terms of the kind of argument participants usually made in the balloon task and in terms of form. The RQ is in disjunctive form, and a derivation of the implied statement can't be done by a shifting the polarity of the entire statement, as it would lead to a neither . . . nor construction, if we take A to represent "he is gonna be kind of generous" and B "he is gonna sell the cure", the RQ would imply a statement of the falsity of both A and B, which is obviously not the case. Instead consider that a rhetorical reading of the question stems from a reading of the disjunction as exclusive (also known as an either/or fallacy), where posing A leads to a negation of B, and posing B leads to a negation of A.

This doesn't mean that the RQ evokes two topoi at the same time, ("if x is generous then x won't sell the cure" and "if x sells the cure then x is not generous") rather that the implied statement is an enthymeme that is an instantiation of one of these topoi or the other, in this case, depending on context (in the general, also on utterance content, prosody and syntactic markers that force an RQ reading). Put differently, provided sufficient context is given as to if a rhetorical reading is obvious in the interaction, then it is obvious whether generous(d) or sell_cure(d) is presupposed – and which is negated in the consequent, as well as in the derivation of the implied statement of the RQ. Consider now the context of the exchange between 1 and 3 where they are providing reasons against saving the doctor, and note the introductory "and" in the RQ, connecting it to the previous utterance as yet another argument for not saving the doctor. In other words, the enthymeme implied by the RQ is given as a reason for why the doctor should not be saved. In light of the context, we can gloss the RQ as: "He isn't gonna be generous about it, because he is gonna sell the cure".

The topos drawn on (21) reflects the rhetorical emphasis on the corresponding part of the RQ, that the doctor is going to sell the cure – in the context its converse interpretation lacks rhetorical power, "He isn't gonna sell the cure, because he is gonna be generous about it" would not be applicable as a reason to not save the doctor. 2 disagrees with 3,

on the grounds that the power to save lives is more important than the notion of morality. 3 reiterates their argument, again drawing upon the topos in (21) since the power to make money is a proxy for not being generous.

(21) $\dfrac{\text{x is not generous}}{\text{x should not be saved}}$

It is another example of an either/or dilemma, like in the RQ, the emphasised alternative stands in exclusive disjunction with the power to save lives.

4 Conclusions

We have shown that RQs can express entire enthymemes, and either the antecedent or consequent of enthymematic inferences. In the latter case, the RQ is linked to another utterance in the surrounding discourse which serves as the other part of the syllogism. This link is enabled by the literal function of RQs as interrogatives. An enthymematic inference can be constructed by an RQ, the RQ can make up its antecedent, or more frequently, its consequent. Moreover, an RQ can provide a reason to invalidate the premises of a topos previously evoked, or make up an enthymeme by making salient the lack of concludable answers to the RQ.

The only major difference observed between yes/no-RQs and wh-RQs in this regard is that wh-RQs often serve as consequents in inferences. However, due to the limited amount of either kind of RQ in the data, this does not warrant any conclusion as to a fundamental difference between them in this regard. One thing that can be said for wh-RQs contra yes/no-RQs is that the wh-element introduces quantifiers in the statements implied, and by making general statements over groups of individuals having a property they introduce topoi in a more explicit way, whereas yes/no-RQs presuppose this implicitly. Again, drawing any hard conclusions in this matter is difficult due to the small amount of cases of RQs analysed.

An interesting phenomenon emerges when examining adjunct wh-RQs, such as how-RQs, conveying scalar implicatures. In these cases, the RQ implies an inference motivating the gradation of a property of an individual under discussion somewhere along a scale. This analysis gives a more detailed account for the chain of reasoning in such examples, than an approach dealing purely with the probability of distribution of answers to RQs in clusters on an extreme end of presupposed pragmatic scales made salient by the context (Rohde, 2006). The approach suggested here is also consistent with formal approaches to dialogue like KoS (Ginzburg, 2012), which enables an analysis of questions where the interpretation is very open-ended, as it is not purely denotational. However, more work needs to be done regarding the topoi evoked by such RQs, because the high variability of statements implied by them presupposes a very high amount of available topoi as warrants, and how mandated these are in the situation itself varies in relation to generalisations of pragmatic scales invoked (like balloons as easily pilotable aircrafts contra Boeings as difficult ones). More work also needs to be done in relation to how RQs function in regards to incrementally updating the state of evoked and accommodated topoi in the dialogue, especially in the case of how-RQs. The only certain conclusion that can be made in this matter as of now is that RQs are very frequently in use in interactional settings, and when used, are met with agreement, succeeding in the purpose of persuasion, the more common the topos they invoke is. More investigation can also be done on why the abundance of RQs in one dialogue contrast with the complete lack of them in another, and how this relates to how common the topoi drawn on are.

Acknowledgments

This work was supported by the Dialogical Reasoning in Patients with Schizophrenia (DRiPS) project funded by Riksbankens jubileumsfond (P16-0805:1). Howes and Breitholtz were additionally supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP).

References

Jean-Claude Anscombre. 1995. La théorie des topoi: Sémantique ou rhétorique? *Hermés*, 15:185–198.

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.).

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

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

Marijana Cerović. 2016. When suspects ask questions: Rhetorical questions as a challenging device. *Journal of Pragmatics*, 105:18–38.

Oswald Ducrot. 1988. Topoï et formes topique. *Bulletin d'Études de la Linguistique Française*, 22:1–14.

Markus Egg. 2007. Meaning and use of rhetorical questions. In *Proceedings of the 16th Amsterdam Colloquium*, pages 73–78. Universiteit van Amsterdam Amsterdam.

Jane Frank. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics*, 14(5):723–738.

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

H.P. Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3(S 41):58.

Chung-Hye Han. 1998. Deriving the interpretation of rhetorical questions. In *Proceedings of West Coast Conference in Formal Linguistics*, volume 16, pages 237–253. Citeseer.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 112(3):201–229.

Cornelia Ilie. 1994. *What else can I tell you? A pragmatic study of English rhetorical questions as discursive and argumentative acts.* Ph.D. thesis, University of Stockholm.

Mary Lavelle, Patrick GT Healey, and Rosemarie McCabe. 2013. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, 39(5):1150–1158.

Emily Norwood Pope. 1972. *Questions and answers in English.* Ph.D. thesis, Massachusetts Institute of Technology.

Hannah Rohde. 2006. *Rhetorical questions as redundant interrogatives*. Ph.D. thesis, UC San Diego: Department of Linguistics.

Julian Schlöder, Ellen Breitholtz, and Raquel Fernández. 2016. Why? In *Proceedings of JerSem*, pages 5–14.

John R Searle. 1975. Indirect speech acts. In *Speech acts*, pages 59–82. Brill.

The role of definitions in coordinating on perceptual meanings

Staffan Larsson
Centre for Linguistic Theory and Studies in Probabillity (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Box 200, SE 40530 Sweden
staffan.larsson@ling.gu.se

Abstract

This paper provides an account of how learning of (and coordination on) perceptual meaning can be initialised by partial definitions given in interaction, assuming that the words used in the definition themselves have perceptual meanings. In brief, the idea is that definitions provide a structure (a Naive Bayes classifier) connecting the defined concept with the concepts used in the definition. We formalise this account in Probabilistic Type Theory with Records, ProbTTR.

1 Introduction

Human first language learners typically learn from demonstrations, where a word becomes associated with a perceptual stimuli. This kind of semantic learning can be modeled as training a perceptual classifier on new perceptually available examples (Larsson, 2015, 2020). However, it also seems clear that at least adult humans can learn tentative new meanings, including perceptual meanings, from verbal descriptions.

This paper explores how learning of perceptual meaning can be initialised by partial definitions given in interaction, provided that the words used in the definition themselves have known perceptual meanings. In brief, the idea is that definitions provide hints on a structure (here, a Naive Bayes classifier) connecting the defined concept with the concepts used in the definition. The defined concept is an unobserved variable (for a classifier, the class variable), and the concepts used in the definition are evidence variables.

Semantic coordination, the process of interactively agreeing on the meanings of words and expressions, can be regarded as a process of reciprocal learning, where agents learn from each other. Semantic coordination can happen tacitly as a side-effect of dialogue interaction, or through more or less explicit discussion and negotiation of meanings of words and expressions – sometimes referred to as Word Meaning Negotiations (WMNs) (Myrendal, 2015; Noble et al., 2021).

An account of probabilistic inference and classification in ProbTTR is introduced in Larsson and Cooper (2021), where it is also demonstrated how probabilistic classification of perceptual evidence can be combined with probabilistic reasoning. Building on Larsson and Cooper (2021), Larsson et al. (2021) propose a probabilistic account of semantic learning from interaction formulated in terms of a Probabilistic Type Theory with Records (ProbTTR) (Cooper et al., 2014, 2015). Starting from a probabilistic type theoretic formulations of naive Bayes classifiers, the account of semantic learning is illustrated with a simple language game (the fruit recognition game).

In the following, we will connect these strands of work in an attempt to provide a formal account of the role of definitions in semantic coordination, and in particular for perceptual meanings. We first provide a brief overview of TTR and ProbTTR. We go on to review earlier work on probabilistic classification and learning from interaction using ProbTTR. Section 3 follows Larsson and Myrendal (2017) in relating dialogue acts involved in WMNs to semantic updates on an abstract level. The main contribution of this paper is Section 4, which explores the idea that the the dependency structure of Bayesian classifiers can be derived (learned) from definitions, and that one effect of a definition can be to update the structure of a Bayesian classifier. We provide examples of several ways in which this can happen in the context of a simple language game, the *fruit fetching game*. We end the paper with conclusions and future work.

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

$$\begin{bmatrix} \ell_1 & = & a_1 \\ \ell_2 & = & a_2 \\ \dots & & \\ \ell_n & = & a_n \\ \dots & & \end{bmatrix} : \begin{bmatrix} \ell_1 & : & T_1 \\ \ell_2 & : & T_2(l_1) \\ \dots & & \\ \ell_n & : & T_n(\ell_1, l_2, \dots, l_{n-1}) \end{bmatrix}$$

Figure 1: Schema of record and record type

$$\begin{bmatrix} \text{ref} & = & \text{obj}_{123} \\ \text{c}_{\text{man}} & = & \text{prf}_{\text{man}} \\ \text{c}_{\text{run}} & = & \text{prf}_{\text{run}} \end{bmatrix} : \begin{bmatrix} \text{ref} & : & \text{Ind} \\ \text{c}_{\text{man}} & : & \text{man(ref)} \\ \text{c}_{\text{run}} & : & \text{run(ref)} \end{bmatrix}$$

Figure 2: Sample record and record type

2 Background

This section reviews the background needed to follow the rest of the paper: TTR, Probabilistic TTR fundamentals, and Bayes nets and Naive Bayes classifiers.

2.1 TTR: A brief introduction

We will be formulating our account in a Type Theory with Records (TTR). We can here only give a brief and partial introduction to TTR; see also Cooper (2005) and Cooper (2012). To begin with, $s : T$ is a judgment that some s is of type T. One *basic type* in TTR is Ind, the type of an individual; another basic type is Real, the type of real numbers.

Next, we introduce *records* and *record types*. If $a_1 : T_1, a_2 : T_2(a_1), \dots, a_n : T_n(a_1, a_2, \dots, a_{n-1})$, where $T(a_1, \dots, a_n)$ represents a type T which depends on the objects $a_1, \dots, a_n$, the record to the left in Figure 1 is of the record type to the right.

In Figure 1, $\ell_1, \dots \ell_n$ are *labels* which can be used elsewhere to refer to the values associated with them. A sample record and record type is shown in Figure 2.

Types constructed with predicates may be *dependent*. This is represented by the fact that arguments to the predicate may be represented by labels used on the left of the ':' elsewhere in the record type. In Figure 2, the type of c_{man} is dependent on ref (as is c_{run}).

If r is a record and ℓ is a label in r, we can use a *path* $r.\ell$ to refer to the value of ℓ in r. Similarly, if T is a record type and ℓ is a label in T, $T.\ell$ refers to the type of ℓ in T. Records (and record types) can be nested, so that the value of a label is itself a record (or record type). As can be seen in Figure 2, types can be constructed from predicates, e.g., "run" or "man". Such types are called *ptypes* and correspond roughly to propositions in first order logic.

2.2 Probabilistic TTR fundamentals

In ProbTTR (as in TTR generally), situations are understood in a sense similar to that of Barwise and Perry (1983). It is also assumed that agents can individuate situations, and that they have a way of judging situations to be of situation types.

The core of ProbTTR is the notion of a probabilistic judgement, where a situation s is judged to be of a type T with some probability.

(1) $p(s : T) = r$, where $r \in$ [0,1]

Such a judgement expresses a subjective probability in that it encodes an agent's take on the likelihood that a situation is of that type.

A *probabilistic Austinian proposition* is an object (a record) that corresponds to, or encodes, a probabilistic judgement. Probabilistic Austinian propositions are records of the type in (2).

(2) $$\begin{bmatrix} \text{sit} & : & \textit{Sit} \\ \text{sit-type} & : & \textit{Type} \\ \text{prob} & : & [0,1] \end{bmatrix}$$

A probabilistic Austinian proposition φ of this type corresponds to the judgement that φ.sit is of type φ.sit-type with probability φ.prob.

(3) $p(\varphi.\text{sit}{:}\varphi.\text{sit-type}) = \varphi.\text{prob}$

We assume that agents track observed situations and their types, modelled as probabilistic Austinian propositions.

We use $p(T_1||T_2)$ to represent the probability that an agent assigns to some situation s being of type T_1, given that s is of type T_2. Note that $p(T_1||T_2)$, the conditional probability for some s of $s : T_1$ given that $s : T_2$, is different from $p(T_1|T_2)$, the probability of there being something of type T_1 given that there is something of type T_2. We refer to the former as the *bound variable* conditional probability, and the latter as the *existential* conditional probability.

Larsson and Cooper (2021) introduce a type theoretic counterpart of a random variable in Bayesian inference. To represent a single (discrete) random variable with a range of possible (mutually exclusive) values, ProbTTR uses a *variable type* V whose range is a set of *value types* $\mathfrak{R}(V) = \{A_1, \dots, A_n\}$ which are all (mutually disjoint) subtypes of V ($A_j \sqsubseteq V$ for $1 \leq j \leq n$).

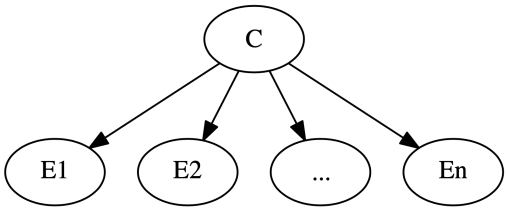


Figure 3: Evidence and Class in a Naive Bayes classifier

2.3 Bayesian nets and the Naive Bayes classifier

A Bayesian Network is a Directed Acyclic Graph (DAG). The nodes of the DAG are random variables, each of whose values is the probability of one of the set of possible states that the variable denotes. Its directed edges express dependency relations among the variables. When the values of all the variables are specified, the graph describes a complete joint probability distribution (JPD) for its random variables. Bayesian Networks provide graphical models for probabilistic learning and inference (Pearl (1990); Halpern (2003)).

A standard Naive Bayes model is a special case of a Bayesian network. More precisely, it is a Bayesian network with a single class variable C that influences a set of evidence variables $E_1, \ldots, E_n$ (the evidence), which do not depend on each other. Figure 2 illustrates the relation between evidence types and class types in a Naive Bayes classifier.

A Naive Bayes classifier computes the marginal probability of a class, given the evidence:

(4)

$$p(c) = \sum_{e_1,\ldots,e_n} p(c \mid e_1, \ldots, e_n)p(e_1)\ldots p(e_n)$$

where c is the value of C, e_i is the value of E_i $(1 \leq i \leq n)$ and

(5) $p(c \mid e_1, \ldots, e_n) =$

$$\frac{p(c)p(e_1 \mid c)\ldots p(e_n \mid c)}{\sum_{C=c'} p(c')p(e_1 \mid c')\ldots p(e_n \mid c')}$$

2.4 A ProbTTR Naive Bayes classifier

Corresponding to the evidence, class variables, and their value types, we associate with a ProbTTR Naive Bayes classifier κ:

(6) a. a collection of n evidence variable types $\mathbb{E}^\kappa_1, \ldots, \mathbb{E}^\kappa_n$

b. n associated sets of evidence value types $\mathfrak{R}(\mathbb{E}^\kappa_1), \ldots, \mathfrak{R}(\mathbb{E}^\kappa_n)$

c. a class variable type $\mathbb{C}^\kappa$, e.g. *Fruit*, and

d. an associated set of class value types $\mathfrak{R}(\mathbb{C}^\kappa)$

We can encode this as a TTR record as seen in Figure 4. (The function lbl takes a type T and returns a label unique to T.)

To classify a situation s using a classifier κ, the evidence is acquired by observing and classifying s with respect to the evidence types. Larsson and Cooper (2021) define a ProbTTR Bayes classifier κ as a function from a situation s (of the meet type[1] of the evidence variable types $\mathbb{E}^\kappa_1, \ldots, \mathbb{E}^\kappa_n$) to a set of probabilistic Austinian propositions that define a probability distribution over the values of the class variable type $\mathbb{C}^\kappa$, given probability distributions over the values of each evidence variable type $\mathbb{E}^\kappa_1, \ldots, \mathbb{E}^\kappa_n$. Formally, a ProbTTR Naive-Bayes classifier is a function

(7) $\kappa : \mathbb{E}^\kappa_1 \wedge \ldots \wedge \mathbb{E}^\kappa_n \rightarrow$

$$\text{Set}(\begin{bmatrix} \text{sit} & : & \textit{Sit} \\ \text{sit-type} & : & \textit{Type} \\ \text{prob} & : & [0,1] \end{bmatrix})$$

such that if $s : \mathbb{E}^\kappa_1 \wedge \ldots \wedge \mathbb{E}^\kappa_n$, then

(8) $\kappa(s) = \{ \begin{bmatrix} \text{sit} = s \\ \text{sit-type} = C \\ \text{prob} = p^\kappa(s : C) \end{bmatrix} \mid C \in \mathfrak{R}(\mathbb{C}^\kappa) \}$

2.5 Semantic classification in the fruit recognition game

Larsson and Cooper (2021) illustrate semantic classification using a Naive Bayes classifier in ProbTTR using the *fruit recognition game*. In this game a teacher shows fruits to a learning agent. The agent makes a guess, the teacher provides the correct answer, and the agent learns from these observations.

We use short-hands *Apple* and *Pear* for the types corresponding to an object being an apple or a pear, respectively[2]. Furthermore, we will assume that

[1] An object a is of the meet type of T_1 and T_2, $a : T_1 \wedge T_2$, iff $a : T_1$ and $a : T_2$.

[2] For details, see Larsson and Cooper (2021).

$$\upsilon(\kappa) = \left[\begin{array}{lcl} \text{cvar} & = & \mathbb{C}^\kappa \\ \text{cvals} & = & \mathfrak{R}(\mathbb{C}^\kappa) \\ \text{evars} & = & \{\mathbb{E}^\kappa_1, \ldots, \mathbb{E}^\kappa_n\} \\ \text{evals} & = & \{\left[\begin{array}{lcl} \text{lbl}(\mathbb{E}^\kappa_1) & = & \mathfrak{R}(\mathbb{E}^\kappa_1) \\ \ldots & & \\ \text{lbl}(\mathbb{E}^\kappa_n) & = & \mathfrak{R}(\mathbb{E}^\kappa_n) \end{array}\right] \end{array}\right]$$

Figure 4: Variables and values associated with a Naive Bayes classifier κ

$$\upsilon(\text{FruitC}) = \left[\begin{array}{lcl} \text{cvar} & = & \textit{Fruit} \\ \text{cval} & = & \{\textit{Apple}, \textit{Pear}\} \\ \text{evars} & = & \{\textit{Col}, \textit{Shp}\} \\ \text{evals} & = & \{\left[\begin{array}{lcl} \text{lbl}(\textit{Col}) & = & \{\textit{Red}, \textit{Green}\} \\ \text{lbl}(\textit{Shp}) & = & \{\textit{AShape}, \textit{PShape}\} \end{array}\right] \end{array}\right]$$

Figure 5: Variables and values associated with a Naive Bayes fruit classifier FruitC

the objects in the Apple Recognition Game have one of two shapes (a-shape or p-shape, corresponding to types *Ashape* and *Pshape*) and one of two colours (green or red, corresponding to types *Green* and *Red*).

The class variable type is *Fruit*, with value types $\mathfrak{R}(\textit{Fruit}) = \{\textit{Apple}, \textit{Pear}\}$. The evidence variable types are (i) *Col*(our), with value types $\mathfrak{R}(\textit{Col}) = \{\textit{Green}, \textit{Red}\}$, and (ii) Shape, with value types $\mathfrak{R}(\textit{Shape}) = \{\textit{Ashape}, \textit{Pshape}\}$.

For a situation s the classifier FruitC(s) returns a probability distribution over the value types in $\mathfrak{R}(\textit{Fruit})$.

(9) $$\text{FruitC}(s) = \{\begin{bmatrix} \text{sit} = s \\ \text{sit-type} = F \\ \text{prob} = p^{\text{FruitC}}(s : F) \end{bmatrix} \mid F \in \mathfrak{R}(\textit{Fruit})\}$$

We follow Larsson and Cooper (2021) in showing how semantic classification (i.e., estimating a probability distribution over class value types) works under the assumption that we can compute conditional probabilities $p(C_j||E_1 \ldots E_n)$ of a class value types C_j given evidence value types $E_1 \ldots E_n$.

In general, for $C_j \in \mathfrak{R}(\mathbb{C}^\kappa)$, we have

(10) $$p^\kappa(s : C_j) = \sum_{\substack{E_1 \in \mathfrak{R}(\mathbb{E}^\kappa_1) \\ \ldots \\ E_n \in \mathfrak{R}(\mathbb{E}^\kappa_n)}} \hat{p}^\kappa(C_j||E_1 \ldots E_n) p(s : E_1) \ldots p(s : E_n)$$

Correspondingly, in the fruit recognition game, for each $F \in \mathfrak{R}(\textit{Fruit})$ we have

(11) $$\hat{p}^{\text{FruitC}}(s : F) = \sum_{\substack{L \in \mathfrak{R}(\textit{Col}) \\ S \in \mathfrak{R}(\textit{Shape})}} p^{\text{FruitC}}(F||L \wedge S) p(s : L) p(s : S)$$

Larsson (2015) shows how perceptual classification can be modelled in TTR, and Larsson (2020) reformulates and extends this formalisation to probabilistic classification. Larsson and Cooper (2021) suggests regarding the non-conditional probabilities (e.g. $p(s : L)$ and $p(s : S)$ above) as resulting from probabilistic classification of real-valued (non-symbolic) visual input, where a classifier assigns to each image a probability that the image shows a situation of the respective type. Such a classifier can be implemented in a number of different ways, e.g. as a neural network, as long as it outputs a probability distribution. The training of perceptual classifiers are outside the scope of this paper, but see Larsson (2013); Fernández and Larsson (2014).

2.6 Semantic learning

For the model of semantic classification that uses conditional probabilities, a central question is of course how to estimate conditional probabilities, of the form $p(C||E_1 \wedge \ldots \wedge E_n)$ (where $C \in \mathfrak{R}(\mathbb{C}), E_i \in \mathfrak{R}(\mathbb{E}_i), 1 \leq i \leq n$). Using Bayes rule and marginalising over the class value types, we get for a Naive Bayes classifier:

(12) $$\hat{p}^\kappa(C||E_1 \wedge \ldots \wedge E_n) = \frac{p(C)p(E_1||C) \ldots p(E_n||C)}{\sum_{C' \in \mathfrak{R}(\mathbb{C}^\kappa)} p(C')p(E_1||C') \ldots p(E_n||C')}$$

For all combinations of evidence value types $E_1, \ldots, E_n$ and class value types C, we need (a) the conditional probability of the evidence value types given the class value type, $p(E_i||C)$, and (b) the prior of the class value type, $p(C')$.

We compute likelihoods and probabilities as ratio of the frequencies of occurrences, summed over all judgements in the history:

(13)

$$p(E_i||C) = \frac{\sum_{j \in \mathfrak{J}, j.\text{sit}=s} p(s:C)p(s:E_i)}{\sum_{j \in \mathfrak{J}, j.\text{sit}=s} p(s:C)}$$

The formula (13) tells us that we can consider probabilities in the history of judgements as fractions of events; and this is justified by interpreting them as fractions of language-community speakers making the corresponding categorical judgement. In this sense, we are providing a frequentist interpretation of epistemic probability. (For the full account and motivation, see [paper under review].)

In addition to conditional probabilities, (12) requires the prior probabilities of the class value types $C \in \mathfrak{R}(\mathbb{C})$. We use $\mathfrak{p}_\mathfrak{J}(T)$ to denote the prior probability that an arbitrary situation is of type T given $\mathfrak{J}$.

(14)

$$\mathfrak{p}_\mathfrak{J}(T) = \frac{\sum_{j \in \mathfrak{J}_T} j.\text{prob}}{\text{P}(\mathfrak{J})} \text{ if } \text{P}(\mathfrak{J}) > 0, \text{ otherwise } 0$$

where $\text{P}(\mathfrak{J})$ is the cardinality of situations in $\mathfrak{J}$, i.e. the total number of situations in $\mathfrak{J}$.

(15) $\text{P}(\mathfrak{J}) = |\{s|\exists j \in \mathfrak{J}, j.\text{sit} = s\}|$

We can encode the relevant conditional probabilities and priors as a TTR record $\pi(\kappa)$, as seen in Figure 6. Accordingly, we replace (12) with (16):

(16) $\hat{p}^\kappa(C||E_1 \wedge \ldots \wedge E_n) =$

$$\frac{\mathfrak{p}^\kappa_\mathfrak{J}(C)p^\kappa(E_1||C)\ldots p^\kappa(E_n||C)}{\sum_{C' \in \mathfrak{R}(\mathbb{C}^\kappa)} \mathfrak{p}^\kappa_\mathfrak{J}(C')p^\kappa(E_1||C')\ldots p^\kappa(E_n||C')}$$

where

$p^\kappa(E||C) = \pi(\kappa).\text{condps.lbl}(C).\text{lbl}(E)$

$\mathfrak{p}^\kappa_\mathfrak{J}(C) = \pi(\kappa).\text{priors.lbl}(C)$

What this buys us is the possibility of updating classifiers by manipulating records encoding them. In Section 4, we will exploit this in formulating semantic updates resulting from word meaning negotiations.

3 Word Meaning Negotiation and semantic updates

In Myrendal (2015, 2019), a taxonomy for dialogue acts involved in WMNs of so-called *trigger words* T in online discussion forum communication is presented. Two central dialogue acts are:

- Explicification: Provides an explicit (partial or complete) definition of T. We will here refer to this as simply *definition*.
- Exemplification: Providing examples of what the trigger word can mean, or usually means.

To describe the effects of these dialogue acts (once they are grounded), Larsson and Myrendal (2017) propose an abstract formalism for conceptual updates, where we assume that a definition D of a word (or expression) T has been provided, or an example situation E. D or E is then used for updating the meaning in question.

- $\delta(T, D)$: T updated with D as a partial definition of T
- $\epsilon(T, E)$: T updated with E as an example of a situation described by T

The abstract meaning update functions[3] serve as a sort of API between dialogue acts and their consequent meaning updates. We can see the learning from examples described above in Section 2.6 as part of the specification of ϵ(T, E). While we leave the exact formulation for future work, updating with an example E in the frequentist learning paradigm amounts to (1) adding example E to $\mathfrak{J}$, (2) recomputing the conditional probabilities and priors based on the updated $\mathfrak{J}$, and (3) updating the probabilities and priors in the classifier record. If we assume that P' is a record like the one shown in Figure 7 but with updated values based on $\mathfrak{J}$ updated with E, step (3) could be formalised thus (taking FruitR to be the union[4] of the records in Figures 5 and 7, so FruitR=υ(FruitC)$\cup\pi$(FruitC)):

(17) FruitR$'$=FruitR[P$'$]

Simplifying somewhat, if r_1 and r_2 are records, then $r_1[r_2]$ is the union of r_1 and r_2 except that if a label ℓ occurs in both r_1 and r_2, the value of ℓ in $r_1[r_2]$ will be $r_2.\ell$. See Cooper (in prep) for details.

[3] We ignore the polarity of the updates here; in general, definitions and examples can be positive or negative.

[4] Records are labelled sets.

$$\pi(\kappa)=\left[\begin{array}{lcl} \text{condps} & = & \left[\begin{array}{lcl} \text{lbl}(C_1) & = & \left[\begin{array}{lcl} \text{lbl}(E_1) & = & p(E_1||C_1) \\ \ldots & & \\ \text{lbl}(E_v) & = & p(E_v||C_1) \end{array}\right] \\ \ldots & & \\ \text{lbl}(C_w) & = & \left[\begin{array}{lcl} \text{lbl}(E_1) & = & p(E_1||C_w) \\ \ldots & & \\ \text{lbl}(E_v) & = & p(E_v||C_w) \end{array}\right] \end{array}\right] \\ \text{priors} & = & \left[\begin{array}{lcl} \text{lbl}(C_1) & = & \mathfrak{p}_{\mathfrak{J}}(C_1) \\ \ldots & & \\ \text{lbl}(C_w) & = & \mathfrak{p}_{\mathfrak{J}}(C_w) \end{array}\right] \end{array}\right]$$

Figure 6: Record containing conditional probabilities and priors for a classifier κ, where for $1 \leq u \leq v, E_u \in \mathfrak{R}(\mathbb{E}_1^\kappa) \cup \ldots \cup \mathfrak{R}(\mathbb{E}_n^\kappa)$, and where for $1 \leq u \leq w, C_u \in \mathfrak{R}(\mathbb{C}^\kappa)$

$$\pi(\text{FruitC}) = \left[\begin{array}{lcl} \text{condps} & = & \left[\begin{array}{lcl} \text{lbl}(\textit{Apple}) & = & \left[\begin{array}{lcl} \text{lbl}(\textit{Red}) & = & 0.63 \\ \text{lbl}(\textit{Green}) & = & 0.37 \\ \text{lbl}(\textit{AShape}) & = & 0.97 \\ \text{lbl}(\textit{PShape}) & = & 0.03 \end{array}\right] \\ \text{lbl}(\textit{Pear}) & = & \ldots \end{array}\right] \\ \text{priors} & = & \left[\begin{array}{lcl} \text{lbl}(\textit{Apple}) & = & 0.64 \\ \text{lbl}(\textit{Pear}) & = & 0.26 \end{array}\right] \end{array}\right]$$

Figure 7: Parts of record containing conditional probabilities and priors for the fruit classifier

4 Learning perceptual meanings from definitions

The work reviewed above showed how probabilistic classifiers can be trained from examples presented in interaction. However, this cannot be the whole story. Indeed, in terms of the dialogue acts for semantic coordination presented in Larsson and Myrendal (2017), we have only covered exemplification. What about partial definitions (explicifications)? What effect do they have on agent's takes on meanings, and how is learning from definitions related to learning from examples?

From the perspective of agents learning how to classify situations probabilistically, one might ask how agents learn the structure of the Bayes net (or as a special case, Naive Bayes classifier) used to classify situations. We propose to connect these two questions, by exploring the idea that the the dependency structure of Bayesian classifiers can be derived (learned) from definitions, and that one effect of a definition can be to update the structure of a Bayesian classifier. (We are not claiming that this is the only way agents can learn such structures.)

In the fruit recognition game, B learns how to take shape (a-shape or p-shape) and colour (red or green) into account when classifying apples and pears, by adjusting conditional probabilities and priors. Before going into learning new meanings from definitions, it might be helpful to show how learning new meanings from examples (demonstrations) could be accounted for.

4.1 Learning a new meaning from example

In Larsson and Cooper (2009), it is shown how *ontological* meaning (e.g. that kumqat is a type of fruit) can be learned from interaction, and how such learning can be modelled in TTR. We can imagine a version of the fruit recognition game where new fruits (i.e., new value types for the fruit variable type) are introduced by demonstration:

A: What fruit is this?
B: A pear.
A: Wrong, it's a Wax Jambu.
B: Okay.

In this example, B can learn both that Wax Jambus are fruits, and what they look like based on being provided with an example Wax Jambu that they can observe. From the context, B can figure out that Wax Jambus are fruits. In the general case such an inference can be based on a variety of factors, including the ongoing activity and linguistic evidence. In terms of a probabilistic classifier, learning this amounts to adding a new value (type) to the Fruit variable (type). This update can be formalised thus:

(18) FruitR$'$=FruitR[$\left[\text{cvals} = F.\text{cvals} \cup \{C\}\right]$]

Furthermore, B could add the new example to $\mathfrak{J}$ and update the conditional probabilities, as detailed above.

In the following, we will see how B can instead learn from partial definitions, which do not provide

perceptually available evidence but do seem to offer help in guiding B's learning of the structure of the classifier, as well as associated probabilities.

4.2 Learning a new meaning from definition

We can imagine another language game where A asks B to fetch different fruits in a fruit storage, where several types are fruits are available, some of them unknown to B:

A: Get me an apple please
B: (fetches apple) there you go
A: Thanks. Now get me a Wax Jambu!
B: A Wax Jambu?
A: They are pear-shaped and red.

We can call this the *fruit fetching game*. Let's assume that our learning agent B from A's second utterance learns that Wax Jambus are fruits. However, B has not been presented with an example fruit to use for training. In this sense B does not yet know what Wax Jambus look like. It seems plausible that B in this case might be able to use A's definition of Wax Jambu to distinguish Wax Jambus from other fruits (even if this ability will not be as developed as it might later be after seeing several Wax Jambus).

How, then, could we model the effects of A's definition, which (with pronoun resolved) can be paraphrased as "Wax Jambus are pear-shaped and red"? Firstly, by adding a value type *WaxJambu* to the fruit classifier:

(19) FruitR′=FruitR[
$\left[\text{cvals} = \text{FruitR.cvals} \cup \{\mathit{WaxJambu}\}\right]$]

Secondly, by recomputing probabilities, assigning high values to $p(\mathit{Pshape}||\mathit{WaxJambu})$ and $p(\mathit{Red}||\mathit{WaxJambu})$, and lowering other probabilities accordingly. For simplicity, we assume here that the high values are 1 and that conditional probabilities for other values of the same variables are lowered to 0.

(20) FruitR″=FruitR′[
$$\left[\text{condps}=\left[\text{lbl}(\mathit{WaxJambu})=\begin{bmatrix}\text{lbl}(\mathit{PShape}) = 1.0\\ \text{lbl}(\mathit{AShape}) = 0.0\\ \text{lbl}(\mathit{Red}) = 1.0\\ \text{lbl}(\mathit{Green}) = 0.0\end{bmatrix}\right]\right]$$
]

Hence, the ProbTTR implementation of the $\delta^+(T, D)$ function should be such that δ^+(⟦Wax Jambu⟧,⟦pear-shaped and red⟧) results in these updates.

Equipped with the updated mental fruit classifier, B now goes off to fetch a Wax Jambu in a storage room, despite never having seen one. One way of finding the right type of fruit in the storage is to simply going through the fruits in storage one by one and classify them, until one is classified as the sought type (here, Wax Jambu)[5].

4.3 Learning new evidence values

Above, A's definition only included evidence values that were already used in the fruit classifier. However, A may also introduce a unknown value previously unknown to B:

A: Get me a Mango please!
B: A Mango?
A: They have an oblong shape.

In this case, B needs to both add the new class value *Mango* and a new evidence value *Oblong* (for the variable *Shp*):

(21) FruitR′=FruitR[
$\left[\text{values} = \text{FruitR.cvals} \cup \{\mathit{Mango}\}\right]$]

(22) FruitR″=FruitR′[
$$\left[\text{evals} = \begin{bmatrix}\text{lbl}(\mathit{Shp}) = \text{FruitR.evals.lbl}(\mathit{Shp})\\ \cup\ \{\mathit{Oblong}\}\end{bmatrix}\right]$$
]

Finally, as before, the conditional probabilities are shifted to favour the evidence variable given in the definition:

(23) FruitR‴=FruitR″[
$$\left[\text{condps}=\left[\text{lbl}(\mathit{Mango}) = \begin{bmatrix}\text{lbl}(\mathit{PShape}) = 0.0\\ \text{lbl}(\mathit{AShape}) = 0.0\\ \text{lbl}(\mathit{Oblong}) = 1.0\end{bmatrix}\right]\right]$$
]

We assume here that B was familiar with the shape value type *Oblong*, but had not previously considered it relevant to fruit classification[6]A more complicated situation arises when a previously unknown value for a known variable is introduced, e.g. a new shape. In such cases, perceptually available

[5]One can imagine a continuation of the game, where B shows the retrieved fruit to A and receives feedback on whether it was right kind of fruit or not, and trains on this example in the normal way.

[6]An agent may know a very high number of shapes but not all of them will be relevant to classifying fruits. For such reasons, one might consider separating a general shape classifier (if such a thing is ever needed) from a classifier-specific one (in this case, specific to the fruit classifier). In general, even many many evidence types are of a general character (e.g. shape, colour and size), generic classifiers may be of less use than evidence classifiers that are adapted to specific tasks.

examples may be necessary to train the updated classifier on.

4.4 Learning new evidence variables

Finally, a definition may introduce a new evidence variable, along with a value:

A: Get me a Kumqat!
B: A Kumqat?
A: They are small

We assume that B is already has a *Size* classifier and knows that *Small* is a *Size* (along with, say, *MidSize* and *Large*). Given this, the resulting updates to B's fruit classifier could be described thus:

(24) FruitR$'$=FruitR[
$\left[\text{values} = \text{FruitR.cvals} \cup \{\textit{Kumqat}\}\right]$]

(25) FruitR$''$=FruitR$'$[
$\left[\text{evars} = \text{FruitR.evars} \cup \{\textit{Size}\}\right]$]

(26) FruitR$'''$=FruitR$''$[
$\left[\text{evals} = \left[\text{lbl}(\textit{Size}) = \{\textit{Large}, \textit{MidSize}, \textit{Small}\}\right]\right]$]

(27) FruitR$''''$=FruitR$'''$[
$\left[\text{condps}=\left[\text{lbl}(\textit{Kumqat}) = \left[\begin{array}{l}\text{lbl}(\textit{Small}) = 1.0\\ \text{lbl}(\textit{Large}) = 0.0\\ \text{lbl}(\textit{Midsize}) = 0.0\end{array}\right]\right]\right]$]

This example also raises the question about partial definitions that only mention a value of one of the evidence variables. What should the conditional probabilities for a situation being of the value types for the other evidence variable types (not mentioned in the definition) given that the situation is of the new class value type? For now, we note that several options are available - assuming uniform distributions, or asking for more information ("What colour is a Kumqat? What shape?") and use the response to infer new conditional probabilities.

5 Definitions vs. examples

If we want to model how meanings are affected by both definitions and examples, we will need to say something about the trade-off between definitions and examples. For example, while a definition may be useful until examples have been observed, at some point the observed examples may override a definition. In the proposed account, definitions affect conditional probabilities only in the short run. Assuming conditional probabilities are recomputed when receiving new relevant observations, the probabilities resulting from proposed definitions (e.g. in the fruit fetching game) will be overwritten as soon as an observation of an instance of the defined concept has been made (an actual fruit of the defined type has been observed). This is perhaps not obviously wrong – it is at least theoretically possible that definitions are categorically superseded by observations – but a more flexible trade-off between definitions and examples (observations) would probably be desirable. There are ways of achieving this in the frequentist approach, e.g. by letting a definition lead to adding some relatively high number N of "fake" observations in line with the definition to $\mathfrak{J}$. By manipulation of N, the relative importance of definitions relative to observations can be regulated. If such approaches are deemed unsatisfying for theoretical or empirical reasons, it may be necessary to move to a different learning method. Future work thus includes working out alternative learning approaches that can better account for the trade-off between definitions and examples.

6 Conclusion

We have shown how (partial) definitions offered in word meaning negotiations can help learners structure probabilistic classifiers that are used to compute probabilistic semantic judgements. Technically, this was achieved by encoding a Naive Bayes classifier as a TTR record structure which can updated by definitions. Beyond what has been mentioned above, future work includes parsing natural language into an appropriate representation for updating classifiers, formulating a general update rule for carrying out such updates (of which several examples were given above), and generalising the account to Bayes nets (and other types of probabilistic classifiers). We also want to study actual definitions from human-human dialogues, rather than invented ones.

Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. Bradford Books. MIT Press, Cambridge, Mass.

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. Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Robin Cooper. in prep. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). Draft available from `https://sites.google.com/site/typetheorywithrecords/drafts`.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 72–79. Gothenburg, Association of Computational Linguistics.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology 10*, pages 1–43.

Raquel Fernández and Staffan Larsson. 2014. Vagueness and learning: A type-theoretic approach. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (∗SEM 2014)*.

J. Halpern. 2003. *Reasoning About Uncertainty*. MIT Press, Cambridge MA.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369. Published online 2013-12-18.

Staffan Larsson. 2020. Discrete and probabilistic classifier-based semantics. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 62–68, Gothenburg. Association for Computational Linguistics.

Staffan Larsson, Jean-Philippe Bernardy, and Robin Cooper. 2021. Semantic learning in a probabilistic type theory with records. In *Proceedings of Workshop on Computing Semantics with Types, Frames and Related Structures 2021*.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 1–9. EACL.

Staffan Larsson and Robin Cooper. 2021. Bayesian classification and inference in a probabilistic type theory with records. In *Proceedings of NALOMA 2021*.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue acts and updates for semantic coordination. *SEMDIAL 2017 SaarDial*, page 59.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. Ph.D. thesis, University of Gothenburg.

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

Bill Noble, Kate Viloria, Staffan Larsson, and Asad Sayeed. 2021. What do you mean by negotiation? annotating social media discussions about word meaning. In *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021)*.

J. Pearl. 1990. Bayesian decision methods. In G. Shafer and J. Pearl, editors, *Readings in Uncertain Reasoning*, pages 345–352. Morgan Kaufmann.

Generating Personalized Dialogue via Multi-Task Meta-Learning

Lee Jing Yang[1], Lee Kong Aik[2], Gan Woon Seng[3]
School of Electrical and Electronic Engineering, Nanyang Technological University[1,3]
Institute for Infocomm Research, A*STAR[2]
jingyang001@e.ntu.edu.sg[1], lee_kong_aik@i2r.a-star.edu.sg[2], ewsgan@ntu.edu.sg[3]

Abstract

Conventional approaches to personalized dialogue generation typically require a large corpus, as well as predefined persona information. However, in a real-world setting, neither a large corpus of training data nor persona information are readily available. To address these practical limitations, we propose a novel multi-task meta-learning approach which involves training a model to adapt to new personas without relying on a large corpus, or on any predefined persona information. Instead, the model is tasked with generating personalized responses based on only the dialogue context. Unlike prior work, our approach leverages on the provided persona information only during training via the introduction of an auxiliary persona reconstruction task. In this paper, we introduce 2 frameworks that adopt the proposed multi-task meta-learning approach: the Multi-Task Meta-Learning (MTML) framework, and the Alternating Multi-Task Meta-Learning (AMTML) framework. Experimental results show that utilizing MTML and AMTML results in dialogue responses with greater persona consistency.

1 Introduction

Personalized dialogue generation involves generating dialogue responses which incorporates the personality of the interlocutors, leading to more natural and human-like dialogue. Thus far, approaches to personalized dialogue generation typically require a large number of persona-specific dialogue examples. Certain approaches also require persona information presented in the form of several predefined persona statements (eg. 'I love dogs', 'I am an engineering student.'). However, in a real-world system, large amounts of persona-specific dialogue are rarely available, and collecting descriptive persona statements from every interlocutor is intractable.

To address these practical issues, Persona Agnostic Meta-Learning (PAML) (Madotto et al., 2019), a framework which aims to train a model capable of rapid adaptation to new unseen personas, was proposed. The PAML framework was based on the popular Model-Agnostic Meta-Learning (MAML) framework (Finn et al., 2017). The recently proposed Customized Model Agnostic Meta-Learning (CMAML) (Song et al., 2020b) framework largely follows the PAML framework, with the exception of an additional network structure optimization component. Both the PAML and CMAML frameworks were benchmarked on the PersonaChat corpus (Zhang et al., 2018), a popular personalized dialogue generation corpus which provides persona statements describing each interlocutor in addition to the persona-specific dialogues. As it is unfeasible to collect persona statements from interlocutors in a real world setting, the PAML framework does not utilize the available persona statements during both meta-learning and inference. However, even though it is impractical to utilize the persona statements during inference, the persona statements can be used *during meta-learning* to further improve model performance.

Hence, we introduce a novel multi-task meta-learning approach which leverages predefined persona statements *only during meta-learning* via an additional persona reconstruction task. Essentially, this task involves generating all corresponding persona statements in its entirety given the dialogue context. We hypothesize that the introduction of the persona reconstruction task would result in parameters capable of effectively inducing the persona information from the dialogue context, which would lead to the generation of persona consistent dialogue. The persona statements are *not* used during inference. Prior usage of multi-task learning for personalized dialogue generation involved the addition of a persona classification task (Yang et al.,

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

2021; Su et al., 2019a) and a distraction utterance binary classification task (Na et al., 2021). To our knowledge, this is the first attempt at incorporating persona statement reconstruction.

Our contributions include 2 multi-task meta-learning frameworks which leverage the persona reconstruction task only during training. The Multi-Task Meta-Learning (MTML) framework, as well as a variant known as the Alternating Multi-Task Meta-Learning (AMTML) framework. While both MTML and AMTML involve the addition of a persona reconstruction task only during meta-learning, MTML involves combining the losses derived from generating the response and reconstructing the persona. AMTML, on the other hand, functions by constantly alternating between both tasks. Experimental results on the PersonaChat corpus reveal that utilizing MTML and AMTML result in responses which reflect the interlocutor's persona to a larger extent compared to prior work.

2 Methodology

Our approach to personalized dialogue generation involves extending the PAML framework (Madotto et al., 2019) by introducing a persona reconstruction task only during the meta-learning stage. The PAML framework is essentially an adaptation of the general MAML framework for personalized dialogue generation. The MAML framework involves learning a parameter initialization capable of generalizing and adapting rapidly to new tasks unseen during the training process via gradient descent. Specifically, this involves obtaining the updated parameters by adjusting the model parameters by utilizing data from several related tasks. The original parameters are then optimized based on the updated parameters and a separate test set by computing the second order derivatives (Hessian matrix). In the PAML framework, each persona is viewed as a unique task. Consequently, the goal of the PAML framework is to obtain a parameter initialization capable of rapidly adapting to new personas. We hypothesize that the introduction of the persona reconstruction task would result in parameters which induce the persona from the dialogue context to a larger extent.

2.1 Multi-task Learning

Unlike traditional approaches that involve utilizing both the dialogue context and persona information as model inputs during training, our framework involves reconstructing the persona statements as well as generating the dialogue response given the dialogue context. The primary task of generating the response and the corresponding loss function L_{res} can be expressed in the following equations:

$$f_\phi(x_t|x_{1:t-1}) = p(x_t|x_{1:t-1};\phi) \quad (1)$$

$$L_{res}(\phi) = -\sum_{t=1}^{T} log\, p(x_t|x_{1:t-1};\phi) \quad (2)$$

where x_t and $x_{1:t-1}$refer to the response and dialogue context respectively, and ϕ represents the model parameters. We hypothesize that the persona reconstruction task would result in model parameters capable of inducing the persona information from the dialogue context to a larger extent. This is due to the increased emphasis on persona consistency in the task of persona reconstruction. Also, rather than generating only selected keywords or phrases, the persona reconstruction task involves generating all corresponding persona statements in its entirety given the dialogue context. This is because generating complete sentences would also require the model to account for fluency in addition to persona consistency, which is also a vital aspect of dialogue generation. The persona statements $\mathcal{P}_{1:N}$ are concatenated to form a single sequence $\overline{\mathcal{P}}$. The auxiliary persona reconstruction task and the corresponding loss function L_{rec} is formalized in the following expression:

$$\overline{\mathcal{P}} = concat(\mathcal{P}_{1:N}) \quad (3)$$

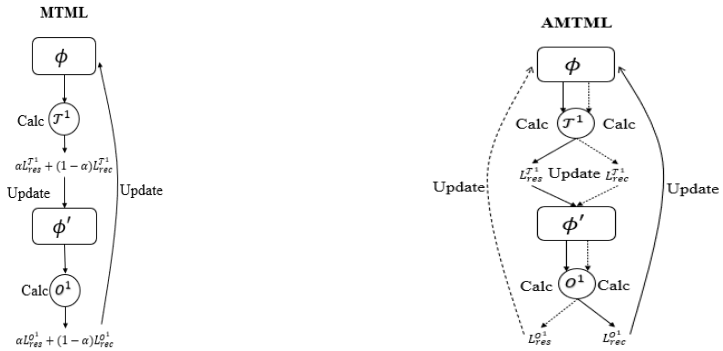
$$f_\phi(\overline{\mathcal{P}}|x_{1:t-1}) = p(\overline{\mathcal{P}}|x_{1:t-1};\phi) \quad (4)$$

$$L_{rec}(\phi) = -\sum_{t=1}^{T} log\, p(\overline{\mathcal{P}}|x_{1:t-1};\phi) \quad (5)$$

During training, the persona reconstruction loss and the response generation loss will be weighted and summed. Hence, the total multi-task loss is expressed as:

$$L(\phi) = \alpha L_{res}(\phi) + (1-\alpha)L_{rec}(\phi) \quad (6)$$

where α determines the contribution of the persona reconstruction and response generation loss to the learning process.



(a) (b)

Figure 1: Diagrams depicting the computation path for MTML and AMTML. A batch size of 1 is assumed. (a): Diagram depicting the MTML framework. The weighted sum of the persona reconstruction loss and the response generation loss is used to update the parameters ϕ. (b): Diagram depicting the AMTML framework. The computation flow alternates between the solid arrows and the dotted arrows after every iteration.

Persona	
i have three guns and love hunting	
my family lives down the street from me	
i go to church every sunday	
i drive a ford pickup truck	
i am very conservative	
Dialogue Context	
User1:	hey there how are you i am shy
User2:	well i am conservative so sounds like a match to me
User1:	i like that , did you go to college ?
User2:	no i just got my high school diploma and opened my owns shooting range
User1:	i went an got a computer science degree
User2:	wish i had continued school but i love what i do hunting , going to church and driving
User1:	gotcha , i build spaceships , the models
User2:	o ok my family lives down from me and they build garages and farms
User1:	that is cool , my mom is a doctor
Responses	
Ref	you ever eat at wendys ? i went to washington school before i quit .
	10-shot
Std	that is cool where are you from ?
Std_p	that is cool what do you do for work ?
$PAML$	he is good to church every week
$MTML_{0.8}$	sure , i pick him up for church every sunday with my ford pickup
$AMTML$	my parents live in the house i grew up in , just down the street .
	5-shot
Std	that is cool what kind of doctor do you do ?
Std_p	that is cool what kind of garages do you drive ?
$PAML$	that is good where do you live ?
$MTML_{0.8}$	that is good what do you do for work ?
$AMTML$	oh okay lol what do you do for work ?

Table 1: Dialogue responses generated by the implemented models.

2.2 Multi-Task Meta Learning Framework

Similar to PAML, MTML aims to learn a set of parameters ϕ capable of quickly adapting to an unseen persona in order to generate contextually appropriate responses which reflect the persona, without relying on user persona statements. However, unlike PAML, which does not utilize the persona statements during both meta-learning and inference, we leverage the persona information *only during meta-learning*. The persona statements are not used during inference. The persona information is incorporated during meta-learning via the multi-task learning (Section 3.1).

We begin by dividing the corpus into a training set, validation set and a test set denoted by $\mathcal{D}_{train}$, $\mathcal{D}_{valid}$ and $\mathcal{D}_{test}$ respectively. For each iteration i, a batch distinct of personas $\mathcal{P}_{1:N}^{1:M}$ and a corresponding set of M dialogues are randomly sampled from the training set $\mathcal{D}_{train}$ to form the meta training set $\mathcal{T}^i$ (support set). Then, another set of M dialogues corresponding to the same batch of personas $\mathcal{P}_{1:N}^{1:M}$ are randomly sampled from the training set $\mathcal{D}_{train}$ to form the meta optimization set $\mathcal{O}^i$ (query set). We refer to the meta training set and meta optimization set collectively as $\mathcal{C}^i$ i.e., $\mathcal{C}^i$ = ($\mathcal{T}^i$, $\mathcal{O}^i$).

During meta-training, the multi-task loss $L_{mul}^{\mathcal{T}^i}$ is computed via a weighted sum between the response generation loss $L_{res}^{\mathcal{T}^i}$ and the persona recreation loss $L_{rec}^{\mathcal{T}^i}$. To calculate $L_{rec}^{\mathcal{T}^i}$, the persona statements $\mathcal{P}_{1:N}^{1:M}$ of each persona were concatenated into a single sequence, resulting in $\overline{\mathcal{P}}^{1:M}$. For an arbitrary persona in the meta training set, the equations for computing the multi-task loss are as follows:

$$L_{res}^{\mathcal{T}^i}(\phi) = -\sum log\, p(x_t, |x_{1:t-1}; \phi) \quad (7)$$

$$\overline{\mathcal{P}}^{1:M} = concat(\mathcal{P}_{1:N}^{1:M}) \quad (8)$$

$$L_{rec}^{\mathcal{T}^i}(\phi) = -\sum log\, p(\overline{\mathcal{P}}|x_{1:t-1}; \phi) \quad (9)$$

$$L^{\mathcal{T}^i}(\phi) = \alpha L_{res}^{\mathcal{T}^i}(\phi) + (1-\alpha) L_{rec}^{\mathcal{T}^i}(\phi) \quad (10)$$

where α accounts for the distribution between the 2 losses. Subsequently, the parameters ϕ are updated via SGD. The updated model parameters ϕ' can be expressed as:

$$\phi' = \phi - \eta_t \nabla_\phi L^{\mathcal{T}^i}(\phi) \quad (11)$$

where η_t refers to the inner loop learning rate and $L^{\mathcal{T}^i}(\phi)$ represents the multi-task training loss attained.

During meta-optimization, the original model ϕ is optimized on the multi-task loss attained on the meta optimization set $\mathcal{O}_i$ and the updated parameters ϕ'. This meta-objective function can be expressed as:

$$\begin{aligned} &\min_{\phi} \sum_{\mathcal{C}^i \sim \mathcal{D}_{train}} L^{\mathcal{O}^i}(\phi') \\ &= \sum_{\mathcal{C}^i \sim \mathcal{D}_{train}} L^{\mathcal{O}^i}(\phi - \eta_t \nabla_\phi L^{\mathcal{T}^i}(\phi)) \end{aligned} \quad (12)$$

where $L^{\mathcal{O}^i}$ refers to the multi-task loss attained on the sampled meta optimization set $\mathcal{O}^i$. To obtain $L^{\mathcal{O}^i}(\phi')$, we first compute $L_{res}^{\mathcal{O}^i}(\phi')$ and $L_{rec}^{\mathcal{O}^i}(\phi')$ by applying Equation 7 - 9 on $\mathcal{O}^i$. Then, we compute the weighted sum between both losses:

$$L^{\mathcal{O}^i}(\phi') = \alpha L_{res}^{\mathcal{O}^i}(\phi') + (1-\alpha) L_{rec}^{\mathcal{O}^i}(\phi') \quad (13)$$

where α determines the contribution of the persona reconstruction and response generation loss respectively. When $\alpha = 1$, MTML is analogous to MAML/PAML. Next, we sum the $L^{\mathcal{O}^i}(\phi')$ attained for every sampled persona $\mathcal{P}_{1:N}^{1:M}$. The original parameters ϕ are then updated using the average of losses obtained by dividing the summed loss by the batch size M. This can be formalized as:

$$\phi = \phi - \eta_o \nabla_\phi \frac{1}{M} \sum_{\mathcal{O}^i} L^{\mathcal{O}^i}(\phi') \quad (14)$$

where η_o refers to the outer loop learning rate and $L^{\mathcal{O}^i}(\phi')$ represents the multi-task training loss attained by the updated parameters ϕ' on the sampled validation set $\mathcal{O}^i$. This computation involves obtaining the gradient of a gradient i.e., second order differentiation. A summary of MTML framework is provided in Algorithm 1. Additionally, an overview of the MTML framework is provided in Figure 1(a).

2.3 Alternating Multi-Task Meta-Learning Framework

Instead of combining the response generation loss and persona reconstruction loss, Alternating-MTML(AMTML) involves constantly alternating between the two loss functions. Essentially, at every iteration, meta-training and meta-optimization are conducted with different loss functions. For AMTML, the α parameter would not be used. At

Algorithm 1 MTML

Require: Hyperparameters α, η_t, η_o
Require: Dataset $\mathcal{D}_{train}$
  for iteration i = 1 to n **do**
    Sample $\mathcal{C}^i = (\mathcal{T}^i, \mathcal{O}^i) \sim \mathcal{D}_{train}$
    for each persona in $\mathcal{T}^i$ **do**
      Calc $L^{\mathcal{T}^i}(\phi) = \alpha L^{\mathcal{T}^i}_{res}(\phi) + (1-\alpha) L^{\mathcal{T}^i}_{rec}(\phi)$
      Update $\phi' = \phi - \eta_t \nabla_\phi L^{\mathcal{T}^i}(\phi)$
      Calc $L^{\mathcal{O}^i}(\phi') = \alpha L^{\mathcal{O}^i}_{res}(\phi') + (1-\alpha) L^{\mathcal{O}^i}_{rec}(\phi')$
    end for
    Update $\phi = \phi - \eta_o \nabla_\phi \frac{1}{M} \sum_{\mathcal{O}^i} L^{\mathcal{O}^i}(\phi')$
  end for

every iteration, during meta-training, either the response generation loss L_{res} or the persona reconstruction loss L_{rec} will be used to update the parameters in the inner loop. Then, the alternate loss would be used to compute the loss for meta-optimization in the outer loop. For example, if the response generation loss is used to update the parameters during meta-training i.e. $L^{\mathcal{T}^i}_{res}$, the persona reconstruction loss would be used to derive the meta-optimization loss i.e. $L^{\mathcal{O}^i}_{rec}$.

In our implementation, this is achieved by utilizing the response generation loss during meta-training when the iteration count is even, and utilizing the persona reconstruction loss when during meta-optimization when the iteration count is odd. Due to the alternating loss functions, the computational complexity and memory requirements for AMTML is lower than MTML, which requires computing both loss functions during both meta-training and meta-optimization. A summary of the AMTML framework is provided in Algorithm 2. Additionally, an overview of the AMTML framework is provided in Figure 1(b). The training set $\mathcal{D}_{train}$ is used to train the model and the validation set $\mathcal{D}_{valid}$ is used to facilitate early stopping.

3 Experiment and Results

3.1 Corpus

The proposed MTML and AMTML frameworks were evaluated on the PersonaChat dialogue corpus (Zhang et al., 2018). The corpus comprises 1155 distinct personas, each consisting of several persona statements. The PersonaChat dialogue corpus was chosen due to the available persona statements which are used to compute the persona reconstruction loss. In our experiment, the corpus is divided into a training, validation and test set. The validation and test sets each consist of 100 unique personas. For our experiments, we utilize the training and validation sets during the meta-learning stage and the test set during the testing stage.

Algorithm 2 Alternating MTML

Require: Hyperparameters η_t, η_o
Require: Dataset $\mathcal{D}_{train}$
  for iteration i = 1 to n **do**
    Sample $\mathcal{C}^i = (\mathcal{T}^i, \mathcal{O}^i) \sim \mathcal{D}_{train}$
    for each persona in $\mathcal{T}^i$ **do**
      if i is even **then**
        Calc $L^{\mathcal{T}^i}_{res}(\phi)$
        Update $\phi' = \phi - \eta_t \nabla_\phi L^{\mathcal{T}^i}_{res}(\phi)$
        Calc $L^{\mathcal{O}^i}(\phi') = L^{\mathcal{O}^i}_{rec}(\phi')$
      else if i is odd **then**
        Calc $L^{\mathcal{O}^i}_{rec}(\phi')$
        Update $\phi' = \phi - \eta_t \nabla_\phi L^{\mathcal{T}^i}_{rec}(\phi)$
        Calc $L^{\mathcal{O}^i}(\phi') = L^{\mathcal{O}^i}_{res}(\phi')$
      end if
    end for
    Update $\phi = \phi - \eta_o \nabla_\phi \frac{1}{M} \sum_{\mathcal{O}^i} L^{\mathcal{O}^i}(\phi')$
  end for

3.2 Implementation

Following Madotto et al., we adopt the standard Transformer architecture (Vaswani et al., 2017) consisting of 6 encoder layers, 6 decoder layers and 4 attention heads is used along with the GloVe embedding (Pennington et al., 2014). The dimensions of the word embedding and hidden dimension of the Transformer are fixed at 300. We use SGD (η_t = 0.005, M = 16) during meta-training and Adam (η_o = 0.003, M = 16) during meta-optimization. For MTML, we define an additional hyperparameter, α, which accounts for the distribution between the persona recreation loss and the response generation loss.

3.3 Evaluation

Automatic Metrics Similar to Madotto et al., we compute the BLEU score (Papineni et al., 2002), the perplexity (PPL) and the C score (Madotto et al., 2019) to evaluate the quality of the generated response. The BLEU score measures the similarity between the generated response and the reference response. PPL is the negative log of the generated response. The C-score reflects the amount of persona information present in the generated response by measuring the persona consistency with respect to the corresponding persona statements via

a BERT-based Natural Language Inference (NLI) model, which is finetuned to indicate if the generated response entails or contradicts the corresponding persona statements. A high C score would imply greater persona consistency.

Human Evaluation We engaged 3 graduated individuals to evaluate 50 responses for each model using 3 criteria: persona consistency, fluency and contextual coherence. Consistency reflects the amount of persona information corresponding to the persona statements are reflected in the generated response. Fluency accounts for any grammatical, spelling and phrasing issues, while Coherence reflects the appropriateness of the dialogue with respect to the dialogue history. Responses were assigned a rating from -1 to 1 for each criteria. For consistency, -1 indicates contradiction, 0 indicates neutrality and 1 indicates persona consistency (i.e. the response accurately reflects information from the corresponding persona statements). For coherence, the individuals were told to assign a score of -1 for incoherent, contextually illogical responses, 0 for a moderate coherent responses and 1 for coherent, contextual responses. Finally, for fluency, -1 is assigned to responses with several fluency issues, 0 for responses with one or two fluency errors and 1 for perfectly fluent responses.

3.4 Experimental Settings

We benchmark MTML/AMTML with the PersonaChat corpus. We train the following Transformer models:

Std: We pretrain a standard Transformer model with only the dialogue context as input.

Std_p: We pretrain a standard Transformer model with both the dialogue context and persona statements as inputs.

$PAML$: We pretrain a standard Transformer via PAML (Madotto et al., 2019).

$MTML_\alpha$: We pretrain a standard Transformer trained via MTML (Section 3.2) where α = [0.9, 0.8, 0.7, 0.6, 0.5].

$AMTML$: We pretrain a standard Transformer via AMTML (Section 3.3).

During testing, the pretrained models described above were further trained, or finetuned, on dialogues corresponding to personas from the test set $\mathcal{D}_{test}$. As highlighted earlier, the models were finetuned using only the dialogue context, which was constructed by concatenating all previous utterances in the dialogue. For our experiment, the length of the dialogue context in each dialogue example would vary according to the number of turns. No restriction was placed on the number of dialogue turns required. No persona statements were used during finetuning. In the 5-shot and 10-shot setting, 5 and 10 dialogues corresponding to a persona was used to finetune the model parameters respectively. Then, the model tasked with generating the response corresponding to the same persona. Samples of the dialogue responses generated by each model is provided in Table 1. Table 2 and 3 depicts the results of automatic evaluation in a 10-shot and 5-shot setting respectively, while Table 4 and 5 depicts the results of the human evaluation in a 10-shot and 5-shot setting respectively.

3.5 Results & Discussion

$MTML_\alpha$ generally achieves higher C-scores and Consistency scores compared to PAML, indicating responses which incorporate a larger amount of persona information. This confirms our hypothesis that the introduction of the persona reconstruction task during meta-learning would result in a model which induces the persona from the dialogue context to a larger extent. However, it can be seen that PPL increases as α decreases. Since PPL scores have been found to correlate negatively with human likeness (Adiwardana et al., 2020), a high PPL score is undesirable. This finding is supported by the human evaluation results, where the Fluency and Coherence scores drop as α increases in both the 5-shot and 10-shot settings. This implies that there is a trade-off between general fluency and persona consistency in the generated response. During meta-learning, if α is too large, the combined loss can be effectively reduced by minimizing the persona reconstruction loss. Hence, the model would be trained to generate responses which contain as much persona information as possible without considering fluency or context. While this would result in persona consistent responses, the responses would be largely incoherent and unnatural.Since $\alpha = 0.8$ strikes a balance between the PPL and C-scores, we conclude that the optimal value of $\alpha = 0.8$.

Both $MTML_{0.8}$ and $AMTML$ improved the persona consistency of the generated responses. In terms of C-score, $MTML_{0.8}$ demonstrated a 78.9%(10-shot) and 100%(5-shot) improvement over $PAML$. In terms of persona consistency, $MTML_{0.8}$ demonstrated a 64.0%(10-shot) and

	PPL	BLEU	C-score
Std	35.87	0.93	0.00
Std_p	38.72	1.66	0.10
$PAML$	41.80	0.71	0.19
$MTML_{0.5}$	77.32	0.53	0.46
$MTML_{0.6}$	57.10	0.53	0.41
$MTML_{0.7}$	52.44	0.57	0.47
$MTML_{0.8}$	43.28	0.42	0.34
$MTML_{0.9}$	40.39	0.71	0.21
$AMTML$	48.66	0.48	0.29

Table 2: Automatic evaluation results (10-shot).

Method	PPL	BLEU	C-score
Std	36.75	1.02	-0.02
Std_p	38.78	1.79	0.09
$PAML$	40.46	0.65	0.15
$MTML_{0.5}$	76.38	0.41	0.50
$MTML_{0.6}$	55.19	0.53	0.48
$MTML_{0.7}$	50.69	0.44	0.45
$MTML_{0.8}$	41.42	0.38	0.30
$MTML_{0.9}$	39.94	0.62	0.13
$AMTML$	44.90	0.42	0.26

Table 3: Automatic evaluation results (5-shot).

	Consistency	Fluency	Coherence
Std	0.16	0.92	0.24
Std_p	0.19	0.89	0.29
$PAML$	0.25	0.77	0.28
$MTML_{0.5}$	0.47	0.24	0.20
$MTML_{0.6}$	0.45	0.57	0.27
$MTML_{0.7}$	0.45	0.60	0.30
$MTML_{0.8}$	0.41	0.80	0.32
$MTML_{0.9}$	0.39	0.88	0.17
$AMTML$	0.42	0.89	0.33

Table 4: Human evaluation results (10-shot).

	Consistency	Fluency	Coherence
Std	0.10	0.87	0.20
Std_p	0.09	0.89	0.21
$PAML$	0.13	0.84	0.27
$MTML_{0.5}$	0.22	0.03	-0.12
$MTML_{0.6}$	0.24	0.65	0.15
$MTML_{0.7}$	0.29	0.65	0.13
$MTML_{0.8}$	0.22	0.78	0.29
$MTML_{0.9}$	0.15	0.77	0.19
$AMTML$	0.23	0.85	0.20

Table 5: Human evaluation results (5-shot).

Automatic	PPL	BLEU	C-score
P^2Bot	18.1	0.61	0.33
Human	Consistency	Fluency	Coherence
P^2Bot	0.39	0.91	0.43

Table 6: Automatic and human evaluation results attained by P^2Bot

15.4%(5-shot) improvement over $PAML$. Similarly, compared to $PAML$, $AMTML$ achieved a 52.6%(10-shot) and 73.3%(5-shot) improvement when it comes to C-score. In terms of Consistency, $AMTML$ demonstrated a 68.0%(10-shot) and 76.9%(5-shot) improvement over $PAML$. Compared to $MTML_{0.8}$, in both the 10-shot and 5-shot settings, $AMTML$ achieved similar Consistency scores and slightly lower C-scores. However, while $MTML_{0.8}$ is comparable to $PAML$ with regard to Fluency and PPL, $AMTML$ outperformed $PAML$, $MTML_{0.8}$ and all other MTML variants in terms of Fluency. When it comes to coherence, $PAML$, $MTML_{0.8}$ and $AMTML$ generally achieved similar results.

Based on the results attained, while responses generated via MTML has a slight edge in terms of persona consistency, responses generated via AMTML are more fluent. On a side note, it should also be highlighted that the BLEU score did not correlate with any aspect of human evaluation. This further emphasizes the unsuitability of the BLEU score as an evaluation metric for dialogue generation (Liu et al., 2016).

3.5.1 PersonaChat SOTA Comparison

Additionally, we compare our proposed frameworks with the current state-of-the-art framework for PersonaChat: P^2Bot (Song et al., 2020a). P^2Bot involves finetuning the GPT pretrained language model on the training set via a transmitter receiver framework. This framework models the user's perception of the other party's persona in addition to the user's own persona. Hence, unlike MTML and AMTML, in the case of P^2Bot, persona statements are provided to the model along with the dialogue context during inference and testing.

From Table 6, it can be observed that the C-score attained by $MTML_{0.8}$, in the 10-shot setting, was comparable to P^2Bot. When it comes to the Consistency score, in the 10-shot setting, both $MTML_{0.8}$ and $AMTML$ outperformed P^2Bot. This implies that the responses generated

by $MTML_{0.8}$ and $AMTML$ generally reflect the corresponding persona information to a greater extent compared to P^2Bot *despite not being provided the persona statements during testing*. However, in terms of Fluency and Coherence, P^2Bot still outperforms both $MTML_{0.8}$ and $AMTML$. This could be partially attributed to the use of the GPT pretrained model, which enhanced the overall quality of the generated responses.

3.5.2 Persona Reconstruction

In this section, we will provide a brief discussion regarding the persona reconstruction task. Based on the results attained, it is clear that the introduction of the persona reconstruction task during meta-learning further incentivizes the model to incorporate more persona information in the generated responses. Under the proposed frameworks, it is challenging to evaluate the performance of the model solely on the persona reconstruction task. However, based on the observed responses and loss values, persona reconstruction tend to be more successful when a longer dialogue context $x_{1:t-1}$ (greater number of turns) is provided. This is expected as a short dialogue context would not contain sufficient persona information for the model to reconstruct.

Persona reconstruction is a interesting and challenging task that could be explored in future work. For this task, the model has to successfully infer the persona from the dialogue context as well as ensure the fluency of the generated description. Finetuning pretrained language models would be a good starting point for future work. Also, to prevent a mix-up between the personas from each interlocutor, only the dialogue utterances of the corresponding interlocutor should be utilized during training and inference.

4 Related Work

Multi-task Learning Multi-task learning broadly refers to the process of learning more than one tasks/objectives concurrently with a shared model. In addition to personalized dialogue generation, multi-task learning has been applied to task-oriented dialogue subtasks including response generation(Zhu et al., 2019), dialogue state tracking(GM and Sengupta, 2019; Trinh et al., 2018; Rastogi et al., 2018), dialogue act selection (McLeod et al., 2019), as well as conditional open-domain dialogue generation(Zeng and Nie, 2021; Ide and Kawahara, 2021).

Meta-learning Meta-learning involves teaching models how to learn efficiently and quickly. There are 3 broad categories of meta-learning algorithms: optimization-based (Finn et al., 2017; Nichol et al., 2018), metric-based (Snell et al., 2017; Vinyals et al., 2016), and model-based (Mishra et al., 2018; Santoro et al., 2016). Optimization-based meta-learning approaches, which involve directly updating the model's parameters to allow for rapid adaptation to unseen tasks, have been applied to various dialogue tasks. Examples of such applications include task-oriented dialogue generation (Mi et al., 2019; Dai et al., 2020; Peng et al., 2020), domain adaptation (Qian and Yu, 2019) and dialogue state tracking (Peng et al., 2020; Huang et al., 2020).

Personalized Dialogue Generation There are numerous forms of personalized dialogue generation. The form covered in this paper requires leveraging both the persona information and dialogue context. Another form of personalized dialogue generation involves conditioning the response on external profile/identity information. Thus far, many different architectures (Wu et al., 2020; Song et al., 2019; Wolf et al., 2019; Kottur et al., 2017; Joshi et al., 2017) and training frameworks (Song et al., 2020a; Liu et al., 2020; Zheng et al., 2019; Su et al., 2019b) which involve utilizing the encoded persona/ or personality information and dialogue context as input have been proposed. For certain dialogue corpora such as DailyDialog (Li et al., 2017) and PERSONALDIALOG (Zheng et al., 2020), the persona descriptions are not provided. Instead, a representation of the interlocutor's personality should be inferred from the dialogue history.

5 Conclusion

In this work, we proposed MTML and AMTML, 2 meta-learning frameworks which adopt our multi-task learning approach involving the addition of a persona reconstruction task. Empirical results demonstrate that both MTML and AMTML effectively increases the amount of persona information reflected in the generated dialogue responses compared to prior work. However, there is still room for improvement when it comes to the fluency and contextual coherence of the generated responses. Future work could involve improving these aspects of the responses by incorporating pretrained language models in meta-learning framework.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 609–618, Online. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

Sushravya GM and Shubhashis Sengupta. 2019. Unsupervised multi-task learning dialogue management. pages 196–202.

Yi Huang, Junlan Feng, Min Hu, Xiaoting Wu, Xiaoyu Du, and Shuo Ma. 2020. Meta-reinforced multi-domain state generator for dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7109–7118, Online. Association for Computational Linguistics.

Tatsuya Ide and Daisuke Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation. *CoRR*, abs/2105.11696.

Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *CoRR*, abs/1706.07503.

Satwik Kottur, Xiaoyu Wang, and Vitor Carvalho. 2017. Exploring personalized neural conversational models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3728–3734.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Sarah McLeod, Ivana Kruijff-Korbayova, and Bernd Kiefer. 2019. Multi-task learning of system dialogue act selection for supervised pretraining of goal-oriented dialogue policies. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 411–417, Stockholm, Sweden. Association for Computational Linguistics.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3151–3157. International Joint Conferences on Artificial Intelligence Organization.

N. Mishra, Mostafa Rohaninejad, Xi Chen, and P. Abbeel. 2018. A simple neural attentive meta-learner. In *ICLR*.

Young Na, Junekyu Park, and Kyung-Ah Sohn. 2021. Is your chatbot perplexing?: Confident personalized conversational agent for consistent chit-chat dialogue. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,*, pages 1226–1232. INSTICC, SciTePress.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *ArXiv*, abs/1803.02999.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word

representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task learning for joint language understanding and dialogue state tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 376–384, Melbourne, Australia. Association for Computational Linguistics.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1842–1850, New York, New York, USA. PMLR.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020a. Generating persona consistent dialogues by exploiting natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885.

Yiping Song, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang. 2020b. Learning to customize model structures for few-shot dialogue generation tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5832–5841, Online. Association for Computational Linguistics.

Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung yi Lee. 2019a. Personalized dialogue response generation learned from monologues. In *INTERSPEECH*.

Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019b. Personalized Dialogue Response Generation Learned from Monologues. In *Proc. Interspeech 2019*, pages 4160–4164.

Anh Duong Trinh, Robert J Ross, and John D Kelleher. 2018. A multi-task approach to incremental dialogue state tracking. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Aix-en-Provence, France. SEMDIAL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona.

M. Yang, W. Huang, W. Tu, Q. Qu, Y. Shen, and K. Lei. 2021. Multitask learning and reinforcement learning for personalized dialog generation: An empirical study. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):49–62.

Yan Zeng and Jian-Yun Nie. 2021. A simple and efficient multi-task learning approach for conditioned dialogue generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4927–4939, Online. Association for Computational Linguistics.

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.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9693–9700.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2019. Multi-task learning for natural language generation in task-oriented dialogue. 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 1261–1266, Hong Kong, China. Association for Computational Linguistics.

Detecting Interlocutor Confusion in Situated Human-Avatar Dialogue: A Pilot Study

Na Li, John D. Kelleher, Robert Ross
School of Computing
Technological University Dublin
{na.li, john.d.kelleher, robert.ross }@tudublin.ie

Abstract

In order to enhance levels of engagement with conversational systems, our long term research goal seeks to monitor the *confusion state* of a user and adapt dialogue policies in response to such user confusion states. To this end, in this paper, we present our initial research centred on a user-avatar dialogue scenario that we have developed to study the manifestation of confusion and in the long term its mitigation. We present a new definition of confusion that is particularly tailored to the requirements of intelligent conversational system development for task-oriented dialogue. We also present the details of our Wizard-of-Oz based data collection scenario wherein users interacted with a conversational avatar and were presented with stimuli that were in some cases designed to invoke a confused state in the user. Post study analysis of this data is also presented. Here, three pre-trained deep learning models were deployed to estimate base emotion, head pose and eye gaze. Despite a small pilot study group, our analysis demonstrates a significant relationship between these indicators and confusion states. We understand this as a useful step forward in the automated analysis of the pragmatics of dialogue.

Keywords

Confusion detection, situated dialogues, emotion recognition, head pose, eye gaze, pragmatics, avatar, wizard-of-oz

1 Introduction

Situated conversation either in the case of Human-Robot Interaction (HRI) or in the virtual world with Avatars provides significant challenges and opportunities for the further development and deployment of dialogue systems. In the case of robotic systems, applications ranging from healthcare assistants (Esterwood and Robert, 2020) to tour guides in a museum (Duchetto et al., 2019) can all take advantage of spoken interaction in a situated setting. Meanwhile within an online setting, applications such as online learning system (Doherty and Doherty, 2018) can also take advantage of the speech channel. However, in each of these scenarios, the need for fluid interaction where users remain engaged is hugely important, and that in order to provide an effective interface, the dialogue system must respond appropriately to the user's words and mental states.

Confusion is a unique mental state that can either precede a high degree of positive engagement in a task, or can also be correlated with negative states such as boredom and subsequent disengagement from a conversation (D'Mello et al., 2014). Estimating the confusion state of a user can hence be a very important step in improving the pragmatics modelling properties of an interactive system. By checking for confusion, or indeed precursors of confusion, we can in principle adjust the dialogue policy or information being presented to the user in order to assist them in the specific task being undertaken. Such monitoring can be seen as a specific form of engagement detection (Sidner et al., 2004; Dewan et al., 2018). In mainstream Human-Computer Interaction (HCI) studies, there have to this point been a number of studies that have investigated the modelling and detection of confusion (Kumar et al., 2019; Grafsgaard et al., 2011; Zhou et al., 2019). However, the majority of study in this area has concerned online learning in specific environments such as AutoTutor, ITS (Intelligent Tutoring Systems) and MOOCs (Massive Open Online Courses) or serious games; little work has focused on general engagement or task-oriented dialogue.

In light of the above, our goal in this paper is to explore the potential manifestations of confusion and investigate whether it is possible to detect the

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

confusion state as a pragmatics analysis task to supplement a multimodal situated dialogue system. While our primary area of interest is in HRI, this study has been executed with a focus on Human-Avatar Interaction (HAI). This is in part due to the relative ease of executing avatar based studies without the physical robot. More specifically, two research questions in this study are presented:

1. Are participants aware they are confused if we give them a specific confusing situation?

2. Do participants express different physical or verbal/non-verbal behaviours when they are confused that we can detect?

To answer these research questions, a wizard-of-oz human-avatar interaction study was designed based around an avatar web application which allowed participants to be both recruited remotely and engage in the interaction remotely. Study stimuli included a series of situated conversations to attempt to trigger confused states in the participants. Participants' behaviours including verbal or non-verbal languages, facial expression and body pose were recorded and subsequently analysed. Before detailing this work, we begin with a review of related work with a particular focus on setting out a relevant framework for engagement and specifically confusion estimation.

2 Related Work

The detection and monitoring of a participant's mental state in conversational interaction is a well-established area of research. In this section, we briefly review a number of works related to our own area of focus, and look in particular at the challenge of defining and identifying confused states during interaction.

2.1 Emotion & Engagement Recognition

The recognition of human emotional states has been noted as a pillar in engaging conversation in domains such as human-robot interaction (Spezialetti et al., 2020). In early work, Cohn (2007) indicated that human emotion may not be directly observable because emotion is a cognitive state, but that emotion can be explained through interaction context and evidenced by user survey, behavioural and physiological indicators. In terms of physiological indicators of emotional state, the facial expression is the most natural manifestation for a human. In terms of analysing facial expressions, the Facial Action Coding System (FACS) (Cohn, 2007; Menne and Lugrin, 2017) is an example of a *part-based method*, which defined the smallest units of muscular activity that can be observed in the human face, called Action Units (AUs). FACS is a well-known analysis tool that has been combined with self-report measurements. More recently of course, Deep Learning based image analysis methods have been used to demonstrate high accuracy for emotion recognition on facial images (Refat and Azlan, 2019). Similarly various recurrent and ensemble network architectures have been built to analyse multimodal datasets including speech (audio) data, text-based data and video data to provide estimates of emotional state (Tripathi and Beigi, 2018; Hazarika et al., 2018).

Beyond facial or verbal expressions, certain behaviours such as head pose, and eye gaze are also noted as non-verbal indicators of engagement and mental state during interaction. In particular, Emery (2000) explained that eye gaze is a component of facial expressions that can be interpreted as a cue to show people's attention to another individual, events or objects – either within spoken interaction or other domains (Khan and Lee, 2019; Mavridis, 2015; Zhang et al., 2020). Similarly, head-pose estimation is also studied extensively in face-related vision research where it is considered related to vision and eye-gaze estimation. Murphy-Chutorian and Trivedi (2009) provided an example of Wollaston illusion, where although eyes are in the same directions, the eye-gazing direction is decided by two differently oriented heads. They indicated that people with different head poses can reflect more emotional information such as dissent, confusion, consideration and agreement. Meanwhile, methods for training models of eye-gaze and head-pose estimation are generally consistent with facial expression analysis.

2.2 Engagement Detection

For us, confusion detection is intended to enhance engagement, and engagement in interaction is widely studied within the fields of psychology and linguistics, where engagement, for example, can be recognized as being broken down into three aspects: social connection (Doherty and Doherty, 2018; Sidner et al., 2004), mental state centric (Sidner et al., 2004), and a motivated or captivated phenomena (Jaimes et al., 2011). For our purposes

here, a key challenge is the detection of engagement. Within HCI and HRI there are three basic categories of engagement detection which are manual, semi-automatic and automatic (Dewan et al., 2018). Manual detection usually refers to tasks such as participant self-reporting or observational check-lists. Semi-automated engagement monitoring utilizes the timing and accuracy of responses such as reaction time for an interaction, or judgements of users' responses to tasks within an interaction. The automatic category meanwhile typically refers to machine learning driven methods operating directly or raw data or automatically extracted features, *e.g.*, Ben Youssef et al. (2019).

In recent years, there have been a wide variety of studies that have attempted to estimate and make use of user engagement in interaction (Tapus et al., 2012). For example, Ben Youssef et al. (2017) studied a human-robot interaction scenario with a fully automated robot (*i.e.*, Pepper Robot) for recognizing users' engagement in spontaneous conversations between users and the robot.

2.3 Confusion Detection

As a psychological state, confusion has been studied mostly to date within the context of pedagogy and related applied fields of learning, and depending on context has been defined as everything from a bonafide emotion through to an epistemological state. When confusion is considered as an effective response, confusion happens in people who are enthusiastic to know or understand something (D'Mello and Graesser, 2014). On the other hand, confusion may be defined as an epistemic emotion (Lodge et al., 2018) that is associated with blockages or impasses in the learning process. Confusion is also triggered by cognitive disequilibrium, where cognitive disequilibrium is itself defined as a state wherein a participant is learning but where obstacles to the normal flow of the learning process are encountered, the participant may feel confused when they encounter contradictory information leading to uncertainties, resulting in cognitive disequilibrium (Yang et al., 2015).

Arguel and Lane (2015) presented two thresholds (T_a and T_b) for levels of confusion in learning. The level of confusion between the two thresholds is termed productive confusion. It indicates that learners are engaged in overcoming their confused state. However, when the level of confusion is over T_b (persistent confusion), it is easy for learners to move to a state of frustration or even boredom. If the level of confusion is less than T_a, then learners may continue to engage in their learning. Lodge et al. (2018) designed a learning event in which the learner was in cognitive disequilibrium, where the disequilibrium was created by a manufactured impasse in the learning process. Similar to the notion of thresholds, in this study learners could be categorised to be in the zone of optimal confusion or sub-optimal confusion. Optimal confusion is productive confusion, which indicates that the learners are engaged in overcoming their confused state. On the other hand sub-optimal confusion is associated with persistent confusion where learners could not resolve the disequilibrium which in turn leads to possible frustration or boredom. D'Mello and Graesser (2014) meanwhile offers a transition oriented model where confusion can be seen as part of emotional transition between engagement/flow and frustration/boredom.

While there have been a number of studies that have touched on confusion in these learning scenarios, we find that there is no well-documented definition of confusion that can assist this research in modelling and mitigating confusion in interaction. In light of this, and for use in the context of dialogue centric human-machine interaction, we offer the following working definition of confusion. *Confusion is a mental state where under certain circumstances, a human experiences obstacles in the flow of interaction. A series of behaviour responses (which may be nonverbal, verbal, and, or non-linguistic vocal expression) may be triggered, and the human who is confused will typically want to solve the state of cognitive disequilibrium in a reasonable duration. However, if the confusion state is maintained over a longer duration, the interlocutor may become frustrated, or even drop out of the ongoing interaction.*

While in an ideal interaction there would be little or no confusion in practice, for the purpose of study it is useful to be able to induce confusion in an interlocutor. Within the literature, at least four types of confusion induction have been considered (Lehman et al., 2012; Silvia, 2010). The first is *complex information* where the material to be understood is genuinely complex and presents challenges to one individual (that might not necessarily apply to another individual). The second is the challenge of *contradictory information* where inconsistencies may push an individual into a confused state. The

third case is the provision of *insufficient information* where confusion is due simply to not enough background or relevant information being provided to an individual. Finally, and related to contradictory information, we have *feedback* inconsistencies where during an interaction one agent provides another with information that is inconsistent with the interaction to date.

3 Study Design

With our working definition of confusion as a guideline, we designed a Wizard of Oz (WoZ) (Riek, 2012) study to investigate: (a) the effectiveness of confusion induction methods in interactions; and (b) the relative performance of a range of manual, semi-automatic and automatic methods for confusion estimation. In the following we describe our overall experiment design; stimuli design, and approach to data analysis.

3.1 Study Overview

While our main focus is in the context of human-robot interaction, this experiment was designed as a human-avatar study to account for some of the study design limitations that were experienced due to the COVID-19 pandemic of 2020-2021. While an avatar does not provide a full situated experience, an avatar has notable benefits over a speech or text only based interaction (Heyselaar et al., 2017).

The experiment was based on a semi-spontaneous one-to-one conversation between an agent (in our case a wizard controlled avatar) and a participant (the user). Participants were recruited from across the university and study programmes, and these participants remained in their own homes while the wizard was similarly in their own work environment. Participants were requested to connect via a camera and audio enabled laptop and with reliable internet connectivity. Typical participation times were designed to be less than 15 minutes in total with 5 minutes for the task centric part of the conversation. At the beginning of the interaction participants were given consistent instructions and consent forms, and following the task (described later) all participants were asked to complete a survey (also detailed later). Finally at the end of this experiment, each participant was invited for a 3-minutes interview with the researcher.

For this study, a web application framework was developed and built on two components: one was a real-time chat application, while the other was

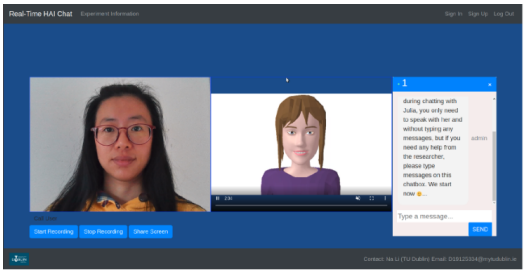


Figure 1: HAI Real-Time Chat Web App

Participant 1		
Stimulus	**Task**	**Condition**
1st	Task 1	A
2nd	Task 2	B
3rd	Task 3	A
Participant 2		
Stimulus	**Task**	**Condition**
1st	Task 1	B
2nd	Task 2	A
3rd	Task 3	B

Table 1: An example of the experiment sequence for two separate study participants.

an avatar application that was embedded within the real-time chat application. The avatar application was based on the framework by Sloan et al. (2020) which provides a sandbox with modules of an e-learning platform with an animated avatar. It integrates animation, speech recognition and synthesis, along with full control of the avatar's facial expressions to convey happiness, sadness, surprise, etc. The real-time chat application meanwhile is a web application for online interaction between the agent/avatar and participant that we developed to handle all survey steps, communication, and enrollment with the user. The application is designed to enable full data recording of both the avatar and the remote participant's audio, text, and camera streams. We illustrate the complete framework in Figure 1.

There were 23 participants in six countries who participated in this study; three of the participants were unable to complete the final experiment due to internet connectivity or equipment problems. All participants were over 18 years of age from different colleges around the world who can have a simple conversion in English at least. We successfully collected video data, user surveys and demographics information from 19 participants (8 males, 11 females) and acquired their permission to use their

data for this research purpose.

3.2 Dialogue Design

To stimulate confusion within a short conversation, we defined two conditions with appropriate stimuli. In condition A, stimuli were designed to invoke confusion in the participant. In condition B, stimuli were designed to allow a participant to complete a similar task in a straightforward way and should avoid confused states. Three separate task sets were defined with each task designed for both conditions. Task 1 was a simple logical question; task 2 was a word problem; while task 3 was a math question. We prepared at least two questions for each conditions in each task. As for the sequence of the experiment, Table 1 shows the sequence of conditions for each participant; for the first participant for example, the sequence of conditions is Task 1 with condition A, Task 2 with condition B and task 3 with condition A. The sequence of conditions between participants was alternated to balance the number of conditions for data analysis.

As for situated dialogues, there are two dialogues corresponding to two conditions, and one dialogue is for one condition; four patterns of confusion for two conditions: the first pattern of complex information and simple information, the second pattern of insufficient information and sufficient information and the last pattern of of correct-negative feedback and correct-positive feedback. For example, below is a word problem with the second pattern, for insufficient information in condition A: *"There are 66 people in the playground including 28 girls, boys and teachers. How many teachers were there in total?"*; while the case for sufficient information *i.e.*, condition B is: *"There are 5 groups of 4 students, how many students are there in the class?"*.

It should be noted that the design of individual stimuli includes both the verbal and non-verbal elements of the interaction. Thus, avatar's responses were mapped to visible behaviours (Cassell and Vilhjálmsson, 2004). Figure 2 shows an example of the mappings of the avatar's facial expressions and body gestures for conversational responses and conversational behaviours corresponding to positive reaction and negative reaction.

3.3 Data Preparation

Frame data was extracted for 19 participants' videos and each video was labelled with the sequence of conditions (*e.g.*, ABA or BAB), such that all frames were labelled as either condition A or condition B. To verify the frame labelling, labelling files with frame names and one condition were manually matched. The image data for condition A had 4084 frames, while the image data for condition B has 3273 frames. Facial recognition and alignment are a significant first step in pre-processing frame data, thus we applied an efficient method to centre crop a 224x224 region for each frame (Savchenko, 2021), and then used a (Multi-task Cascaded Convolutional Neural Networks) MTCNN-based face detection algorithm without frame margins. Figure 3 shows the original frame on the left, with the processed image on the right.

In addition to making use of the raw frame data, we also involved use of the post interaction survey questions. Here the user survey consisted of 10 questions using a 5-level Likert scale. Three of the questions were specific to the three tasks (logical questions, word problems and math questions) including the scores of the both conditions. Each user survey contains the results of the two conditions. Thus, the results of the survey were separated into two independent groups by the two conditions and then collected into one file for analysis with a flag

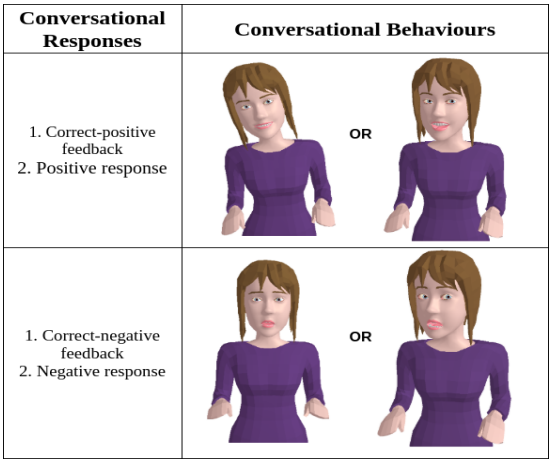


Figure 2: The mapping of the reaction status and visible traits for the avatar.

Figure 3: Frame and aligned facial image

noting "condition", as well as additional parameters such as the average of scores of two tasks under the same condition.

4 Data Analysis

To address our research questions introduced earlier, we applied a number of feature analysis algorithms to our data and analysed the interactions between these and our experimental conditions and the results of the survey questions. Below we detail these methods and present the results of this analysis.

4.1 Frame Data Measurement

Our primary form of analysis was based around the automated processing of video data to determine if automatically extracted indicators of emotion, head pose and eye gaze have a significant correlation with confusion state. For emotion detection we made use of a visual emotion detection algorithm (Savchenko, 2021) based on the MobileNet (Howard et al., 2017) architecture and trained on the AffectNet dataset (Mollahosseini et al., 2017) for 8 target classes, namely the 7 main facial expressions: Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, and an 8th: Contempt. Table 2 shows the number of each of the 7 primary emotion categories predicted grouped by condition A and condition B. It is notable that the predicted results in condition A for 4 negative emotion categories (anger, disgust, fear and sadness) are considerably more than for condition B. In contrast, as for 2 positive emotion categories (happiness and surprise), the number of predicted results in condition A are lower than condition B. Undoubtedly, for neutral emotion we see that the condition A is higher than condition B. Figure 4 presents a comparison of the results of emotion prediction for three categories (negative, positive and neutral) grouped by the conditions. In order to deep understand the correlation relationship between the three emotional categories and conditions, a statistical analysis of whether there is a statistically significant relationship between three emotional categories and the two experimental condition classes A and B. The result of an independent-sample t-test is that there is a significant difference in the three emotional categories (negative, positive and neutral) and two conditions ($M = 0.77, SD = 0.94$ for condition A, $M = 0.48, SD = 0.60$ for condition B), $t(715) = 5.05, \rho - value < 0.05$.

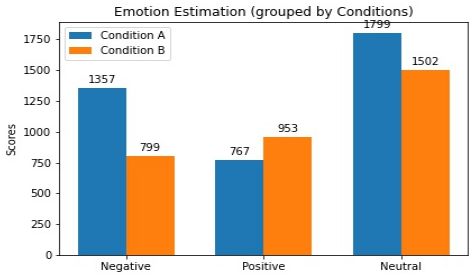


Figure 4: Comparison of three emotional categories grouped by condition A and condition B

For head-pose estimation, we applied use of the model due to Patacchiola and Cangelosi (2017) that makes use of CNNs, dropout and adaptive gradient methods trained on the three public datasets, namely the Prima head-pose dataset (Gourier et al., 2004), the Annotated Facial Landmarks in the Wild (AFLW) (Köstinger et al., 2011) and the Annotated face in the Wild (AFW) dataset (Zhu and Ramanan, 2012). The predicted results are the angles of pitch, yaw and roll for each image. We calculated the sum of absolute values of the three angles as a new feature for statistical analysis because only sum of values of pitch, yaw and roll will be the canceling effect of the positive and negative values, even the sum of values may be 0 as a person has different angles of direction with different positive or negative values of pitch, yaw and roll. Using this metric our related research question is whether there is a statistically significant relationship between the sum of absolute values of these three angles and our two experimental condition classes A and B. The result of an independent-sample t-test is that there is a significant difference in the sum of absolute values of these three angles and two conditions ($M = 21.96, SD = 9.46$ for condition A, $M = 27.40, SD = 12.21$ for condition B), $t(703) = -6.61, \rho - value < 0.05$.

We also plotted the sum of the three angles of the two conditions (see Figure 5). From this we can see that the values of condition A form a less discrete distribution than condition B. While we cannot draw conclusions from it, we also analysis the specific yaw, roll and pitch angle for individuals. To illustrate Figure 6 shows the labelled time for condition A (read lines) and condition B (blue lines), thus this shows for one individual the fluctuations of the pitch angle, yaw angle and roll angle in the time series. This indicates that the angle of the participant's head posture angle in condi-

Condition	Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral	Overall
A	262	282	136	677	702	65	1799	3923
B	77	165	57	480	858	95	1502	3234

Table 2: Result of emotion estimation grouped by condition A and condition B

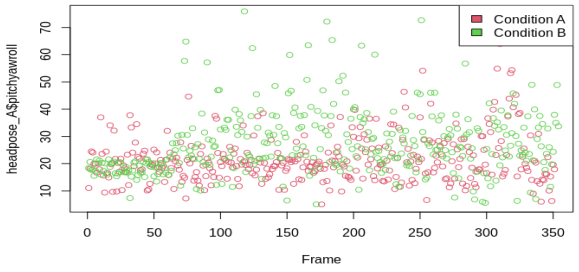


Figure 5: Head-pose estimation: plot the sum of angles values for condition A and condition B

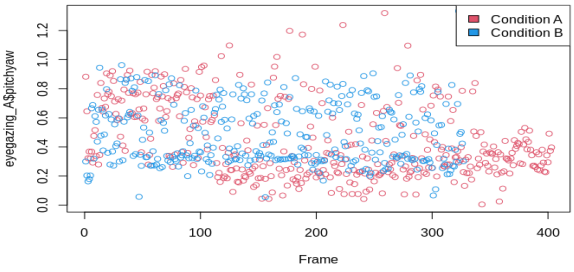


Figure 7: Eye-gaze estimation: plot the sum of angles values for condition A and condition B

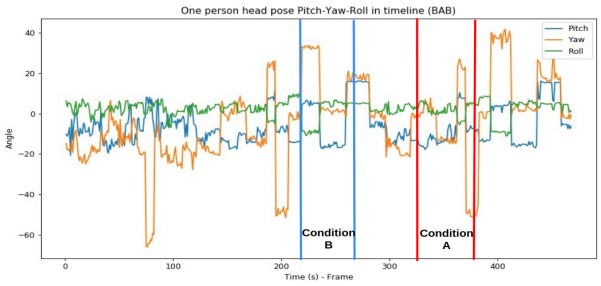


Figure 6: Head-pose estimation: plot the change of one person's pitch, yaw and roll angles at an experiment

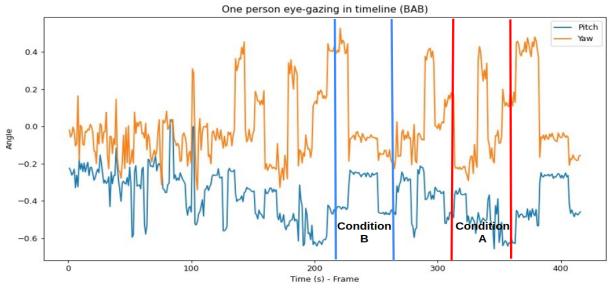


Figure 8: Eye-gaze estimation: plot the change of one person's pitch and yaw angles at an experiment

tion A was generally smaller than the angle of the participant's head posture angle in condition B.

For eye-gaze estimation we applied a state-of-art gaze estimation model that had been trained on the recently published ETH-XGaze dataset (Zhang et al., 2020). The GTH-XGaze dataset includes more than one million high-resolution images of different gazes in extreme head poses from 11 participants. The predicted results are angles of pitch and yaw for relative eyes directions. Again in this case, we summed up the absolute angles of two results, and we ask whether there is a relationship between this metric and our two experimental conditions A and B. An independent-samples t-test again was conducted to compare the two sets. There is a significant difference in the sum of absolute values of pitch and yaw and two conditions was found ($M = 0.44, SD = 0.26$ for condition A, $M = 0.49, SD = 0.22$ for condition B), $t(728) = -2.58, \rho - value < 0.05$.

In addition, we used the same method as with head-pose estimation to demonstrate these results. Figure 7 shows that the eye-gaze values of condition A form a more discrete distribution than those for condition B. Meanwhile in figure 8, we can see that the fluctuations of the same individual participant's pitch angle and yaw angle plotted in the time series. Here we see in this example case that the gaze angle of the participant in condition A is greater than that of the participant in condition B.

4.2 Subjective Measurement

For our survey results we analysed self-reported scores with respect to the two stimuli control conditions A and B. We can break down this analysis into two sub-questions. The first of these is whether there is a statistically significant relationship between the average of self-reported confusion scores for each of the three performed tasks and the conditions. The second three sub-questions are whether there is a statistically significant relationship between confusion scores for each of the three performed tasks and the two conditions.

With respect to the first question, an independent-

samples t-test was used and found that there is no significant difference between the average of confusion scores of the three tasks and two conditions ($M = 3.50, SD = 1.40$ for condition A, $M = 2.97, SD = 1.12$ for condition B), $t(36) = 1.28$, $\rho - value = 0.21$. However, with respect to the second three questions: firstly, there is no significant difference in the confusion scores for task 1 with two conditions was found ($M = 3.00, SD = 1.07$ for condition A, $M = 2.44, SD = 1.33$ for condition B), $t(15) = 0.94, \rho - value = 0.36$; secondly, there is no significant difference in the confusion scores for task 2 with two conditions was found ($M = 3.09, SD = 1.22$ for condition A, $M = 3.10, SD = 1.29$ for condition B), $t(19) = -0.02, \rho - value = 0.99$; lastly, the result indicated that there is a significant difference in the confusion scores for task 3 was found ($M = 4.38, SD = 0.74$ for condition A, $M = 3.00$, $SD = 1.12$ for condition B), $t(15) = 2.94$, $\rho - value < 0.05$.

5 Discussion

Based on the results provided in the previous section, we note that the following holds with respect to the specific questions that we identified:

1. Participants are not always aware they are confused if we gave them a specific confusing situation.

2. When they are confused, their emotion is more negative than when they are not confused.

3. When they are confused, the range of angles of eye gazing is more than when they are not confused.

4. When they are confused, the range of the angles of head shaking is less than when they are not confused.

Due to size and scope limitations, this is in essence a pilot study of confusion induction and detection. Notable limitations are not only on sample size but a number of technical challenges with the study. First, the qualities of videos of participants varied because of the quality of network connection, camera specification, and camera position, etc. Second, the sample size and range of participant backgrounds are a major limitation which limits the conclusions that can be drawn from this work. Third, as noted in Section 2.3, confusion is a unique mental state which can transit to positive states or negative states; in this pilot study, there are no clear dialogues' boundaries and time frames to distinguish the level of confusion. Finally, it should be mentioned that during the 3-minutes interview, many participants reported that they expected to have a wonderful conversation with the avatar, but this experiment lacked casual conversation and even a participant's expectations of the avatar's abilities were often not met.

Nevertheless, we believe that the study results do demonstrate that the approach to data collection and analysis are worthwhile, moreover we intend to build upon this in future work. At a minimum we intend to introduce audio and linguistic content analysis to expand beyond the visual and self-reporting data made use of in the current work. Second, and importantly, having established the general framework we wish to conduct further in-person studies to build upon our framework but with fewer constraints in place due to the COVID-19 pandemic. Ultimately our goal is also to study mitigation factions in confusion situations, and as such we will also be expanding our studies to study the effects of different clarification strategies on the confusion state.

6 Conclusion

In this paper we have proposed the study, detection, and mitigation of confusion as an important factor in improving dialogue centric human computer interaction. We also proposed a new working definition of confusion for our purposes and outlined a study that we conducted to determine if confusion could be induced and detected in a human-avatar task oriented interaction. While we did not find a significant relationship between self-reported confusion scores and induced confusion states, we did find significant differences between observed physical states, *i.e.*, facial emotion, head pose, eye gaze and our induced confused states. Although a small sample size is insufficient for generalisation, we see this work as a crucial initial step down the path to a computational model of confusion in multimodal dialogue.

Acknowledgments

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

Amaël Arguel and Rod Lane. 2015. Fostering deep understanding in geography by inducing and managing confusion: An online learning approach. *ASCILITE 2015 - Australasian Society for Computers in Learning and Tertiary Education, Conference Proceedings*, (November):374–378.

Atef Ben Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions. pages 464–472.

Atef Ben Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics*, 11.

J. Cassell and H. Vilhjálmsson. 2004. Fully embodied conversational avatars: Making communicative behaviors autonomous. *Autonomous Agents and Multi-Agent Systems*, 2:45–64.

Jeffrey F. Cohn. 2007. Foundations of human computing: Facial expression and emotion. In *Artifical Intelligence for Human Computing*, pages 1–16, Berlin, Heidelberg. Springer Berlin Heidelberg.

M. A. Dewan, M. Murshed, and F. Lin. 2018. Engagement detection in online learning: a review. *Smart Learning Environments*, 6:1–20.

Sidney D'Mello and Art Graesser. 2014. Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica*, 151:106–116.

Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170.

Kevin Doherty and Gavin Doherty. 2018. Engagement in hci: Conception, theory and measurement. *ACM Comput. Surv.*, 51(5).

Francesco Duchetto, Paul Baxter, and Marc Hanheide. 2019. Lindsey the tour guide robot - usage patterns in a museum long-term deployment. pages 1–8.

N.J. Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience Biobehavioral Reviews*, 24(6):581–604.

Connor Esterwood and Lionel Robert. 2020. Personality in healthcare human robot interaction (h-hri): A literature review and brief critique.

Nicolas Gourier, D. Hall, and J. Crowley. 2004. Estimating face orientation from robust detection of salient facial structures.

Joseph F Grafsgaard, Kristy Elizabeth Boyer, and James C Lester. 2011. Predicting Facial Indicators of Confusion with Hidden Markov Models. Technical report.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. page 2122–2132.

Evelien Heyselaar, Peter Hagoort, and Katrien Segaert. 2017. In dialogue with an avatar, language behavior is identical to dialogue with a human partner. 49(1):46–60.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

Alejandro Jaimes, Mounia Lalmas, and Yana Volkovich. 2011. First international workshop on social media engagement (some 2011). In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, page 309–310, New York, NY, USA. Association for Computing Machinery.

Muhammad Qasim Khan and Sukhan Lee. 2019. Gaze and eye tracking: Techniques and applications in adas. *Sensors*, 19(24).

Harsh Kumar, Mayank Sethia, Himanshu Thakur, Ishita Agrawal, and Swarnalatha P. 2019. Electroencephalogram with Machine Learning for Estimation of Mental Confusion Level. *International Journal of Engineering and Advanced Technology*, 9(2):761–765.

Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151.

Blair Lehman, Sidney D'Mello, and Art Graesser. 2012. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, 15(3):184–194. Emotions in online learning environments.

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.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

Isabelle M. Menne and Birgit Lugrin. 2017. In the face of emotion: A behavioral study on emotions towards a robot using the facial action coding system. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 205–206, New York, NY, USA. Association for Computing Machinery.

Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *CoRR*, abs/1708.03985.

Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626.

Massimiliano Patacchiola and Angelo Cangelosi. 2017. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143.

Chowdhury Mohammad Masum Refat and Norsinnira Zainul Azlan. 2019. Deep learning methods for facial expression recognition. In *2019 7th International Conference on Mechatronics Engineering (ICOM)*, pages 1–6.

L. Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. In *HRI 2012*.

Andrey V. Savchenko. 2021. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *CoRR*, abs/2103.17107.

Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: A study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, IUI '04, page 78–84, New York, NY, USA. Association for Computing Machinery.

P. Silvia. 2010. Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, 4:75–80.

John Sloan, Daniel Maguire, and Julie Carson-Berndsen. 2020. Emotional response language education for mobile devices. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA. Association for Computing Machinery.

Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:145.

Adriana Tapus, Andreea Peca, Aly Amir, Cristina Pop, Lavinia Jisa, Sebastian Pintea, Alina Rusu, and Daniel David. 2012. Children with autism social engagement in interaction with nao, an imitative robot – a series of single case experiments. *Interaction Studies*, 13.

Samarth Tripathi and Homayoon S. M. Beigi. 2018. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *CoRR*, abs/1804.05788.

Diyi Yang, Robert E Kraut, Carolyn P Rosé, and Ros´ Rosé. 2015. Exploring the Effect of Student Confusion in Massive Open Online Courses. Technical report.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision (ECCV)*.

Yun Zhou, Tao Xu, Shaoqi Li, and Ruifeng Shi. 2019. Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game. *Universal Access in the Information Society*, 18(3):551–563.

Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886.

The Language of Persuasion, Negotiation and Trust

José Lopes
Heriot-Watt University
Edinburgh, United Kingdom
jd.lopes@hw.ac.uk

Helen Hastie
Heriot-Watt University
Edinburgh, United Kingdom
h.hastie@hw.ac.uk

Abstract

There is a need to clearly understand the effect that interactive systems can have on users in the real world. This study explores whether aspects of social interaction (persuasion and negotiation) can be predicted purely from linguistic, politeness and collaborative features. Amongst other findings, we show that politeness cues (such as expressing gratitude) are important for successful negotiation dialogues and that collaborative features (such as repeated content between consecutively turns) are important for effective persuasion. We report here accuracy for automatic prediction methods based purely on interaction features using logistic regression, but also explore more opaque methods including neural models trained with dialogue embeddings. The two scenarios explored both involve economic decision-making, thus the subject has some stake in the outcome of the interaction, which is important for investigating trust.

1 Introduction

As interactive systems become more sophisticated, we can now look to various social aspects of interaction such as persuasion, negotiation and building of trusting relationships. However, there is a lack of understanding of how successful dialogues in this regard, manifest and what linguistic phenomena are observed. Designing and conducting studies to measure trust (subjectively or objectively) is particularly difficult because, in order to instill varying levels of trust in subjects, they have to be involved in the task and feel vulnerable to the outcome (Rousseau et al., 1998). One way to try to emulate this is to involve subjects in some kind of financial commitment to decisions made in the experimental set-up. The underlying assumption is that choices in such scenarios provide a reliable approximation of success in terms of persuasion, negotiation and consequently trust and trustworthiness (Camerer, 2011).

This paper reports an investigation into linguistic cues in two datasets that involve such economic decision-making: 1) where participants negotiate the price of items from real Craigslist advertisements (He et al., 2018); and 2) where one of the participants has to convince the other to donate part of their experimental reward to charity (Wang et al., 2019). The first of these datasets looks at a negotiation setting where one participant plays the role of a buyer and the other a seller and for the second dataset, one subject has the persuader role while the other is the persuadee. We posit that for both types of situations, in order for the interaction and transaction to succeed, there needs to be a trusting relationship between participants because these scenarios involve some emotional and financial investment. If we can establish trends and phenomena in language and dialogue that enable persuasion, effective negotiation and trust, then these can be used to inform dialogue management and natural language generation.

The importance of trust in human-robot interaction and conversational systems is a topic of much recent research (Kok and Soh, 2020). Levels of conflict of interest have been shown to be important for negotiation success (Cadilhac et al., 2013) and the role trust plays when coming to an agreement (Balliet and Van Lange, 2013). This form of cooperation depends on whether trust is conditional or unconditional. Conditional trust represents the minimum level of trust to facilitate social and economic exchanges toward a common goal (Jones and George, 1998). Rempel et al. (1985) state that trust evolves over time in interpersonal relationships, nurtured through interaction. However, trust can fall away rapidly, for example following an error (Nesset et al., 2021). In order to create interactive systems that are able to react and mitigate against

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

over-trust or undertrust/distrust (when perhaps the system is incorrect or misguided), we need to be able to monitor and infer the user's level of trust. Currently, measures of trust and trustworthiness are mostly collected from subjective questionnaires after the interaction (Schaefer, 2013; Jian et al., 2000; Ullman and Malle, 2018) or during the interaction (Khalid et al., 2019). Even if such methods have been validated, they can be considered intrusive as they break the flow of the interaction and are thus impractical for actually deployed systems. By observing linguistic phenomena, we hope to be able to develop an automatic method for predicting persuasion or whether a deal has been achieved without intrusive measures. In the future, we aim to extend the approach to predict trust in dialogues. This would allow for monitoring interaction, thus providing seamless mitigation through dialogue and language.

In this paper, we address the following research questions:

1. RQ1: Can we identify linguistic indicators present trustworthy interaction, in particular in successful persuasion/negotiation dialogues?

2. RQ2: Do role-specific linguistic indicators influence the outcome of the dialogue in a particular way?

3. RQ3: Can we use data-driven methods to predict the outcome of a persuasion/negotiation dialogue?

The contributions of the paper are thus two-fold: an in-depth analysis of linguistic indicators for successful dialogues, breaking this down by role, and providing discussion on how they may also influence trust in interaction. Secondly, we present data-driven methods, of varying transparency, for automatically predicting success in dialogue in terms of persuasion and whether a deal has been reached.

The paper starts by reviewing previous work on detecting trust and using linguistic indicators in predicting human behaviour (Section 2). The data used are described in detail in Section 3. The methodology is described in Section 4 and the results achieved in Section 5. We then discuss less transparent methods trained with dialogue embedding features and neural modelling in Section 6. In Section 7, we discuss the results and implications, and finally conclude the paper with Section 8.

2 Related Work

To achieve trustful interactions, systems have to become trustworthy. In order to do that, systems need to be equipped with resources to monitor the impact of their actions and how they affect the user's perception of the trustworthiness of the system. Therefore, there have been a number of studies where researchers have investigated specific cues that could be associated with trustworthiness. In Lucas et al. (2016), non-verbal cues have been studied in the context of negotiation dialogues between humans. Their goal was to predict both the perceived trustworthiness (i.e. partner perceptions of trustworthiness) and the reported perceived trustworthiness (i.e. if participants are honest). This study showed humans were actually poor predictors of trustworthiness, when compared with the proposed machine learning approach that used multimodal data. Still in the negotiation domain, Mell and Gratch (2017) found that negotiations were more likely to be successful when agents behaved aggressively. However, even if this strategy could lead to successful outcomes in the short-term, it is not necessarily advisable for human-robot interaction to display aggression in the long-term. Similar to our approach, Mell et al. (2019) used machine learning to predict the outcome of a negotiation using interaction features (e.g. number of turns), which were fed into both a linear model and a deep neural network. They do not, however, explore lexical features.

The above-mentioned approaches follow the intuition described in (DeSteno et al., 2012), that trust-related signals will likely emerge dynamically within the context of an interpersonal situation between individuals who are unfamiliar with each other. In addition, DeSteno et al. (2012) found that the accuracy of individuals in predicting trustworthy behaviour was higher when they had access to non-verbal cues. Examples of such cues were leaning forward or head nods. Lisetti et al. (2013), used a similar intuition to test whether different behaviours implemented in virtual agents were perceived more trustworthy. They found that the empathic version of the agent was generally preferred to its non-empathic counterpart on a number of dimensions related to trust, such as willingness to follow the agent's advice or politeness and willingness to continue the interview. Torre et al. (2018) also manipulated the agent's behaviour and measured the perceived trust. The virtual character

with a smiling face was perceived more trustworthy, knowledgeable and appealing. Kraus et al. (2020) modified the robot pro-activity and measured trust. The pro-active (contrasting with the reactive) version of the robot had a higher acceptance rate when it was possible to have natural dialogue, exemplifying the importance of dialogue and for acceptance and trust. Rapport building is also known to be a persuasion strategy that will likely increase trust. Therefore, Zhao et al. (2018) combined social dialogue with a model for task-oriented dialogue, including a first phase intended to build rapport. Examples of strategies used were self-disclosure, shared experience and praise.

2.1 Language and Trust

So far, we have focused on negotiation and persuasion as a means to maintain and manage trust, however, the above-mentioned studies mostly focus on non-verbal behaviour. Of specific interest here is whether we can observe linguistic indicators of these phenomena and use these to automatically predict varying levels of trust. Example studies looking into this area include Scissors et al. (2008), who investigated lexical mimicry (i.e. repetition of words or word phrases by both partners). They found that higher levels of mimicry were present in high-trusting pairs than low-trusting pairs. With regards to lexicon items, Rashkin et al. (2017) show that first-person and second person pronouns are used more in less reliable or deceptive news texts. On the other hand, Newman et al. (2003) found fewer self-references in people telling lies (so less trustworthy) about their personal opinions. These differences can perhaps be explained by the fact that the former is in relation to written facts, whilst the latter is about storytelling. With regards the use of superlatives and comparatives, Rashkin et al. (2017) found that trusted sources are more likely to use assertive words and less likely to use hedging words.

Continuing the theme of trustworthy news sources, Glenski et al. (2018) performed a study where they labelled bot and human users' reactions to (mis)information posted by various news sources and measured how bot and human users reacted to deceptive news sources compared to trusted news sources. However, the language aspect was not analysed. Volkova et al. (2017), on the other hand, found that incorporating linguistic and network features via a "late fusion" technique boosted performance of their suspicious tweet classifier. They found that verified news tweets contain significantly fewer bias markers, hedges and subjective terms.

Recent work has tried to use linguistic indicators to predict behaviours in interactive settings. Constructiveness has been one of the behaviours investigated in the context of an exploration game (Niculae and Danescu-Niculescu-Mizil, 2016) and disputes about Wikipedia articles (De Kock and Vlachos, 2021). In Zhang et al. (2018), politeness markers were used to predict if conversations were likely to fail at early stages. A conversation failure could be seen as a loss of conditional trust between interlocutors. In Niculae et al. (2015), sentiment, argumentation and discourse, politeness, subjectivity and talkativeness were used as linguistic cues to identify betrayal in a competitive game. As stated in Peskov et al. (2020), trust can be betrayed through deception, therefore some of these features might be relevant to our study. The most similar to our work is (Chawla et al., 2020), where BERT and linguistic features were used to predict the final price of successful negotiation in the Craigslist Bargain dataset. In our work, we use different lexical features and dialogue embeddings and have different tasks, namely the binary prediction of persuasion and whether a deal has been achieved. We believe this is an easier task for the model and thus would lead to further insights through the use of simpler more transparent modelling methods.

In this paper, we make use of some of the above-mentioned interaction cues, however, we apply them to negotiation and to the new domain of persuasion dialogues in scenarios of economic decision-making, where subjects in these types of scenarios have been shown to exhibit conditional trust.

3 Data

Two datasets were used in our analysis: the Craigslist Bargain dataset (He et al., 2018) and the Persuasion for Good dataset (Wang et al., 2019). In this section, we will provide a high-level description of these datasets. Further details can be found in the respective papers.

3.1 Craigslist Bargain

This dataset contains 6555 negotiation dialogues collected through crowd-sourcing. During data collection, crowd-workers were provided a real

Craigslist advertisement and were assigned roles of the buyer or the seller. They had to converse with another participant in order to negotiate an agreed price and thus close the transaction. Each participant was trying to push for a target price specified in the job (HIT) description. The datasets include information about these prices and the final closing price, and if participants eventually reach an agreement. This dataset has established partitioning for train/test/dev, which we have used in the research we present in this paper, in line with other work on the same dataset (He et al., 2018).

3.2 Persuasion for Good

The Persuasion for Good dataset is composed of 1017 dialogues between crowd-workers. Each participant had a specific role in the conversation. One crowd-worker, the persuader, had to convince the crowd-worker they were paired with, the persuadee, to donate a fraction of the amount they would receive for performing the task to a given charity (the same charity was used throughout the whole data collection). The persuader could also opt to donate part of their financial reward to the same charity at the end of the dialogue. The amount donated by each participant was recorded. The dataset also includes personality information gathered through pre-screening tests, in addition to demographics. A subset of dialogues was manually annotated for specific persuasion strategies and also for the intended donation verbalised by the persuadee during the dialogue (note that some persuadees actually did not donate the amount they verbally committed).

4 Method

In this section, we describe the method followed to perform two tasks: 1) predict the outcome of the dialogue and 2) identify the most relevant features in this prediction. Because we use a transparent method for prediction, we can do both of these tasks simultaneously. In both datasets, we have used the same features and extracted them from the conversations. We have drawn inspiration from the approach proposed in (De Kock and Vlachos, 2021) for feature-based models. We include the feature groups described below (a full reference of the features used can be found in Appendix A):

- Politeness strategies from (Zhang et al., 2018) for capturing tokens associated with greetings, apologies, directness and other politeness markers;
- Markers for collaboration from (Niculae and Danescu-Niculescu-Mizil, 2016) such as mutual pronoun usage or linguistic style accommodation (COLL).
- LIWC (Pennebaker, 2001) that provides counts of words associated with a given sentiment using pre-built lexicons.

All of the above-mentioned features were extracted at the turn-level, using Convokit (Chang et al., 2020). Similarly to De Kock and Vlachos (2021), at the end of the dialogue, for each feature we take the average (avg) and the gradient of a straight line fit of the feature value throughout the conversation (fit). The latter was done to assess how the feature value evolved throughout the dialogue. Then we have used the features, which will henceforth be called lexicon-based features, to train logistic regressions (LR). The LR method was chosen as it is reasonably transparent and allows the interpretation of the model by examining the weights of each feature.

5 Results

In this section, we present results for each of the datasets used. We used accuracy and F1-score as metrics, as all our tasks are binary classifications and the labels can be unbalanced (see the majority baselines in the results tables). We also report the McFadden R^2 score, the coefficient of determination, to provide a measure of how well the learned model fits the data. For the case of models trained with lexicon-based features, we report the 5 features with the highest absolute coefficients in the trained regressor.

5.1 Negotiation Dialogues

The task for this dataset is to automatically predict whether an agreement had been reached (binary deal/no-deal) and understand what features could help lead to this. We have used the dataset splits (5147 for train, 582 for validation, 826 for test) available in the data release. All sets of features used were able to improve over the majority baseline both in terms of accuracy and F1-score (it is a strong baseline given the dataset is highly unbalanced), as seen in Table 1. Regarding feature types, out of the 5 features in the best performing lexicon-feature based model, 4 were politeness features. In addition, dialogues with a trend of increasing turn length (fit_n_words) were more likely to lead

Features	Accuracy	F1-score	R^2	Top-5 features
Baseline Majority	0.769	0.869	-	-
COLL	0.810	0.886	-0.121	-avg_agree +fit_gap -fit_n_repeated_pos_bigram -avg_n_repeated_content -fit_n_repeated_stop
LIWC	0.815	0.886	0.016	-fit_n_words +avg_certain -avg_geo +avg_n_introduced_w_hedge +avg_n_introduced
Politeness	0.833	0.896	-1.123	-avg_has_negative -avg_apologising -avg_indicative -avg_direct_start -avg_indirect_greeting
COLL + LIWC + Politeness	**0.847**	**0.904**	**0.489**	-fit_n_words -avg_has_negative +avg_has_positive +avg_gratitude -avg_apologising
Buyer Features	0.832	0.896	**0.380**	-fit_pron_me +fit_pron_we +fit_1st_person +fit_indicative +avg_subjunctive
Seller Features	0.834	0.898	-0.222	+fit_n_introduced -avg_direct_start -fit_pron_you -fit_hedges +fit_indicative
Buyer+Seller Features	**0.857**	**0.910**	-0.519	-avg_seller_1st_person -avg_buyer_2nd_person_start +fit_seller_apologising +fit_buyer_please_start +fit_seller_n_adopted_w_hedge

Table 1: Accuracy, F1-score and McFadden's R^2 for predicting negotiation success in the Craigslist Bargain dataset. The speaker-independent features are in the top part of the table. Speaker-dependent features are in the bottom part of the table where the buyer and seller features include LIWC+Politeness separated out and calculated per role. The top-5 features are sorted according to the absolute coefficient value.

to a no-deal. This could indicate the use of longer, more elaborate utterances in an attempt to convince the other party. Dialogues where negative words were identified combined with a high number of apologetic words were also more likely to lead to no deal. On the other hand, dialogues where positive words were identified, combined with high rates of gratitude words (e.g. 'thank you') were more likely to result in a dialogue with a deal.

To further understand the impact of the behaviour of each participant in their various roles in the negotiation, speaker-dependent features were computed, specifically the LIWC and Politeness features for each speaker, be they a buyer or a seller. Since COLL features are meant to capture markers for collaboration, they are viewed as speaker-neutral. Thus in the lower part of Table 1, the results are split into the two buyer/seller roles. An interesting aspect when comparing results in the top half and bottom half of Table 1 is that the model trained with Buyer+Seller features from both speakers (i.e. LIWC+Politeness speaker-dependent features) has a better performance, both in terms of accuracy and F1-score, than the best model trained with speaker-independent features (COLL+LIWC+Politeness). Nevertheless, from the models trained with speaker-dependent features, only the buyer features achieved a R^2 above 0.2, the threshold to be considered a good fit between the trained model and the data. Therefore, when looking at the top speaker-dependent features for best performing model, some caution is warranted.

5.2 Persuasion Dialogues

For this dataset, we trained a LR to predict the persuasiveness, i.e., whether a donation was made by the persuadee. The Persuasion for Good dataset was not released with fixed splits, therefore we adopted a 5-fold cross-validation procedure following previous work with this dataset (Wang et al., 2019). In Table 2, we present the average accuracy and F1-score for all folds and their standard deviation. For each fold, we have saved the respective feature coefficients. Given that the metrics computed for the models have a small standard deviation, we assume that models in each fold are relatively similar and thus averaged the coefficient values for every feature across the 5 folds. The features presented in the tables are those with the highest absolute average coefficient values across folds. Similarly to the negotiation dataset, we also report the average R^2 across the different folds and respective standard deviation.

Using lexicon-based features, we observed a marginal improvement in terms of accuracy, when compared with the baseline majority, except when using only LIWC features. In the set of COLL features, the number of repeated content (avg_n_repeated_content) and stop words (avg_n_repeated_stop) in consecutive turns, and the number of agreement words (avg_agree) contributed to predicting persuasiveness. A high number of direct questions was one of the most valuable features to predict unpersuasive dialogues in the model trained with Politeness features (this feature

Features	Accuracy	F1-score	R^2	Top-5 features
Baseline Majority	0.536 (0.001)	**0.698 (0.001)**	-	-
COLL	**0.571 (0.029)**	0.653 (0.022)	-0.088 (0.063)	+avg_agree +avg_n_repeated_content +avg_n_repeated_stop -fit_disagree +fit_repeated_stop
LIWC	0.500 (0.044)	0.553 (0.033)	-0.107 (0.125)	-avg_geo +coordination_score +avg_n_adopted +avg_n_introduced +avg_n_introduced_w_hedge
Politeness	0.568 (0.031)	0.626 (0.049)	-0.088 (0.160)	+avg_has_positive -avg_direct_question -avg_has_negative +avg_gratitude +avg_2nd_person_start
COLL + LIWC + Politeness	0.556 (0.039)	0.591 (0.038)	**0.025 (0.058)**	-avg_geo -avg_has_negative +avg_has_positive +avg_agree -avg_direct_question

Table 2: Mean Accuracy, F1-score and McFadden's R^2 for predicting persuasion in the Persuasion for Good Dataset in the 5-folds. The figure between brackets represent the standard deviation across the different folds. The top-5 features are sorted according to the mean of absolute coefficient values.

was automatically detected by the occurrence of the initial wordings of "what, why, who or how").

6 Opaque Models for Prediction of Persuasion and Negotiation

As well as the traditional lexicon-based features described above, we have also used embedding-based features, specifically: RoBERTa-SE sentence embeddings (Reimers and Gurevych, 2019) trained for the STS task[1]; and a dialogue vector representation extracted from a ConvERT model (Henderson et al., 2019). For sentence-based models (RoBERTa-SE), for each turn an embedding was generated. The dialogue representation is then the average of the sentence embeddings for all dialogue turns. In the ConvERT model, given the context and the current utterance, the model would provide a dialogue embedding. We compare these two models in order to assess the impact of using a model that attempts to keep the sequential structure of the data (ConvERT) versus a model trained with a larger amount of data (RoBERTa-SE).

These embeddings were given as inputs either to a LR or a neural model composed by a linear layer and a softmax layer, which provides the probability distribution of the different classes (Linear-NN). The reasoning behind this was to see if the neural model was better at predicting whether the dialogue resulted in successful negotiation, even though this method is less transparent than LR.

Results from embedding-based dialogue representations predicting negotiation success in the Craigslist Bargain dataset are shown in Table 3. The fact that ConvERT keeps the sequential structure of the data seems to provide an advantage over RoBERTa-SE in terms of F1 and accuracy. It is interesting to observe a drop in performance from the LR-models to the NN-models. In any case, models based on pre-trained dialogue representations seem to improve the performance over models trained with lexicon-based features using LR, as well as observing a higher R^2 (as reported in Table 1).

For persuasiveness prediction, the neural models trained with ConvERT (see Table 4) outperform those using LR with embedding features and also LR with linguistic features (see Table 2). However, again, the disadvantage of this approach is that these models are less transparent.

7 Discussion

As we look at the features, we find some interesting results. Tables 5 and 6 show an example dialogue and corresponding features, from a Craigslist Bargain and a Persuasion for Good dialogue respectively. One of the features emerging as potentially contributing to no deal was an increasing number of words per utterance (fit_n_words) as the dialogue progresses (see Table 1). In the example of a dialogue where a deal was reached, shown in Table 5, there is a tendency for short utterances as the dialogue unfolds. One of the factors associated with increasing trustworthiness is transparency (Nesset et al., 2021). However, a direct consequence of increasing transparency in dialogue is an increase in the number of words per sentence. This seems an interesting avenue for future research, to instill the appropriate amount of trust while keeping the utterance short, along with the appropriate level of transparency.

Another interesting outcome is that the average number of apologetic words were higher in no deal dialogue compared to dialogues where a deal was reached. This may be due to the fact that people

[1] https://github.com/UKPLab/sentence-transformers

Features	Model	Accuracy	F1-score	R^2
RoBERTa-SE	LR	0.854	0.906	**0.560**
ConvERT	LR	**0.895**	**0.932**	0.533
RoBERTa-SE	Linear-NN	0.843	0.904	-
ConvERT	Linear-NN	0.859	0.913	-

Table 3: Accuracy, F1-score and McFadden's R^2 (in the LR models) for prediction negotiation success in the Craigslist Bargain dataset using dialogue embeddings.

Features	Model	Accuracy	F1-score	R^2
RoBERTa-SE	LR	0.611 (0.038)	0.638 (0.052)	0.050 (0.331)
ConvERT	LR	0.602 (0.022)	0.665 (0.027)	0.120 (0.003)
RoBERTa-SE	Linear-NN	0.607 (0.010)	**0.724 (0.013)**	-
ConvERT	Linear-NN	**0.622 (0.018)**	0.715 (0.004)	-

Table 4: Average accuracy, F1-score and McFadden's R^2 (for the LR models) in the 5 folds for predicting persuasion in the Persuasion for Good dataset using dialogue embeddings. Number between brackets is the standard deviation in the 5 folds.

Buyer: I am interested in purchasing this item, but since it is used I can only afford to pay about 25 **Seller**: I mean, we can work out a deal, but that is way too low. how about 60? **Buyer**: Shoot, I only have about 40 in my account right now.	avg_agree = 0.0 fit_gap = 0.016 fit_n_repeated_pos_bigram = −0.333 avg_n_repeated_content = 0.0 fit_n_repeated_stop = −0.333 **fit_n_words** = −0.214 avg_n_certain = 0.0 avg_n_geo = 0.0 avg_n_introduced_w_hedge = 0.0 avg_n_introduced = 0.0 **avg_has_negative** = 0.0 **avg_apologising** = 0.0 avg_indicative = 0.0 avg_direct_start = 0.0 avg_indirect_greeting = 0.0 **avg_has_positive** = 1.0 **avg_gratitude** = 0.0

Table 5: Example of a dialogue where a deal was reached from the Craigslist Bargain dataset with corresponding feature values. Underlined words have direct impact in the feature values reported. Top-5 features of COLL+LIWC+Politeness model in bold from Table 1.

apologise for the negotiation not being successful or being unable to adjust the price to the other person's requested price.

Collaborative features seem to be more important for success in persuasion than negotiation dialogues (see Tables 1 and 2) when compared to other lexicon-based features. This could be explained by the fact that the persuasion needs a high amount of collaboration, where both participants could benefit from a positive outcome, whereas in Craigslist Bargain the task of negotiation is competitive and both users have to compromise to achieve a trade-off.

The number of geographical-related words, given by the avg_geo feature seems to be influential, which is perhaps non-obvious. Since none of the datasets are likely to have a large number of geographical references (even if there is a section of Craigslist Bargain about housing). It could be that the geographical lexicon has several polysemic words (e.g. the word physical in part of this lexicon) and therefore introduces some extra noise in the data. This could be an example of lexicon that could be filtered out in future work.

In both corpora, the number of positive words and agreement expressions contributed to predicting the success of the dialogue. Interestingly, some of the most predictive features are not explicitly related to the success of task (such as the use of agreement words), but rather relate more to the language common to the various roles, e.g., whether the seller uses the first person or whether the buyer starts a sentence with a second person pronoun. On the other hand, gratitude has been one of the features which contributed to a correct prediction of a deal. Social norms might explain this, since it is considered polite to thank the other person at the end of the negotiation. Attempting to predict the outcome of the dialogue in earlier stages could help us to confirm this hypothesis.

The performance of the models is generally higher in the Craigslist than the Persuasion for Good dataset, which could likely be due to the larger size of the Craigslist dataset. Nevertheless, the set of features used by the best performing models with lexicon-based features (COLL+LIWC+Politeness and Buyer+Seller Features) was rather unexpected. One perhaps would have thought that features such as agreements/disagreements would be natural candidates to be relevant features for the task at hand, but these were not part of the top-5 most relevant feature when combining all the lexicon-based features. Also the best performance for lexicon-based feature

Persuader: Hello **Persuadee**: Hi there how are you? **Persuader**: Great and you **Persuadee**: I am doing fine, thanks for asking. **Persuader**: Can I ask you a question? **Persuadee**: You sure can **Persuader**: Do you have any kids? **Persuadee**: I don't have any children but there are many children in my family. **Persuader**: Im sure you love them as your own, I have a big family do you? **Persuadee**: I do have a big family, I enjoy it at times, other times I don't **Persuader**: I truly understand. But I know I would do anything for them especially if it was for their wellbeing, do you feel the same way **Persuadee**: Yes I do. **Persuader**: It's great to know that there are others who feel the same way. Have you heard of Save the Children? **Persuadee**: No I haven't. Could you QUICKLY tell me about it **Persuader**: Save the Children is an international non-governmental organization that promotes children's rights, provides relief and helps support underlinechildren in developing countries. **Persuadee**: Okay that sounds nice and an important service **Persuader**: And the money raised helps feed and clothe them. Its a lot of underlinechildren that are starving and need our help. Would like to help? **Persuadee**: I would like to help in the future when I am more financially stable. **Persuader**: I understand but even the smallest amount would be a BIG help. **Persuadee**: I am sure but I just am not able at this time	**avg_agree** $= 0.053$ avg_n_repeated_content $= 0.053$ avg_n_repeated_stop $= 2.053$ fit_disagree $= NaN$ fit_n_repeated_stop $= 1.285$ **avg_geo** $= 0.0$ coordination_score $= NaN$ avg_n_adopted $= 0.150$ avg_n_introduced $= 0.150$ avg_n_introduced_w_hedge $= 0.0$ **avg_has_positive** $= 2.2$ **avg_direct_question** $= 0.0$ **avg_has_negative** $= 0.0$ avg_gratitude $= 0.050$ avg_2nd_person_start $= 0.05$

Table 6: Example of unsuccessful dialogue from the Persuasion for Good dataset with corresponding feature values. Underlined words have direct impact in the feature values reported. Top-5 features of COLL+LIWC+Politeness model in bold from Table 2.

models in the Craigslist was achieved by separating the seller from the buyer features. This is an interesting outcome and reinforces the different roles of each speaker in the dialogue.

Finally, initial results suggest that dialogue embeddings are powerful representations that can be used to predict the outcome of the dialogue. In fact, for LR trained with dialogue embeddings, the R^2 was above 0.2 for negotiations, unlike most of the cases using lexicon-based features, which shows a better fit to the data. However, interpretable features and models can provide more explainable and transparent cues.

8 Conclusion and Future Work

We have investigated linguistic indicators that reflect two tasks, namely a successful negotiation and persuasion of a donation. These two interaction outcomes can be seen as examples of conditional trust (Jones and George, 1998), since they involve social and/or economic exchanges. In the case of negotiation, the task is competitive, whereas persuasion dialogues can be considered more of a collaboration. Various lexicon-based features were identified as being indicators of success through our method of training regressors. However, a role-based analysis showed differences in the relevant features. Therefore, considering the role will be important when designing trustworthy conversational agents. Future work will look into individual differences more deeply and explore variations of personality and propensity to trust of individual users.

Methods based on dialogue embeddings achieved the best performance in both problems, however these methods are opaque. Future work would involve combining recent work on transparent NLP methods for explaining embedding models (Hoover et al., 2020) and explainable AI (Ribeiro et al., 2016), so as to provide further insight into linguistic and dialogue features for these opaque but high performing features and models.

In the introduction, we mentioned that in both datasets used in this research we were using proxies for trust assuming that financial transaction between subject would only occur when a certain level of trust is achieved. This is a limitation of our work, which are trying to address at the moment by collecting trustworthiness ratings at turn level. This will allow us to confirm whether our assumption is correct and develop a fine-grained strategy to increase trustworthiness in conversational agents.

Finally, a discussion on the ethical implications is needed of using interactive systems for these types of interactions, where trust is conditional on the perceived behaviour of system.

Acknowledgements

This work was funded and supported by the EPSRC ORCA Hub (EP/R026173/1) and UKRI Node on Trust (EP/V026682/1).

References

Daniel Balliet and Paul AM Van Lange. 2013. Trust, conflict, and cooperation: a meta-analysis. *Psychological bulletin*, 139(5):1090.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Colin F Camerer. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Kushal Chawla, Gale M. Lucas, Jonathan Gratch, and Jonathan May. 2020. BERT in negotiations: Early prediction of buyer-seller negotiation outcomes. *CoRR*, abs/2004.02363.

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.

David DeSteno, Cynthia Breazeal, Robert H. Frank, David Pizarro, Jolie Baumann, Leah Dickens, and Jin Joo Lee. 2012. Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, 23(12):1549–1556. PMID: 23129062.

Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. How humans versus bots react to deceptive and trusted news sources: A case study of active users. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '18, page 654–661. IEEE Press.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Matthew Henderson, Inigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A visual analysis tool to explore learned representations in Transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71.

Gareth R. Jones and Jennifer M. George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, 23(3):531–546.

Halimahtun Khalid, Wei Shiung Liew, Bin Sheng Voong, and Martin Helander. 2019. Creativity in measuring trust in human-robot interaction using interactive dialogs. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, pages 1175–1190, Cham. Springer International Publishing.

Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports*, pages 1–13.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 107–116, New York, NY, USA. Association for Computing Machinery.

Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Trans. Manage. Inf. Syst.*, 4(4).

Gale Lucas, Giota Stratou, Shari Lieblich, and Jonathan Gratch. 2016. Trust me: Multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 5–12, New York, NY, USA. Association for Computing Machinery.

Johnathan Mell, Markus Beissinger, and Jonathan Gratch. 2019. An expert-model & machine learning hybrid approach to predicting human-agent negotiation outcomes. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 212–214, New York, NY, USA. Association for Computing Machinery.

Johnathan Mell and Jonathan Gratch. 2017. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, pages 401–409.

Birthe Nesset, David A. Robb, José Lopes, and Helen Hastie. 2021. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 313–317, New York, NY, USA. Association for Computing Machinery.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675. PMID: 15272998.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California. Association for Computational Linguistics.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1650–1659, Beijing, China. Association for Computational Linguistics.

James W Pennebaker. 2001. Linguistic inquiry and word count: LIWC 2001.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

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.

John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404.

Kristin Schaefer. 2013. *The perception and measurement of human-robot trust*. Ph.D. thesis.

Lauren E. Scissors, Alastair J. Gill, and Darren Gergle. 2008. Linguistic mimicry and trust in text-based cmc. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, page 277–280, New York, NY, USA. Association for Computing Machinery.

Ilaria Torre, Emma Carrigan, Killian McCabe, Rachel McDonnell, and Naomi Harte. 2018. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 423–427, New York, NY, USA. Association for Computing Machinery.

Daniel Ullman and Bertram F. Malle. 2018. What does it mean to trust a robot? steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 263–264, New York, NY, USA. Association for Computing Machinery.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Ran Zhao, Oscar J. Romero, and Alex Rudnicky. 2018. Sogo: A social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, page 239–246, New York, NY, USA. Association for Computing Machinery.

A Linguistic indicators reference

Tables 7, 8 and 9 has a complete reference of all lexical features used.

Feature	Definition
n_repeated_pos_bigram	number of repeated POS bigrams in consecutive turns
n_repeated_content	number of repeated content words in consecutive turns
agree	whether there is an agreement expression
disagree	whether there is a disagreement expression
n_repeated_stop	number of repeated stop words in consecutive turns.
coordination score	coordination score between the two speakers

Table 7: Collaborative (COLL) indicators reference.

Feature	Definition
n_adopted_w_hedge	number of w ords re-used from hedges lexicon
n_words	number of words per utterance
n_introduced	total number of words re-used
n_adopted_w_certain	number of words re-used from certain lexicon
n_introduced_w_hedge	number of newly introduced words from the hedges lexicon
pron_we	number of usages of words from the we lexicon
geo	number of usages of words from the geographic terms lexicon
hedge	number of words from the hedges lexicon
n_introduced_w_certain	number of newly introduced words from the certain lexicon
pron_you	number of words from the you lexicon
meta	number of words from the meta lexicon
pron_me	number of words from the me lexicon
n_adopted	number of re-used words
pron_3rd	number of words from the

Table 8: LIWC indicators reference.

Feature	Definition
please_start	if utterance starts with please
factuality	if utterance has factuality expressions (e.g. actually)
apologising	if utterance contains apologetic words
2nd_person	if utterance contains second person words
please	if utterance contains please
direct_question	if utterance starts with what, why, who or how
gratitude	if utterance contains gratitude words
has_positive	if utterance as positive words
1st_person_start	if utterance starts with a first person pronoun
1st_person	if utterance has first person pronouns
1st_person_pl.	if utterance contains first person plural pronouns
subjunctive	if utterance includes 'could' or 'would' before 'you'
indicative	if utterance includes 'can' or 'will' before 'you'
direct_start	if utterance has a direct start
indirect_(greeting)	if utterance starts with 'hi', 'hello' and 'hey'
has hedge	if utterance has hedges
indirect_(btw)	if utterance contains expression 'by the way'
has_negative	if utterance has negative words
deference	if utterance has deference words
2nd_person_start	if uttreance starts with a second person pronouns

Table 9: Politeness Linguistic indicators reference.

Dialogue act classification is a laughing matter

Vladislav Maraev
University of Gothenburg
vladislav.maraev@gu.se

Bill Noble
University of Gothenburg
bill.noble@gu.se

Chiara Mazzocconi
Aix-Marseille University
chiara.mazzocconi@live.it

Christine Howes
University of Gothenburg
christine.howes@gu.se

Abstract

In this paper we explore the role of laughter in attributing communicative intents to utterances, i.e. detecting the dialogue act performed by them. We conduct a corpus study in adult phone conversations showing how different dialogue acts are characterised by specific laughter patterns, both from the speaker and from the partner. Furthermore, we show that laughs can positively impact the performance of Transformer-based models in a dialogue act recognition task. Our results highlight the importance of laughter for meaning construction and disambiguation in interaction.

1 Introduction

Laughter is ubiquitous in our everyday interactions. In the Switchboard Dialogue Act Corpus (SWDA, Jurafsky et al., 1997a) (US English, phone conversations where two participants that are not familiar with each other discuss a potentially controversial subject, such as gun control or the school system) non-verbally vocalised dialogue acts (whole utterances that are marked as non-verbal, 66% of which contain laughter) constitute 1.7% of all dialogue acts. Laughter tokens[1] make up 0.5% of all the tokens that occur in the corpus. Laughter relates to the discourse structure of dialogue and can refer to a *laughable*, which can be a perceived event or an entity in the discourse. Laughter can precede, follow or overlap the laughable, and the time alignment between them depends on who produces the laughable, the form of the laughter, and the pragmatic function performed (Tian et al., 2016).

Bryant (2016) shows how listeners are influenced towards a non-literal interpretation of sentences when accompanied by laughter. Similarly, Tepperman et al. (2006) shows that laughter can act as a contextual feature for determining the sincerity of an utterance, e.g. when detecting sarcasm.

Nevertheless there is a dearth of research exploring the use of laughter in relation to different dialogue acts in detail, and therefore little is known about the role that laughter may have in facilitating the detection of communicative intentions.

Based on previous work and the corpus study presented in this paper, we argue that laughter is tightly related to the information structure of a dialogue. In this paper, we investigate the potential of laughter to disambiguate the meaning of an utterance, in terms of the dialogue act it performs. To do so, we employ a Transformer-based model and look into laughter as a potentially useful feature for the task of *dialogue act recognition* (DAR). Laughs are not present in a large-scale pre-trained models, such as BERT (Devlin et al., 2019), but their representations can be learned while training for a dialogue-specific task (DAR in our case). We further explore whether such representations can be additionally learned, in an unsupervised fashion, from dialogue-like data, such as a movie subtitles corpus, and if it further improves the performance of our model.

The paper is organised as follows. We start with some brief background in Section 2. In Section 3 we observe how dialogue acts can be classified with respect to their collocations with laughs and discuss the patterns observed in relation to the pragmatic functions that laughter can perform in dialogue. In Section 4 we report our experimental results for the DAR task depending on whether the model includes laughter. We further investigate whether non-verbal dialogue acts can be classified as more specific dialogue acts by our model. We conclude with a discussion and outlining the directions for further work in Section 5.

[1] Switchboard Dialogue Act Corpus does not include speech-laughs.

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

2 Background

2.1 Laughter

Laughter does not occur only in response to humour or in order to frame it. It is crucial in managing conversations in terms of dynamics (turn-taking and topic-change), at the lexical level (signalling problems of lexical retrieval or imprecision in the lexical choice), but also at a pragmatic (marking irony, disambiguating meaning, managing self-correction) and social level (smoothing and softening difficult situations or showing (dis)affiliation) (Glenn, 2003; Jefferson, 1984; Mazzocconi, 2019; Petitjean and González-Martínez, 2015).

Moreover Romaniuk (2009) and Ginzburg et al. (2020) discuss how laughter can answer or decline to answer a question, and Mazzocconi et al. (2018) explore laughter as an object of clarification requests, signalling the need for interlocutors to clarify its meaning (e.g., in terms of what the "laughable" is) to carry on with the conversation.

2.2 Dialogue act recognition

The concept of a dialogue act (DA) is based on that of the speech act (Austin, 1975). Breaking with classical semantic theory, Speech Act Theory considers not only the propositional content of an utterance but also the actions, such as *promising* or *apologising*, it carries out. Dialogue acts extend the concept of the speech act, with a focus on the interactional nature of most speech. DAMSL (Core and Allen, 1997), for example, is an influential DA tagging scheme where DAs are defined in part by whether they have a *forward-looking* function (expecting a response) or *backward-looking* function (in response to a previous utterance).

Dialogue act recognition (DAR) is the task of labelling utterances with the dialogue act they perform, given a set of dialogue act tags. As with other sequence labelling tasks in NLP, some notion of context is helpful in DAR. One of the first performant machine learning models for DAR was a Hidden Markov Model that used various lexical and prosodic features as input (Stolcke et al., 2000).

Recent state-of-the-art approaches to dialogue act recognition have used a hierarchical approach, using large pre-trained language models like BERT to represent utterances, and adding some representation of discourse context at the dialogue level (e.g., Ribeiro et al., 2019; Mehri et al., 2019). However Noble and Maraev (2021) observe that without fine-tuning, standard BERT representations perform very poorly on dialogue, even when paired with a discourse model, suggesting that certain utterance-internal features missing from BERT's textual pre-training data (such as laughter) may have an adverse effect on dialogue act recognition.

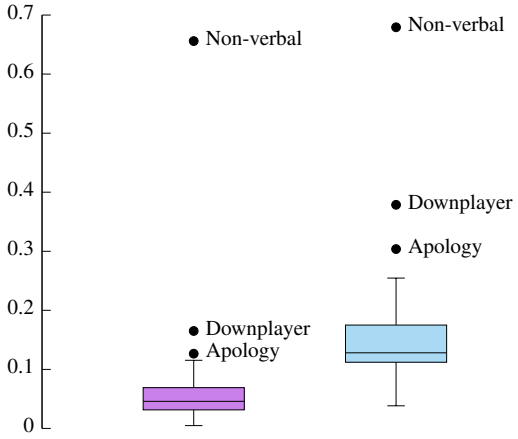


Figure 1: Box plots for proportions of dialogue acts which contain laughs in SWDA. On the left: proportion of DAs containing laughter, on the right: proportion of DAs having laughter in one of the adjacent utterances.

3 Laughs in the Switchboard Dialogue Act Corpus

In this section we analyse dialogue acts in the Switchboard Dialogue Act Corpus according to their collocation with laughter and provide some qualitative insights based on the statistics.

SWDA is tagged with a set of 220 dialogue act tags which, following Jurafsky et al. (1997b), we cluster into a smaller set of 42 tags.

The distribution of laughs in different dialogue acts has a rather uniform shape with a few outliers (Figure 1). The most distinct outlier is the *Non-verbal* dialogue act which is misleading with respect to laughter, because utterances only containing a single laughter token fall into this category. However isolated laughs can serve, for example, to acknowledge a statement, to deflect a question, or to show appreciation (Mazzocconi, 2019). We will further conjecture on this class of DAs in Sec. 4.5.

3.1 Method

Let us illustrate our comparison schema using the other two outliers, *Downplayer* (make up 0.05% of all utterances) and *Apology* (0.04%), comparing them with the most common dialogue act in SWDA – *Statement-Non-Opinion* (33.27%). We consider laughter-related dimensions of an utterance, and

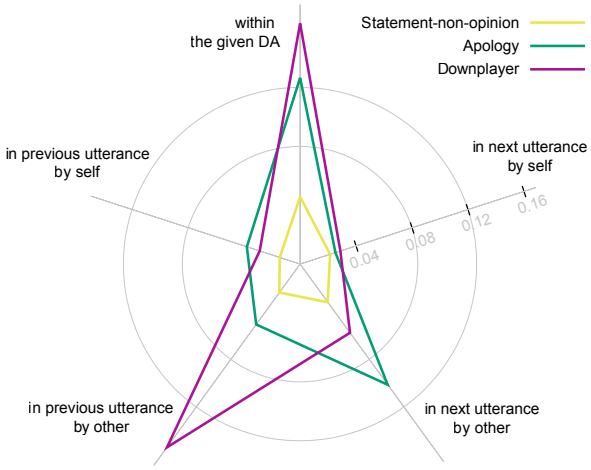


Figure 2: Comparison between the pentagonal representations of laughter collocations of dialogue acts.

create 5-dimensional (pentagonal) representations of DAs according to them. Each dimension's value is equal to the proportion of utterances of a given type which contain laughter:

- ↑ current utterance;
- ↖ immediately preceding utterance by the same speaker;
- ↗ immediately following utterance by the same speaker;
- ↙ immediately preceding utterance by the other speaker;
- ↘ immediately following utterance by the other speaker.

For instance, (1) is an illustrative example of the phenomenon shown in Figure 2.

(1) [2] **A:** I'm sorry to keep you waiting #<laughter>.# *Apology*
B: #Okay# <laughter>. / *Downplayer*
A: Uh, I was calling from work *Statement (n/o)*

We show the representations of all dialogue acts on Figure 3. We believe that such a depiction helps the reader form impressions about similarities between DAs based on their laughter collocations and notice the ones that stand out in some respects.

To further assess the similarity between dialogue acts based on their collocations with laughs we factorise their pentagonal representations into 2D space using singular value decomposition (SVD). We can see that dialogue acts form some distinct clusters. The resulting plot is shown in Figure 6 in Appendix A.1. Let us now proceed with some qualitative observations.

[2]Overlapping material is marked with hash signs.

3.2 Observations

Laughter and modification or enrichment of the current DA We observe a higher proportion of laughter accompanying the current dialogue act (↑) when the laughter is aimed at modifying the current dialogue act with some degree of urgency to smooth or soften it (*Action-directive*, *Reject*, *Dispreferred answer*, *Apology*), to contribute to its enrichment stressing the positive disposition towards the partner (*Appreciation*, *Downplayer*, *Thanking*), or to cue for the need to consider a less probable meaning, therefore helping in non-literal meaning interpretation (*Rhetorical question*).

While *Apology* and *Downplayer* have rather distinct and peculiar patterns (Fig. 6) discussed in more detail below, we observe *Dispreferred answers*, *Action directives*, *Offers/Options/Commits* and *Thanking* to constitute a close cluster when considering the decomposed values of the pentagonal used for DA representation.

Laughter for benevolence induction and laughter as a response The patterns observed in relation to the preceding and following turns reflect the multitude of functions that laughter can perform in interaction, stressing the fact that it can be used both to induce or invite a determinate response (dialogue act) from the partner (*Downplayer*, *Agree/Accept*, *Appreciation*, *Acknowledge*) as well as being a possible answer to specific dialogue acts (e.g. *Apology*, *Offers/Options/Commits*, *Summarise/Reformulate*, *Tag-question*).

A peculiar case is the one of *Self-talk*, often followed by laughter by the same speaker. In this case the laughter may be produced to signal the incongruity of the action (in dialogue we normally speak to others, not to ourselves), while at the same time function to smooth the situation, for instance, when having issues of lexical retrieval, as in (2), or some degree of embarrassment from the speaker, when questioning whether a contribution is appropriate or not, as in (3).

(2) **A:** Have, uh, really, -
A: what's the word I'm looking for, *Self-talk*
A: I'm just totally drawing a blank <laughter>. *Statement (n/o)*

(3) **B:** Well, I don't have a Mexi-, - *Statement (n/o)*
B: I don't, shouldn't say that, *Self-talk*
B: I don't have an ethnic maid <laughter>. *Statement (n/o)*

Apology and Downplayer It is interesting to comment on the parallelisms of laughter usage in

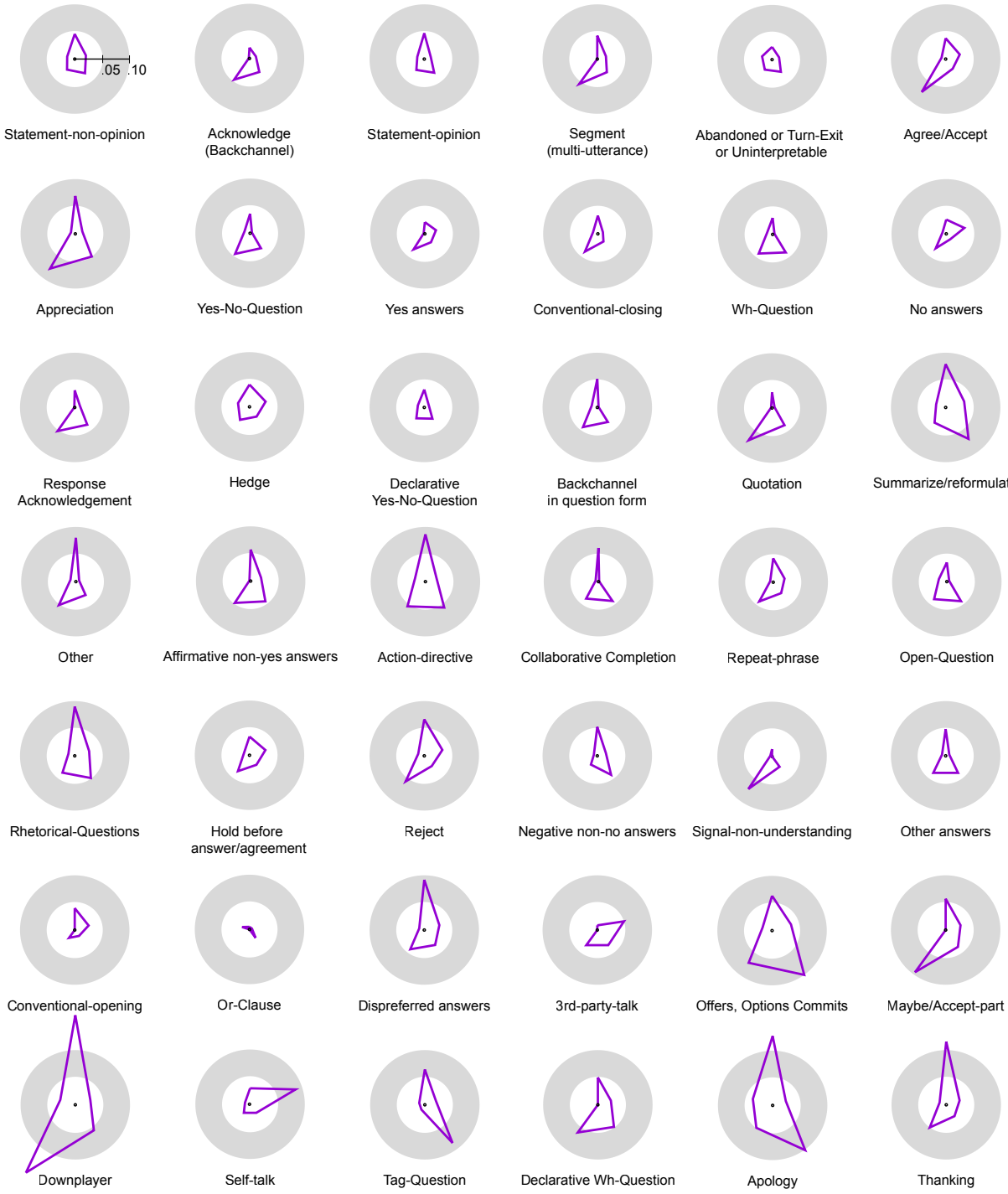


Figure 3: Pentagonal representation of dialogue acts: proportions of utterances which include laughter. Dimensions: ↑ current utterance; ↖ immediately preceding utterance by the same speaker; ↗ immediately following utterance by the same speaker; ↙ immediately preceding utterance by the other speaker; ↘ immediately following utterance by the other speaker. DAs are ordered by their frequency in SWDA (left-to-right, then top-to-bottom).

relation to *Apology* and *Downplayer*, represented in Fig. 2 in contrast to *Statement-non-opinion*, in as much as their graphic representations are more or less mirror-images of each other and show how the dialogue acts are linked by the pragmatic functions laughter can perform in dialogue.

In both *Apology* and *Downplayer* we observe a rather higher proportion of occurrences in which the dialogue act is accompanied by laughter (↑) in comparison to other DAs (Fig. 3). In the case of *Apology*, laughter can be produced to induce benevolence from the partner (Mazzocconi et al., 2020), while in the case of *Downplayer* the laughter can be produced to reassure the partner about some

situation that had been appraised as discomforting (classified as *social incongruity* by Mazzocconi et al., 2020) and somehow signal that the issue should be regarded as not important (Romaniuk, 2009; **?**), as in (4).

(4) **A:** I don't, I don't think I could do that <laughter>. # *Statement (n/o)*
B: Oh, it's not bad at all. *Downplayer*
A: It's, it's a beautiful drive. *Statement (n/o)*

The interesting mirror-image patterns observable in the lower part of the graph can therefore be explained by considering the relation between the two dialogue acts. We observe cases in which an *Apology* is accompanied by a laughter, and then followed by a *Downplayer*, showing that the laughter's positive effect was attained and successful. This allows us to explain both the bottom left spike (↙) observed for *Downplayer* (often preceded by an utterance by the partner containing laughter) and the bottom right spike (↘) observed for *Apology* (often followed by an utterance by the partner containing laughter). In example (1) both the apology and the downplayer are accompanied by laughter, while in (5) a typical example of a laughter accompanying an *Apology* is followed by a *Downplayer*.

(5) **B:** I'm sorry <laughter>. # *Apology*
A: That's all right. / *Downplayer*
B: You, you were talking about, uh, uh, *Summarise*

We now turn to the question of whether our qualitative observations of patterns between laughs and dialogue acts can be used to improve a dialogue act recognition task.

4 The importance of laughter in artificial dialogue act recognition

4.1 Data

We perform experiments on the Switchboard Dialogue Act Corpus (SWDA, 42 dialogue act tags), which is a subset of the larger Switchboard corpus, and the dialogue act-tagged portion of the AMI Meeting Corpus (AMI-DA). AMI uses a smaller tagset of 16 dialogue acts (Gui, 2005).

Preprocessing We make an effort to normalise transcription conventions across SWDA and AMI. We remove disfluency annotations and slashes from the end of utterances in SWDA. In both corpora, acronyms are tokenised as individual letters. All utterances are lower-cased.

Utterances are tokenised using a word piece tokeniser (Wu et al., 2016) with a vocabulary of 30,000. We add a special laughter token to the vocabulary and map all transcribed laughter to that token. We also prepend each utterance with a speaker token that uniquely identifies the corresponding speaker within that dialogue.

Switchboard	**AMI Corpus**
Dyadic	Multi-party
Casual conversation	Mock business meeting
Telephone	In-person & video
English	English
Native speakers	Native & non-native speakers
2200 conversations	171 meetings
1155 in SWDA	139 in AMI-DA
400k utterances	118k utterances
3M tokens	1.2M tokens

Table 1: Comparison between Switchboard and the AMI Meeting Corpus

4.2 The model

To test the effectiveness of BERT for DAR, we employ a simple neural architecture with two components: an encoder that vectorises utterances, and a sequence model that predicts dialogue act tags from the vectorised utterances (Figure 4). Since we are primarily interested in comparing different utterance encoders, we use a basic RNN as the sequence model in every configuration.[3] The RNN takes the encoded utterance as input at each time step, and its hidden state is passed to a simple linear classification layer over dialogue act tags. Conceptually, the encoded utterance represents the context-agnostic features of the utterance, and the hidden state of the RNN represents the full discourse context.

As a baseline utterance encoder, we use a word-level CNN with window sizes of 3, 4, and 5, each with 100 feature maps (Kim, 2014). The model uses 100-dimensional word embeddings, which are initialised with pre-trained gloVe vectors (Pennington et al., 2014). For the BERT utterance encoder, we use the $\text{BERT}_{\text{BASE}}$ model with hidden size of 768 and 12 transformer layers and self-attention heads (Devlin et al., 2018, §3.1). In our implementation, we use the un-cased model provided by Wolf et al. (2019).

4.3 Experiment 1: Impact of laughter

In the first experiment we investigated whether laughter, as an example of a dialogue-specific signal, is a helpful feature for DAR. Therefore, we

[3] We have experimented with LSTM as the sequence model, but the accuracy was not significantly different compared to RNN. It can be explained by the absence of longer distance dependencies on this level of our model.

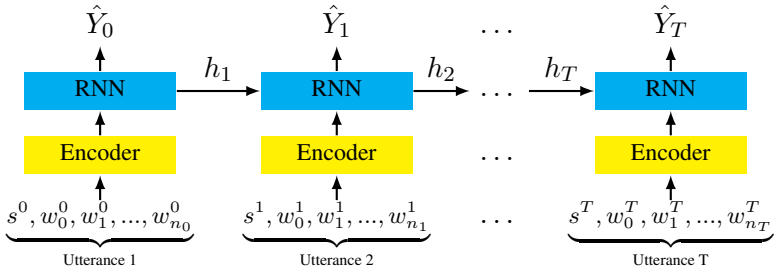


Figure 4: Simple neural dialogue act recognition sequence model

	SWDA		AMI-DA	
	F1	acc.	F1	acc.
BERT-NL	36.48	76.00	44.75	68.04
BERT-L	36.75	76.60	43.37	64.87
CNN-NL	36.95	73.92	38.00	63.18
CNN-L	37.59	75.40	37.89	64.27
Majority class	0.78	33.56	1.88	28.27

Table 2: Comparison of macro-average F1 and accuracy depending on using laughter on the training phase.

train another version of each model: one containing laughs (L) and one with laughs left out (NL), and compare their performances in DAR task. Table 2 compares the results from applying the models with two different utterance encoders (BERT, CNN).

BERT outperforms the CNN on AMI-DA. On SWDA, the two encoders are more comparable, though BERT has a slight edge in accuracy, suggesting that it relies more heavily on defaulting to common dialogue act tags. On SWDA, we see small improvements in accuracy and macro-F1 for models that included laughter. For AMI-DA, the effect of laughter is small or even negative – the impact of laughter on performance becomes more clear in the disaggregated performance over different dialogue acts. Indeed, laughter improves the accuracy of the model even on some dialogue acts in which laughter occurs rarely in the current and adjacent utterances (see Figure 7 in Appendix A).

Confusion matrices (Figure 5) provide some food for thought. Most of the misclassifications fall into the majority classes, such as *sd* (Statement-non-opinion), on left edge of the matrix. However, there are some important exceptions, such as *rhetorical questions*, that are misclassified as other forms of questions due to their surface question-like form. Importantly, laughter helps to classify rhetorical questions correctly, this is because in a conversation it can be used as a device to cancel seriousness or reduce commitment to literal meaning (Ginzburg et al., 2015; Tepperman et al., 2006)

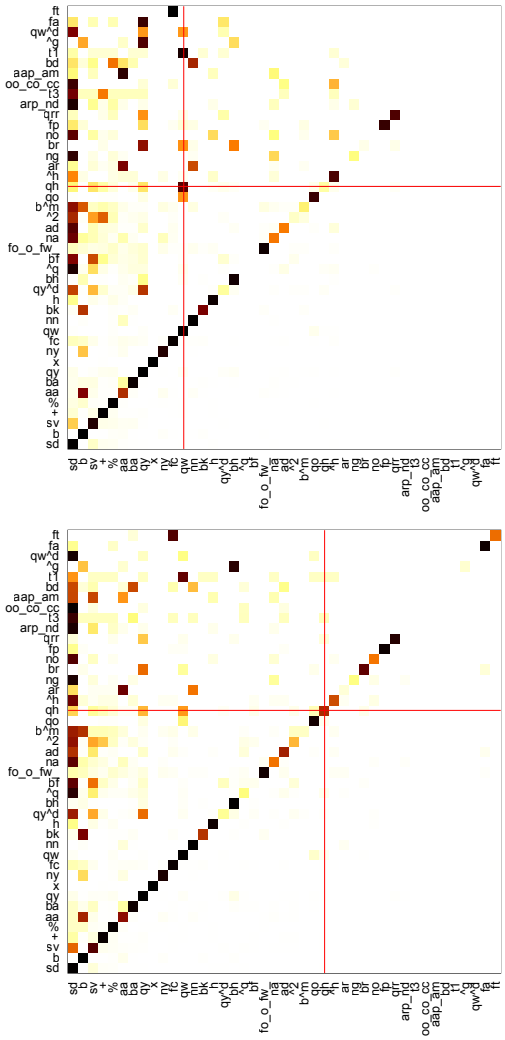


Figure 5: Confusion matrices for BERT-NL (top) vs BERT-L (bottom); SWDA corpus. Solid lines show classification improvement of rhetorical questions.

Therefore, questions, like the one we show in example (6), are easier to disambiguate with laughter.

(6) **B:** Um, as far as spare time, they talked about, *Statement (n/o)*
B: I don't, + I think, *Statement (n/o)*
B: who has any spare time <laughter>? *Rhetorical Quest.*
A: <laughter>. *Non-verbal*

4.4 Experiment 2: laughter and pre-training

As previously noted, training data for BERT does not include features specific to dialogue (e.g. laughs). We therefore experiment with a large and more dialogue-like corpus constructed from OpenSubtitles (Lison and Tiedemann, 2016) (350M tokens, where 0.3% are laughter tokens). We used a manually constructed list of words frequently used to refer to laughter in subtitles and replaced every occurance of one of these words with the special laughter token. We then collected every English-language subtitle file in which at least 1% of the utterances contained laughter (about 11% of the total). Because utterances are not labelled with speaker in the OpenSubtitles corpus, we randomly assigned a speaker token to each utterance to maintain the format of the other dialogue corpora.

The pre-training corpus was prepared for the combined masked language modelling and next sentence (utterance) prediction task, as described by Devlin et al. (2018).

We analyse how pre-training affects BERT's performance as an utterance encoder. To do so, we consider the performance of DAR models with three different utterance encoders: i) `FT` – pre-trained BERT with DAR fine-tuning; ii) `RI` – randomly initialised BERT (with DAR fine-tuning); iii) `FZ` – pre-trained BERT without fine-tuning (frozen during DAR training). For the pre-trained (`FT`, `FZ`) conditions we perform two types of pre-training: i) `OSL` – pre-training on the portion of OpenSubtitles corpus ii) `OSNL` – same as `OSL`, but with all the laughs removed. We fine-tune and test our models on the corpora containing laughs (`L`).

We observe that dialogue pre-training improves performance of the models. Fine-tuned models also perform better than the frozen ones because the latter provide less opportunities for the encoder to be trained for the specific task.

Including laughter in pre-training data improves F1 scores in most cases, except for the SWDA in the fine-tuned condition. The difference is especially pronounced for AMI-DA corpus in the fine-tuned condition (4.97 p.p. difference in F1). The question of relevance of movies subtitle data for either SWDA or AMI-DA can be a subject for further study, including the types of laughs in the corpora. It might be the case that nature of AMI-DA is congruent with those of movie subtitles, since participants in AMI-DA basically are role-playing being in a focus group rather than being involved in a natural dialogue. People might produce laughs in places only where they intuitively expected by them to be produced (i.e. humour related), just as in scripted movie dialogues.

	SWDA		AMI-DA	
	F1	acc.	F1	acc.
`BERT-L-FT`	36.75	76.60	43.37	64.87
`BERT-L+OSL-FT`	41.42	76.95	48.65	68.07
`BERT-L+OSNL-FT`	43.71	77.09	43.68	64.80
`BERT-L+OSL-FZ`	9.60	57.67	17.03	51.03
`BERT-L+OSNL-FZ`	7.69	55.29	16.99	51.46
`BERT-L-RI`	32.18	73.80	34.88	60.89
Majority class	0.78	33.56	1.88	28.27
SotA	-	83.1[4]	-	-

Table 3: Comparison of macro-F1 and accuracy with further dialogue pre-training.

4.5 Experiment 3: Laughter as a non-verbal dialogue act

In this experiment, following the observations regarding the misleading character of *Non-verbal* dialogue acts, we looked at the predictions that the model would give this class of dialogue acts if it wasn't aware of the *Non-verbal* class. To do so, we mask the outputs of the model where the desired class was *Non-verbal* and do not backpropagate these results. We used the `BERT-L-FT` for this experiment. After training we tested the resulting model on the test set containing 659 non-verbal dialogue acts, 413 of which contain laughter.

For 314 (76%) of such dialogue acts the model has predicted the *Acknowledge (Backchannel)* class and for 46 (11%) – continuations of the previous DA by the same speaker. The rest were classified as either something uninformative (the *Abandoned or Turn-Exit or Uninterpretable* class) or, from manual observation, clearly unrelated.

Acknowledge (Backchannel) can cover some uses of laughter, for instance, to show to the interlocutor acknowledgement of their contribution, implying the appreciation of an incongruity and inviting continuation, functioning simultaneously as a continuer and assessment feedback (Schegloff, 1982), as in example (7).

(7) (We mark continuations of the previous DA by the same speaker with a plus, and indicate misclassified dialogue acts with a star. Laughs shown in bold constitute *Non-verbal* dialogue acts)

[4] Kozareva and Ravi (2019)

B:	Everyone on the boat was catching snapper, snappers except guess who.	*Statement (n/o)*
A:	<laughter> It had to be you.	*Summ./reform.*
B:	<laughter> I ca-, I, -	*Uninterpretable*
A:	Couldn't catch one to save your life. Huh.	*Backchannel**
B:	That's right,	*Agree/Accept*
B:	I would go from one side of the boat to the other,	*Statement (n/o)*
B:	and, uh,	+
A:	**<laughter>.**	*Backchannel*
B:	the, uh, the party boat captain could not understand, you know,	+
B:	he even, even he started baiting my hook <laughter>,	*Statement (n/o)*
A:	**<laughter>.**	*Backchannel*
B:	and holding, holding the, uh, the fishing rod.	+
A:	How funny,	*Appreciation*

Nevertheless, these two cases clearly cannot account for all the examples discussed in the literature (e.g. standalone uses of laughter as signal of disbelief or negative response to a polar question Ginzburg et al., 2020) and above in Sec. 3.2. Future models will therefore require a manual assignment of meaningful dialogue acts to standalone laughs.

5 Discussion

The implications of the results obtained are twofold: showing that laughter can help a computational model to attribute meaning to an utterance and help with pragmatic disambiguation, and consequently stressing once again the need for integrating laughter (and other non-verbal social signals) in any framework aimed to model meaning in interaction (Ginzburg et al., 2020; Maraev et al., 2021).

Our results provide further evidence (e.g. Torres et al. (1997); Mazzocconi et al. (2021)) for the fact that non-verbal behaviours are tightly related to the dialogue information structure, propositional content and dialogue act performed by utterances. Laughter, along with other non-verbal social signals, can constitute a dialogue act in itself conveying meaning and affecting the unfolding dialogue (Bavelas and Chovil, 2000; Ginzburg et al., 2020).

In this work we have shown that laughter is a valuable cue for DAR task. We believe that in our conversations laughter is informative about interlocutors' emotional and cognitive appraisals of events and communicative intents. Therefore, it should not come as a surprise that laughter acts as a cue in a computational model.

On the question of laughter impact on the dialogue act recognition (DAR) task, this study found that laughter is more helpful in SWDA corpus than in AMI-DA. Due to the nature of interactions over the phone, SWDA dialogue participants can not rely on visual signals, such as gestures and facial expressions. Our results support the hypothesis that in SWDA, vocalizations such as laughter are more pronounced and therefore more helpful in disambiguating dialogue acts. This may also explain why our best models perform better on SWDA: more of the information that interlocutors and dialogue act annotators rely on is present in SWDA transcripts, whereas AMI-DA annotators receive clear instructions to pay attention to the videos (Gui, 2005). This finding is consistent with that of Bavelas et al. (2008) who demonstrate that in face-to-face dialogue, visual components, such as gestures, can convey information that is independent from what is conveyed by speech.

Laughter can be used to mark the presence of an incongruity between what is said and what is intended, coined as *pragmatic incongruity* by Mazzocconi et al. (2020). In those cases laughter is especially valuable for disambiguating between literal and non-literal meaning, as we have shown for rhetorical questions, a task which is still a struggle for most NLP models and dialogue systems.

There is abundant room for further study of how laughter can help to disambiguate communicative intent. Stolcke et al. (2000) showed that the specific prosodic manifestations of an utterance can be used to improve DAR. With respect to laughter, the form (duration, arousal, overlap with speech) can be informative about its function and position w.r.t. the laughable (Mazzocconi, 2019). Incorporating such information is crucial if models pre-trained on large-scale text corpora are to be adapted for use in dialogue applications.

Acknowledgments

Maraev, Noble and Howes were supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP). Mazzocconi was supported by the "Investissements d'Avenir" French government program managed by the French National Research Agency (reference: ANR-16-CONV-0002) and from the Excellence Initiative of Aix-Marseille University—"A*MIDEX" through the Institute of Language, Communication and the Brain.

References

2005. Guidelines for Dialogue Act and Addressee Annotation Version 1.0.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.

Janet Bavelas, Jennifer Gerwing, Chantelle Sutton, and Danielle Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2):495–520.

Janet Beavin Bavelas and Nicole Chovil. 2000. Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and social Psychology*, 19(2):163–194.

Gregory A Bryant. 2016. How do laughter and language interact? In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Mark G Core and James F Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Boston, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*, pages 137–146.

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

Phillip Glenn. 2003. *Laughter in Interaction*. Cambridge University Press, Cambridge, UK.

Gail Jefferson. 1984. On the organization of laughter in talk about troubles. In *Structures of Social Action: Studies in Conversation Analysis*, pages 346–369.

D Jurafsky, E Shriberg, and D Biasca. 1997a. Switchboard dialog act corpus. *International Computer Science Inst. Berkeley CA, Tech. Rep*.

Daniel Jurafsky, Liz Shriberg, and Debra Biasca. 1997b. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *arXiv:1408.5882 [cs]*.

Zornitsa Kozareva and Sujith Ravi. 2019. ProSeqo: Projection Sequence Networks for On-Device Text Classification. 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 3894–3903, Hong Kong, China. Association for Computational Linguistics.

Pierre Lison and Jorg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, page 7.

Vladislav Maraev, Jean-Philippe Bernardy, and Christine Howes. 2021. Non-humorous use of laughter in spoken dialogue systems. In *Linguistic and Cognitive Approaches to Dialog Agents (LaCATODA 2021)*, pages 33–44.

Chiara Mazzocconi. 2019. *Laughter in interaction: semantics, pragmatics and child development*. Ph.D. thesis, Université de Paris.

Chiara Mazzocconi, Vladislav Maraev, and Jonathan Ginzburg. 2018. Laughter repair. *Proceedings of SemDial*, pages 16–25.

Chiara Mazzocconi, Vladislav Maraev, Vidya Somashekarappa, and Christine Howes. 2021. Looking at the pragmatics of laughter. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

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

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining Methods for Dialog Context Representation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Bill Noble and Vladislav Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 166–172, Groningen, The Netherlands (online). Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Cécile Petitjean and Esther González-Martínez. 2015. Laughing and smiling to manage trouble in french-language classroom interaction. *Classroom Discourse*, 6(2):89–106.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. Deep Dialog Act Recognition using Multiple Token, Segment, and Context Information Representations. *arXiv:1807.08587 [cs]*.

Tanya Romaniuk. 2009. The 'clinton cackle': Hillary rodham clinton's laughter in news interviews. *Crossroads of Language, Interaction, and Culture*, 7:17–49.

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.

Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. " yeah right": Sarcasm recognition for spoken dialogue systems. In *Ninth International Conference on Spoken Language Processing*.

Ye Tian, Chiara Mazzocconi, and Jonathan Ginzburg. 2016. When do we laugh? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 360–369.

Obed Torres, Justine Cassell, and Scott Prevost. 1997. Modeling gaze behavior as a function of discourse structure. In *First International Workshop on Human-Computer Conversation*. Citeseer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

A Supplementary materials

A.1 Collocations of laughs and dialogue acts

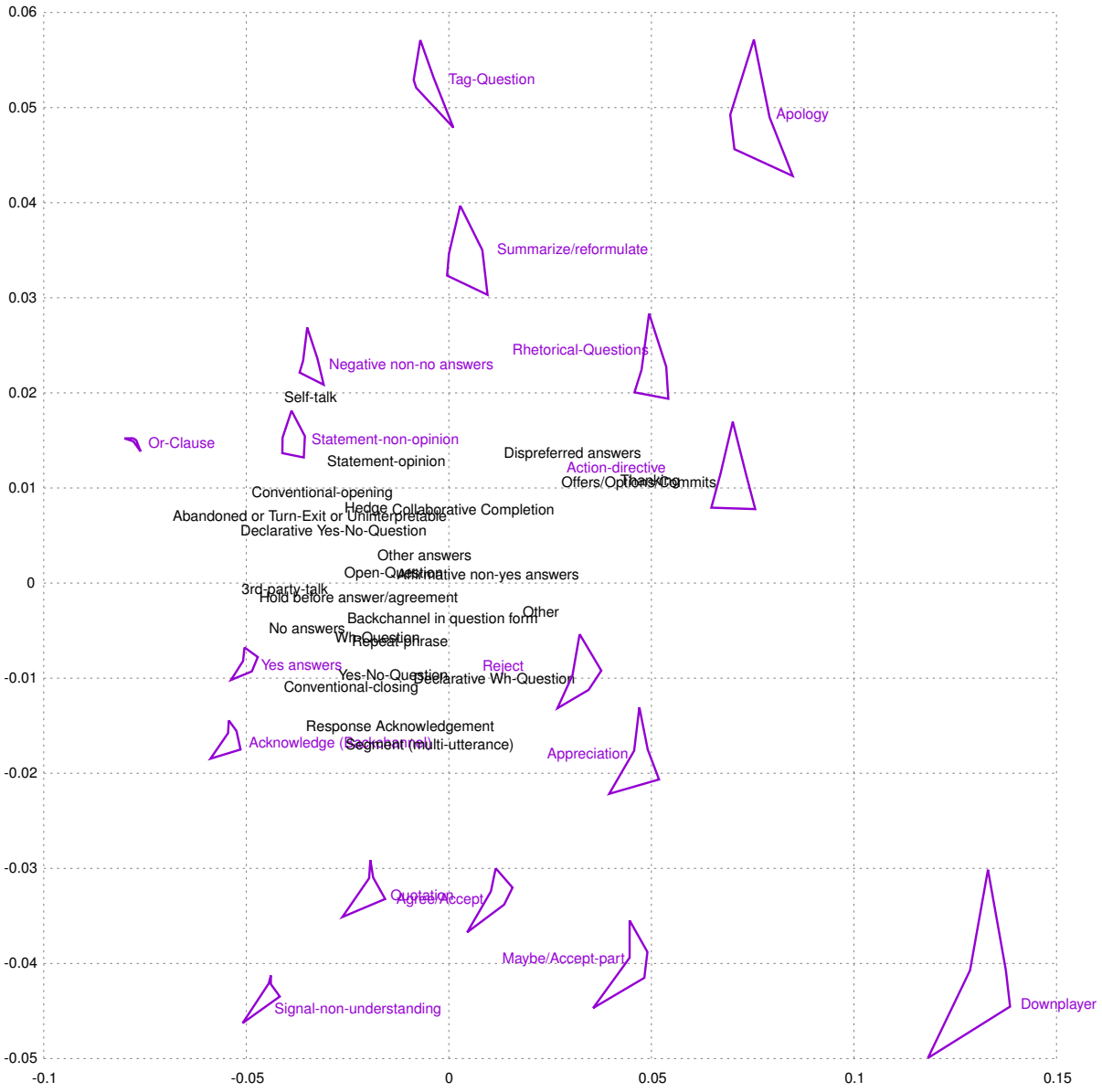


Figure 6: Singular value decomposition of pentagonal representations of dialogue acts. For a selection of dialogue acts (in purple) we depict their pentagon representations.

A.2 Model performance in DAR task

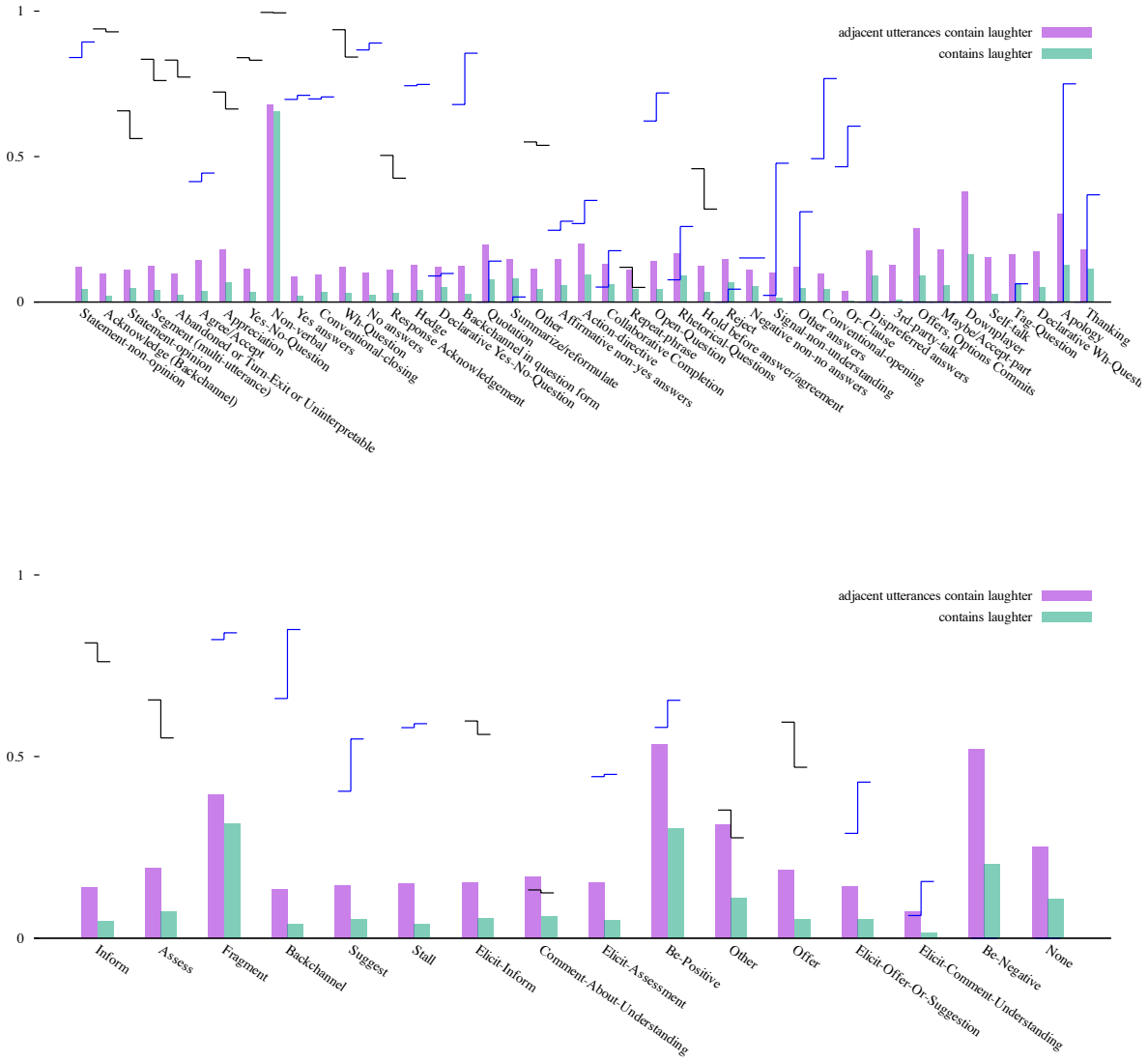


Figure 7: Change in accuracy for each dialogue act (`BERT-NL` vs `BERT-L`). Positive changes when adding laughter (`BERT-L`) are shown in blue. Vertical bars indicate how often dialogue act is associated with laughter. Top chart: SWDA, Bottom chart: AMI-DA.

What do you mean by *negotiation*? Annotating social media discussions about word meaning

Bill Noble and **Kate Viloria** and **Staffan Larsson** and **Asad Sayeed**
Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg
bill.noble@gu.se
kateviloria@outlook.com
staffan.larsson@ling.gu.se
asad.sayeed@gu.se

Abstract

We present a formalisation and annotation protocol for *word meaning negotiation* (WMN), a conversational routine in which speakers explicitly discuss the meaning of a word or phrase. WMN is formalised as an interaction game with a shared game board and rules for subsequent contributions, as well as a semantic update function based on the state of the game board. We develop an annotation schema based on this formalisation and present the results of annotating 150 Twitter conversations as WMNs.

1 Introduction

Meaningful dialogue requires some degree of alignment between participants' lexico-semantic resources. When misalignments are discovered, participants may choose to explicitly engage with the discrepancy in a metalinguistic discussion where the meaning of a misaligned word or phrase is at issue. These discussions—termed *word meaning negotiations* (WMN)—exhibit a certain structure, which we attempt to characterise and put to use by annotating WMNs collected from Twitter.

The opportunity for a WMN arises whenever a dialogue participant finds that they disagree with—or do not understand—what another speaker meant by a certain *trigger word* or phrase. They may ignore the discrepancy or silently deal with on their own (Larsson, 2010), or they may *indicate* it to their interlocutor (perhaps in the form of a clarification request). If the interlocutor responds to the indicator, a WMN has been initiated. As the WMN progresses, participants may propose, accept, reject, or raise the question of particular semantic relations between the word that triggered the WMN and other entities, which we refer to as *anchors*.

We start by discussing previous work on WMN that underpins this contribution (Section 2). Then, we develop the formal model of WMN, including a semantic update rule that can be integrated in a game board model of dialogue (Section 3). After that, we introduce an annotation schema, based on our WMN model, and present the results of an annotation study using that schema (Section 4). Finally, we discuss insights into the phenomenon of WMN resulting from the annotation study and suggest avenues for future work (Section 5).

2 Background and Related Work

There is surprisingly little work on word meaning negotiation as such. WMNs, in the form of corrective feedback, have been studied as an aspect of first language acquisition (Clark, 2007). There has also been work that teaches artificial agents the meaning of novel terms based on defintions and grounded perceptual examples (Mohan et al., 2012; Krause et al., 2014). WMNs have also been studied in conversations between non-native language learners (Varonis and Gass, 1985; Long, 1996). Myrendal (2015, 2019) has taken a more in-depth look at WMNs between adult speakers, focusing on conversations in Swedish online discussion forums.

The model and annotation scheme we develop in this work builds on the structural model of Varonis and Gass (1985) and the classificatory schemas of Myrendal (2015, 2019). The semantic update fuction we define in Section 3.4 extends the dialogue acts proposed by Larsson and Myrendal (2017). We discuss this foundation in more depth below.

TIR model In the Trigger-Indicator-Response model, when an interlocutor recognizes a non-understanding and chooses to address it overtly, the discourse enters a "subroutine" in which participants attempt to repair the non-understanding and align their semantic common ground. These subroutines are embedded in the regular linear flow of dialogue in such a way that the current line of con-

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

versation is suspended Furthermore, WMNs may be nested if, in the course of resolving one non-understanding, another non-understanding occurs and is indicated by one of the participants.

A WMN has three key elements:

Trigger – an utterance by a speaker, S_1, that contains a lexical item resulting in non-understanding by another participant, S_2.

Indicator – an utterance in which S_2 explicitly indicates their non-understanding of the trigger.

Response – an utterance in which S_1 overtly acknowledges the non-understanding.

A trigger can occur at any point in a dialogue (e.g., in a question *or* in a response). The non-understanding is only made part of the common ground once it has been indicated by S_2—thus, the trigger can only be identified retrospectively, with respect to its indicator. Likewise, the response refers back to the indicator: it may attempt to rectify the non-understanding, or merely acknowledge that a discrepancy was indicated.

Although the T-I-R model was developed for WMNs in a language learning context, Myrendal (2015) found it to be a good model for the initiation of WMNs in discussion forums as well.

Non-understanding vs. disagreement Myrendal (2015) categorises WMNs as those resulting from *misunderstanding* (NON), when one dialogue participant doesn't understand the meaning of a word uttered by another participant, in the context in which it was used, or *disagreement* (DIN), when a participant disagrees with how someone else used a word,(Myrendal, 2015). NONs are generally initiated with a *metalinguistic clarification request*, whereas DINs are initiated with a *metalinguistic objection*.

WMN dialogue acts Myrendal (2019) inventories types of WMN contributions, including *generic* and *specific explicifications*[1] (which we refer to as *partial definitions*), *exemplification*, *contrasting*, *metalinguistic objections* (which can be used in an ongoing WMN, as well as to initiate one), and *endorsement* (of a using a particular word in a given context).

Larsson and Myrendal (2017) propose dialogue acts based on these contribution types, and propose semantic update functions for exemplification partial definition and contrasting, that apply to the meaning of the trigger word, in the event that the dialogue act is grounded. In this paper, we expand on that work by using the act-level update functions to define an update that takes the entire WMN into account.

[1] (see also Ludlow, 2014).

3 Formal model

The model presented in this section has a dual purpose. First, it affords the precise formulation of hypotheses about WMNs (in general or in a particular domain) that can be tested in terms of the model. Second, the model itself implies a certain structure to the phenomenon of WMNs which may, to a greater or lesser degree, capture what is observed. As is often the case, these two roles are not entirely separable: What is expressible in the model affects the hypotheses that can be tested; How well the model aligns empirically with the phenomenon it seeks to describe affects the reliability of the conclusions one can draw.

In addition to the descriptive goal, we want the model to support a semantic *update function* that computes the change in shared lexical resources resulting from a WMN (section 3.4). The rule we define builds on the work of (Larsson and Myrendal, 2017), taking their dialogue act-specific rules and extending them to operate over a whole WMN.

Our model of word meaning negotiation depends on the notion of semantic *anchors* and speaker commitments to semantic *relations* between those anchors. This is motivated by the intuition that when speakers discuss the meaning of a word, they do so by triangulating it in reference to other points (or regions) of semantic space. In a successful WMN, the meaning of the word in question is "anchored" by the participants as a result of joint commitment to relations between the word and reference points (i.e., *anchors*) that *are* grounded.

When the project of aligning on meaning has started, it is not uncommon to discover that further discrepancies exist; that is, it can be that some of the anchors introduced to negotiate the meaning of the trigger word are themselves lacking semantic common ground (as in Varonis and Gass, 1985). This shouldn't be surprising: First of all, once a WMN has begun, discrepancies that might have gone unnoticed or un-remarked-upon are suddenly difficult to ignore. Furthermore, new anchors are introduced precisely *because* one of the participants thinks they have an elucidating relation to

the trigger. Where one semantic misalignment exists, misalignment on related terms may be lying in wait. What makes something eligible as an anchor is not that its *meaning* is common ground and fully specified, but that it can be grounded as a shared *discourse referent*, available for participants invoke anaphorically (or by name or description) and put in relation to other anchors as well as to the trigger.

We represent a word meaning negotiation, between a set of speakers S taking place over N turns, as sequence of tuples:

$$\mathsf{WMN} = \langle s_i, A_i, R_i \rangle_{i \leq N} \tag{1}$$

where s_i is the speaker at turn i, A_i is the set of anchors introduced in that turn (we let $t \in A_0$ be the trigger), and R_i is the set of relations between anchors that s_i publicly commits (Asher and Lascarides, 2008) to during that turn.

3.1 Anchors

Once introduced, anchors are available for the remainder of the WMN, accessible by co-referring expressions, including anaphora. Thus, the set of common ground anchors at turn i is defined as the union of anchors introduced so far:

$$\boldsymbol{A}_i = \bigcup_{j \leq i} A_i \tag{2}$$

We let $[\![a]\!]$ denote the meaning of a, given the context of the dialogue and the semantic common ground of the speakers, without yet considering any updates resulting from the WMN.[2]

3.2 Semantic relations

Word meaning negotiation depends on a commonly understood set of possible semantic *relation types* between anchors, $\mathcal{R}$. In the remainder of the formalisation and in the annotation study (Section 4), we assume two semantic relations, *example* and *partial definition*:

$$\mathcal{R} = \{\mathsf{Exa}, \mathsf{Def}\} \tag{3}$$

We also make use of a set of *polarities*:

$$\mathcal{O} = \{+, -, ?\} \tag{4}$$

Polarity correspond to an attitude (or commitment) that speakers may express towards a given relation between two anchors. This set of polarities indicate whether a relation holds ($+$) or its converse holds ($-$), or if the matter is in question ($?$).

In the model, $R_i \subseteq \mathcal{R} \times \mathcal{O} \times \mathbf{A}_i \times \mathbf{A}_i$ is a set of semantic relations. We will write $\mathsf{R}^o(a, b)$ for $\langle \mathsf{R}, v, a, b \rangle$. For example, $\mathsf{Def}^+(a, b) \in R_i$ means that speaker s_i has publicly committed to a as a (positive) partial definition of b.

Given WMN, we can compute a speaker's current commitments. For a pair of anchors (a, b) and relation R, we consider the speaker to be committed to the most recent polarity that has been part of their public commitments. Formally, this is defined as follows:

$$\boldsymbol{R}_{s,0} = \begin{cases} R_0 & \text{if } s = s_0 \\ \varnothing & \text{otherwise} \end{cases} \tag{5}$$

and

$$\boldsymbol{R}_{s,i+1} = \begin{cases} \boldsymbol{R}'_{s,i} \cup R_{s,i+1} & \text{if } s = s_i \\ \boldsymbol{R}_{s,i} & \text{otherwise} \end{cases} \tag{6}$$

where

$$\boldsymbol{R}'_{s,i} = \{\mathsf{R}^o(a, b) \in \boldsymbol{R}_{s,i} \mid \neg \exists o'.\, \mathsf{R}^{o'}(a, b) \in R_s\} \tag{7}$$

Finally, we define the common ground relations at turn i as those relations to which all speakers have publicly committed:

$$\boldsymbol{R}_i = \bigcap_{s \in S} \boldsymbol{R}_{s,i} \tag{8}$$

3.3 Interaction rules

Now that we have a structure for representing the state of a WMN at each turn and a way to compute what is common ground based on the history of those states, we characterise the rules of the WMN as an interaction game.

Formally, there are very few conditions on what A_i and R_i can include. Any number of anchors can be introduced in a turn, although practically the number is usually quite small (see Section 4.4). The main restriction on R_i is that it must not result in a cycle in s_i's public commitments; that is, $\{(a, b) \mid \mathsf{R}^o(a, b) \in \boldsymbol{R}_{i,s_i}\}$ must not contain a cycle. This means that $\boldsymbol{R}_{i,s}$, considered as a labeled directed graph, is acyclic, a condition that is necessary for the semantic update function (Section 3.4) to be well-defined. Intuitively it would be very

[2] Note that this interpretation, as with the negotiated meaning defined in Section 3.4, may be different for different speakers, since speakers can of course be wrong about what is common ground.

strange for speakers to ground such a cycle for exactly that reason—indeed we did not see any such cycles in speaker commitments (let alone grounded cycles) in our annotation study, although the annotation protocol would have allowed it. There are three ways of contributing to R_i:

Propose (or raise) a relation For any two anchors in $\boldsymbol{A}_i$, the speaker either proposes a relation between them ($o \in \{+, -\}$) or poses the question of their relation without asserting anything one way or the other ($o \in \{?\}$).

Ground a relation The speaker makes some indication of their stance (or negative grounding) regarding a relation that another speaker has just committed to. For some $\mathsf{R}^o(a, b) \in R_{i-1}$, $\mathsf{R}^{o'}(a, b) \in R_i$, where $o, o' \neq ?$. If $o = o'$, then it is *positive grounding*, otherwise it is *negative grounding*.

Positive grounding can be accomplished more or less implicitly, though what counts as grounding may depend on the WMN type (NON or DIN), as well as other factors such as the medium of the dialogue and social context.

Answer a question Finally, for $\mathsf{R}^?(a, b) \in R_{i-1}$, s_i can add $\mathsf{R}^o(a, b)$ to R_i for any $o \neq ?$ by answering the question posed by s_i. Note that *grounding a relation* and *answering a question* don't formally add to the possible elements of R_i beyond *posing a relation*, but we characterise them separately because they usually take the form of grounding statements or polar answers which don't include explicit co-reference to an anchor. For that reason, we also annotate them differently (Section 4.2).

3.4 Semantic update

Our goal is to define a semantic update function that takes WMN as input. We define update functions that apply to the meaning of an anchor, based on a relation with another anchor, if that relation is grounded. Then, we recursively define the update for a whole WMN based on those functions in a straightforward way:

For $a \in \boldsymbol{A}_N$, let

$$\{\mathsf{R}_1^{o_1}(b_1, a), ..., \mathsf{R}_n^{o_n}(b_n, a)\} \subseteq \boldsymbol{R}_N$$

be the common round relations anchoring a at turn N. Then the semantic update given by WMN for a is defined as:

$$\Delta(a) = [I(\mathsf{R}_1, o_1, \Delta(b_1)) \circ ... \circ I(\mathsf{R}_n, o_n, \Delta(b_n))](\llbracket a \rrbracket) \quad (9)$$

Here, I is the interpretation of R (we assume that for a semantic relation to be common ground implies the existence of an update function):[3]

$$I = \begin{cases} \lambda x.\epsilon^o(b, x) \text{ if } \mathsf{R} = \mathsf{Exa} \\ \lambda x.\delta^o(b, x) \text{ if } \mathsf{R} = \mathsf{Def} \end{cases} \quad (10)$$

In essence, Δ, as defined in (9) applies the update implied by the semantic relations recursively on $\mathbf{R}_N$ in a straightforward way: the updated meaning of an anchor is computed by sequentially applying each its grounded relations to other anchors, with the caveat that each of those anchors should first have *their* meaning updated, if they were also negotiated as part of the WMN.

4 Annotation study

4.1 Data

We collected exchanges on Twitter that, based on search heuristics, were likely to involve WMN. In particular, we used the Twitter filtered stream API to find tweets that were in reply to another tweet and that used the indicator phrase *what do you mean by*.[4] This heuristic method is based on that of Myrendal (2015), who used similar phrases in Swedish to build a corpus of WMNs from online discussion forums. The search resulted in a total of 1783 candidate indicator tweets, collected over a 24-hour period (May 5–6, 2021).

After 48 hours (to wait for replies), we used the Twitter search API to collect the rest of the thread, retrieving tweets both upwards and downwards in the reply chain. Since the reply structure on Twitter is a tree (each tweet can be *in reply to* at most one other tweet, but can *have* multiple replies), retrieving the upwards context is easy—we just followed the replies up to the root of the thread. For the downward search (replies to the indicator), we initially look for a reply from the author that the indicator was a reply to, alternating back and forth between these two users for further replies and taking the first reply in case there were multiple.[5] This resulted in 671 threads with at least one reply after the candidate indicator (38% of threads), of which we randomly sampled 150 for annotation.

[3] We let ϵ^+, ϵ^-, δ^+, and δ^- be as defined in Larsson and Myrendal (2017).

[4] We used a regular expression to allow for some variation in the exact wording (see supplementary materials for details).

[5] This is a somewhat brittle heuristic that could improved upon. For example, it breaks if a user makes a "double reply" or if the conversation is between more than two users.

4.2 Annotation protocol

The annotation protocol, which was developed over a series of pilot studies, aims to be comprehensible for annotators with no linguistic background (see the annotation guide in the supplementary materials). In the pilot studies, small sets of data collected from Twitter were manually annotated using initial drafts of the annotation schema by two annotators (both with a linguistic background). Error analysis sessions were conducted in order to discuss and clarify unclear definitions and inconsistent judgments between annotators. The schema was then refined based on these discussions.

Two additional annotators were added to annotate more data, which we report on in Section 4.4. All four annotators are linguists familiar with WMNs. As in the pilot studies, an error analysis was conducted, which we discuss in Section 4.5).

Annotators were shown text of the tweets, one thread at a time, in the BRAT annotation tool (Stenetorp et al., 2012). Tweets were separated by a header that included the time of the tweet and the username of the tweet author. We displayed a maximum of 10 context tweets on either side of the candidate indicator.

Annotators were instructed to read the Twitter threads and select and classify text spans as different components of a WMN——as well as to determine whether or not an exchange as a whole was in fact a WMN. The four main points of interest, meant to be evaluated in order, during annotation were the WMN Type, Trigger spans, Anchors (Examples and Definitions), and instances of Grounding. While it was recommended that annotators examine these four points in order, we noted that it is completely acceptable and sometimes necessary to go back and forth to gain a better understanding of the thread.

WMN Type The search phrase (e.g., *what do you mean by*) was automatically pre-labeled as an Indicator to help the annotator find the intended focus of the example. Annotators were instructed to tag the Indicator span with the WMN Type of the dialogue as a whole. WMN Type consists of two decision points: First, the annotator must decide whether the thread is a WMN or not. If it *is* a WMN, it must then be classified as a non-understanding (NON) or disagreement (DIS).

Trigger The second task is to identify the word or phrase in question as the Trigger. Annotators must also label every other instance of the Trigger in the discussion, including anaphoric references. It is not necessary to link Triggers together with co-reference relations since it is implied.

Anchors The next step is to find the Trigger's Anchors and to distinguish between an Anchor's two types, Examples and Definitions. Relations are annotated with a link between the anchor and the Trigger o another Anchor, and the are marked with the polarity of the relation. An Anchor can also appear multiple times within a WMN, including anaphoric reference. In this case, these anchors are linked together using the co-reference relation. Annotators are instructed to try and leave negations out of the anchor and instead annotate the relationship as having negative polarity. When linking anchors, it is important which instance of the Anchor the link originates from, since this indicates which speaker is making the commitment and when. It is recommended that annotators use their best guess when identifying whether or not a URL (which could be an image or external link) is an Anchor and if so, its type based on the textual context.

Grounding Spans of text that explicitly state the speaker does (or does not) understand or agree with the previously offered example or definition must be annotated as Grounding. This span must be linked to the Anchor it refers to. The polarity link of a Grounding statement can be either positive or negative and between an Anchor and a Trigger or between two Anchors. In a non-understanding WMN, a grounding statement with a positive link indicates that the speaker understands the proposed relationship between the Anchor and the Trigger (or another Anchor). A negative link indicates that the speaker does not (or may not) understand the proposed relationship between the Anchor and the Trigger. In a disagreement WMN, a positive link indicates that the speaker agrees with or has adopted the proposed relationship between the Anchor and the Trigger. With a negative link, the grounding statement indicates that the speaker does not agree with or has not adopted the proposed relationship between the Anchor and Trigger.

4.3 Post-processing annotations

There are some discrepancies between the annotation schema and the WMN formalisation described in Section 3, mainly due to the fact the formalisation is comprised of abstract semantic units, while

the annotation is performed directly on the surface form of the WMN.

Text spans annotated as an *Anchor* (Example or Definition) were divided into equivalence classes, based on the co-reference annotations, which constitute the set of anchors in the formalisation. Spans annotated as Trigger were assumed to co-refer and the set of Triggers also constitutes an Anchor.

Relation type (Exa or Def in the formalisation) is coded as property of anchors in the annotation schema. In the pilot studies, we found that it was easier to decide the relational role of the anchor span before determining the polarity and target anchor. It is also more visually legible to separate the relation type (indicated by the color of the anchor span) and polarity (indicated by the color of the relation arrow). In theory, it would be possible for an anchor have multiple relational roles (imagine, for example, a WMN in which *insect* is used as both a partial definition of a *locust* and as an example of an *invertebrate*), but in practice this seems to be vanishingly rare (we have never observed it).

4.4 Results

In this section we report the results of the annotation study, particularly inter-annotator agreement.

We measured annotator agreement at two levels of description: the surface-form annotation, and then on the formal WMN representation extracted from the annotations. For agreement statistics, we report the proportion of agreed-upon items (A_0), as well as Cohen's kappa (κ) and Scott's pi (π).[6] Cohen's kappa computes expected agreement (the denominator) using annotator-level priors for the label distribution, whereas Scott's pi assumes a uniform distribution across annotators. Significantly higher κ compared to π would suggest that annotators have different priors for the category labels (Artstein and Poesio, 2008), but we don't observe that to be the case in any of the agreement statistics we measured.

First, we measured agreement on the dialogue level, namely, the classification of whether or not the dialogue was a WMN and if so, what type. Agreement was above chance, but (Table 1) with a substantial amount of disagreement. We discuss potential sources of disagreement in Section 5.

We measured agreement on span type at the token level. Tokenisation was performed post-hoc—annotators selected spans from the raw character-level text—but we consider a token to be part of a span if a majority of characters in the token overlap with it. This eliminates any artificial disagreements caused by, for example, missing the final letter in a word when selecting a span. We also consider it to be more representative than character-level agreement, which would be biased by longer words.[7]

	A_0	π	κ
WMN/Not	0.71	0.40	0.40
NON/DIN	0.79	0.47	0.48

Table 1: WMN type agreement. *WMN/Not* measures agreement on whether or not the dialogue was a WMN, while *NON/DIN* (restricted dialogues both annotators agreed were WMNs) measures agreement on whether the WMN resulted from *non-understanding* or *disagreement*.

We found a moderate level of agreement on all span types except *grounding* (Table 2). Error analysis suggests that this may be primarily due to how much of a tweet the annotator considered to be a part of the grounding span. Additional guidance on this point in the annotation guide may help to raise the level of agreement.

	A_0	π	κ
Anchor	0.93	0.59	0.60
Trigger	0.98	0.63	0.63
Grounding	0.98	0.22	0.22
Overall	0.87	0.64	0.64

Table 2: Token-level span type agreement. *Anchor* (both Definition and Example are considered *Anchor* here), *Trigger*, and *Grounding* only consider the binary choice of whether or not a token is of that type. *Overall* considers all three possibilities together.

At the level of the formal WMN representation, we are interested in whether annotators agree on whether and what kind of relations between anchors participants commit to at each turn, and when they explicitly indicate grounding of those relations. Computing agreement for relations and grounding requires that we align the anchors identified by the two annotators. For this, we take the bijection that maximizes token-level overlap of the spans associated with the anchors. This anchor mapping aligned an average of 89.1% (σ=19.2%) of anchors per dialogue (that is, on average 10.9% of anchors

[6] A_0 is the numerator for both κ and π.

[7] We used the NLTK (v.3.6.2) regex-based TweetTokenizer.

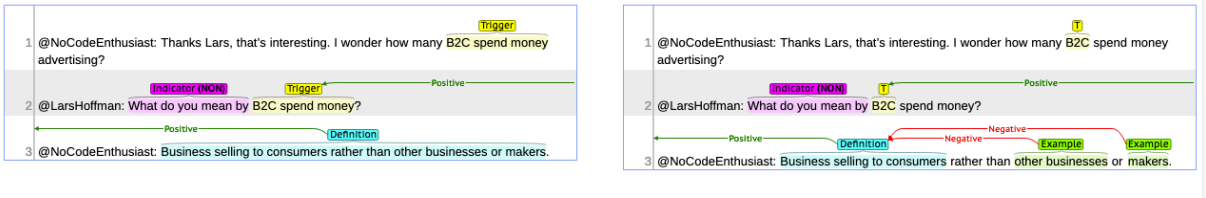


Figure 1: Annotations from two annotators, showing disagreement in the extent of the trigger phrase and anchor structure.

had no counterpart in the other annotation).

For relations, we considered each *potential* relation at each turn; that is, for turn i, we consider each pair of anchors (including the trigger), $\{(a, b) \in \boldsymbol{A}_i \times \boldsymbol{A}_i \mid a \neq b\}$ (with the caveat that $\boldsymbol{A}_i$ only includes *aligned* anchors, since there is no possibility for agreement on unaligned anchors). Annotators agree if they both created a relation (with the same relation type and polarity) from an a-span originating in turn i to an b-span (regardless of where)—or if they both created no relation at all for that pair. As with the token-level statistics, A_0 is quite high, since relations are sparse, relative to all the opportunities for a relation to be created, but the chance-adjusted scores are also reasonably high (Table 3).

For grounding, for each turn i (starting with $i = 1$), we considered each aligned anchor that both annotators agreed was mentioned in turn $i - 1$. Annotators agree if they both thought that the current speaker grounded (with the same polarity) a relation originating in that anchor—or if they both thought no such grounding occurred. Agreement is lower than for anchor relations, but still well above chance (Table 3).

	A_0	π	κ
Relation	0.93	0.69	0.69
Grounding	0.88	0.58	0.59

Table 3: Turn-level agreement on relation type and grounding polarity for possible relations and grounding.

4.5 Error analysis

After annotating the examples, we conducted some post-hoc discussions in which the annotators attempt to ascertain the reason for certain discrepancies. Based on these discussions, we make suggestions for improvements to the annotation protocol, which should aid in future efforts to annotate WMN. Further observations about the phenomenon of WMN, which came to light in these conversations, can be found in Section 5.

WMN Type The phrase *what do you mean by* is often used in a rhetorical way (i.e., not as a genuine question or clarification request), but it can be difficult to determine whether the speaker's objection to using a word to describe some situation under discussion is a disagreement about the meaning of the word (DIN) or a disagreement about the nature of the situation under discussion (not a WMN). The decision could be clarified by emphasizing the *results* of the indicator phrase: Does the meaning of the word subsequently become at-issue? When non-understanding or disagreement is indicated but no meaning negotiation results, this is typcially not considered a WMN (Varonis and Gass, 1985; Myrendal, 2015), but giving such "declined WMNs" their own category could result in better agreement.

Anchor spans Analysis revealed two kinds of discrepancy in anchor spans: (1) where the annotators disagreed on whether something was an anchor, or how much of the text referred to the anchor (reflected in token-level agreement, Table 2), and (2) where the annotators disagreed on whether something was one anchors or two (reflected mainly in the failure to find a bijection between the two annotated sets of anchors).

A particularly notable discrepancy of the first kind involves the extent of the trigger phrase, since the speaker will sometimes repeat some context around the trigger to help locate it in the previous utterance. This can raise the question of how much of what they repeated is context and how much is the trigger. One strategy for annotators could be to observe what is *actually negotiated* subsequent to the indicator, although this too can be ambiguous.

Another common discrepancy was that one annotator would annotate multiple anchors, where another would find only one (see Figure:1).

Relation types While agreement on relation type (annotated as anchor span type) was fairly good, there were a few cases where adding more relation types could improve clarity. *Contrasting* is a common pattern in WMNs where the trigger word is compared to an alternative that the speaker thinks better describes the situation under discussion (Myrendal, 2019): *x is really more of a Y than a Z*. In the annotation guide, we suggested such examples be annotated with two relations: $\mathsf{Exa}^{+}(Y, x), \mathsf{Exa}^{-}(Z, x)$, but it could also be its own ternary relation that is interpreted using δ and ϵ, as in Larsson and Myrendal (2017).

5 Discussion and conclusion

We conclude by offering some observations on the WMNs in our Twitter corpus, and discussion on the implications these observations may have for negotiated meaning more broadly.

Speaker meaning/token meaning As mentioned in Section 4.5, it was often unclear whether *what do you mean by X* was asking what the speaker understands *X* to mean in general, or what they were *using X* to mean in a particular context.[8] This is perhaps related to the phenomenon where the indicator repeats a whole sentence, but the negotiation focuses on one word or short phrase: Since questions about sentence meaning are necessarily about speaker meaning, including the sentence in the indicator may clarify that the question is about speaker meaning. Clark (1996)'s hierarchical grounding schema, makes the distinction between grounding on the level of *signal meaning* and grounding on the level *uptake* (speaker meaning or illocutionary act). When a WMN is focused on resolving a non-understanding (NON), the issue can be either with the signal meaning or with uptake, however a disagreement (DIN) about how a word is used is necessarily a disagreement about its *meaning potential* (Linell, 2009)—it doesn't make sense to disagree *that* someone meant something, only *how* they went about meaning it.

Social and cultural context Many of the WMNs in our corpus involved politically or socially controversial topics and the moves made by the participants often required some understanding of the social context in which the conversation was taking place. Consider the example in figure 2: Interpreting *convicted by the court* as providing a negative example of a *delusional conspiracy*, requires understanding the role of *conspiracy* in Indian political discourse, what *Congress* refers to (a political party) and even the political alignment implied by *Saffron* in one of the usernames.

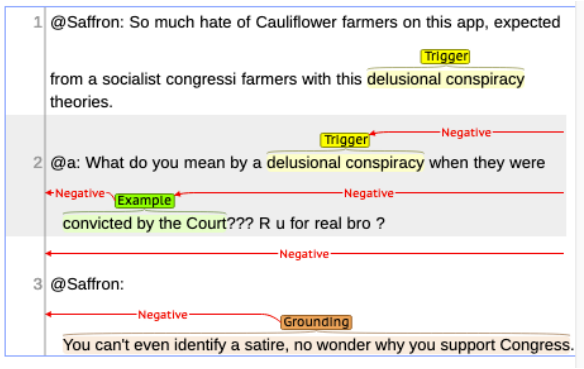


Figure 2: Post-hoc annotation provided by an annotator familiar with Indian social media political discourse. The original annotators of this example, lacking the background knowledge, had different interpretations.

Agreement and reliability As the cultural context example demonstrates, annotator disagreement doesn't *necessarily* imply that the annotation schema is incorrect or doesn't reflect the underlying phenomenon. In that case, one of the annotators lacked the context to interpret the WMN correctly, but it is possible for WMNs to be ambiguous (open to multiple possible interpretations), even when both have sufficient background knowledge. Reflecting these different interpretations can make this formalisation a useful tool for analysis, just as first-order logic is a useful tool for analysing certain classes of ambiguous sentences.

Taking that for granted, and considering our somewhat mediocre annotator agreement scores, what can we conclude about this formalisation and annotation schema? Is it in some sense *correct*? The only way to know is probably to continue using it (and where possible, improve upon it)—to carry out further annotation studies on conversational data from different sources, formulate and test hypotheses, and eventually attempt to train artificial agents capable of WMN.

As explicit meta-linguistic discussions, WMNs have potential as window into the processes of semantic alignment, acquisition, and change more generally. By modeling WMNs, we hope to develop conceptual frameworks that apply to the dynamics of lexical semantic resources more broadly.

[8] See also: Myrendal (2019) *general* versus *specific explicifications*.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher and Alex Lascarides. 2008. Commitments, Beliefs and Intentions in Dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Eve V. Clark. 2007. Young Children's Uptake of New Words in Conversation. *Language in Society*, 36(2):157–182.

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

Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz. 2014. Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue Acts and Updates for Semantic Coordination. In *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 52–59. ISCA.

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

Michael H. Long. 1996. The Role of the Linguistic Environment in Second Language Acquisition. In William C. Ritchie and Tej K. Bhatia, editors, *Handbook of Second Language Acquisition*, pages 413–468. Academic Press, San Diego.

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*, first edition edition. Oxford University Press, Oxford, United Kingdom.

Shiwali Mohan, Aaron Mininger, James Kirk, and John E. Laird. 2012. Learning Grounded Language through Situated Interactive Instruction. In *2012 AAAI Fall Symposium Series*.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. PhD Thesis, University of Gothenburg, University of Gothenburg.

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

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

E. M. Varonis and S. Gass. 1985. Non-native/Non-native Conversations: A Model for Negotiation of Meaning. *Applied Linguistics*, 6(1):71–90.

Challenging evidential non-challengeability

Vesela Simeonova
University of Tübingen
vesela.simeonova@uni-tuebingen.de

Abstract

Whether and how evidential markers can be challenged in discourse is theoretically consequential and yet it is not studied in depth: (i) only direct challenges are tested, not indirect; (ii) different evidential bases are not compared explicitly. This paper informs both gaps, providing a novel methodological framework for testing for challengeability based on the specifics of the evidential base. The results show that: (i) evidentials can be indirectly challenged, supporting a presuppositional account and falsifying alternative ones; (ii) different evidential bases are challenged in different ways, even within the same language; (iii) some evidential bases cannot be challenged, but this is due to the nature of the evidence they represent, not a consequence of the nature of the grammatical category of evidentiality, as assumed before.

1 Evidential (non-)challengeability

1.1 Direct challenges

The question whether evidentials are challengeable in discourse has been of interest since the first formal semantic work on evidentiality, (Izvorski, 1997), as it can inform whether they are interpreted as at-issue (AI) or not-at-issue (NAI) information.

Izvorski (1997) argues that evidentials in Bulgarian are not directly challengeable in conversation, as in (1): the proposition that Ivan passed the test can be felicitously contested by an interlocutor, but the evidential grounds for uttering it cannot be.[1]

(1) A: Ivan izkara-**l** izpit-a.
Ivan pass-REP exam-DEF
'*Apparently*, Ivan passed the exam. '

B: This isn't true.
='It's not true that Ivan passed the exam.'
≠'It's not true that *it is said* that Ivan passed the exam.'
[Bulgarian], (Izvorski, 1997): (16)

In other languages (to the exception of Basque, see Korta and Zubeldia, 2014), this test has yielded the same results, which have been taken to support various types of NAI accounts: presuppositional (Izvorski, 1997; McCready and Asher, 2006; Matthewson et al., 2007): sincerity condition (Faller, 2002), NAI-assertion (Murray, 2010), evidentials as tenses (Smirnova, 2013) or as conventional implicatures (Koev, 2016); and also an AI account of evidentials as subjective content (Korotkova, 2016a,b).

1.2 Indirect challenges

The value of direct challengeability as a diagnostic for AI/NAI status has been disputed (Matthewson et al., 2007; Murray, 2010; Korotkova, 2016b), since the theories cited above do not make different predictions about it (they all predict patterns as in (1)); furthermore, even some AI content is not challengeable (Korotkova, 2016a, 2020a).[2]

However, when evidentials are attested to be not **directly** challengeable, an implicature is left behind that they may be indirectly so. This is also one of the theoretical predictions of presuppositional accounts, while it is not the case for others, such as NAI assertion and AI subjectivity. Therefore, while it may be true that direct challenge impossibility is not informative, indirect challenges are.

And yet whether and how evidential utterances can be indirectly challenged has not yet been explored. The primary empirical task of this paper, carried out in §2, is to fill that gap and address

[1] Evidential markers are formatted in **bold** in examples, and their approximate translation — in *italics*. The English translation is not intended to represent their (not-)at-issueness status. Glosses used: REP=reportative evidential; DIR=direct evidential; DEF=definite; REFL=reflexive; ADJ=adjective; SBJ=subjunctive; VOC=vocative.

[2] By the end of this paper, novel concerns are raised about the direct challengeability diagnostic as stated in (1).

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

whether (and which) evidentials can be challenged indirectly and how. The theoretical goal, explored in §3, is to compute some of the implications of the findings. The findings also provide a methodological blueprint for testing indirect denials and a new way to interpret direct ones.

1.3 Terminological conventions

The following terminology is adopted in this paper:
EVIDENTIAL ('ev') $\stackrel{\text{def}}{=}$ the evidential marker in the sentence, e.g. the morpheme *l* in (1).
EVIDENTIAL BASE ('ev base') $\stackrel{\text{def}}{=}$ the type of evidence that the evidential marker denotes, e.g. direct, reportative, or inferential evidence (Willett, 1988).
SCOPE SITUATION $\stackrel{\text{def}}{=}$ the situation which the evidence is about, e.g. the situation of Ivan passing the exam in (1).
Expanding on Smirnova (2013); Koev (2016), who propose that evidentials carry a temporal component (the evidence acquisition time), I propose that each ev base corresponds to an **event** of (respectively) witnessing, hearing a report about, or inferring about the scope situation. I refer to this event as the EVIDENTIAL EVENT ('ev event').

2 How to challenge evidentials

This section develops the methodology for indirectly challenging evidentials. The resulting empirical findings are: (i) overall, evidentials are indirectly challengeable; (ii) different ev bases have different challengeability profiles (even in the same language); (iii) thirdhand reportatives are not challengeable even indirectly, but this is due not to the nature of evidentiality as a grammatical category, rather to the nature of the ev event that that particular type of evidential represents (rumors).

The novel data provided here are from Bulgarian (other languages are identified individually), but the diagnostics are not language specific — they are ev base specific.

2.1 Challenging a direct evidential base

For an interlocutor to challenge the ev base means that they refuse to accept that the ev event occurred. The scope situation may be true, but the interlocutor submits that it is false that the speaker has that type of evidence about it. In what circumstances could such challenge occur? — When the ev event is impossible or at least highly improbable to have occurred. This is most intuitive to demonstrate with direct evs, which denote an event of witnessing the scope situation (Willett, 1988).[3]

The example below is naturally occurring, uttered by a tween English-dominant heritage speaker of Bulgarian ('S'). His mom ('M'), a native speaker of Bulgarian, corrects him for his use of the direct evidential to describe something that happened when he was a baby, because it's impossible that he remembers his behavior (even though he was of course physically present when the scope situation occurred). As one of the (adult native) consultants commented when presented with this scenario, "one has to reach a certain age in order to be able to use that form".[4]

(2) S: Kogato **bjax** bebe, **placheh** mnogo.
when was.DIR baby cried.DIR a.lot
'When I was a baby, I used to cry a lot.'
M: Siakash pomnish kolko problemi
as.if remember.2SG how-many problems
suzdavashe
created.2SG.DIR
'As if you could remember how much trouble you gave us!'

This example also shows that it is possible to agree on the scope situation and only disagree about the evidential event.

The next example is also natural, this time from the comments section of a news article about terrorist attacks.[5] One commenter (A) disputes the news article's claim that the attacks were spontaneous and by few armed men; he uses direct evidential in his comment. Two other commenters (B, C) confront A on the basis that A couldn't have been an eyewitness of the events.

(3) A: Ataki-te **biaha** dosta dobre
attacks-DEF were.DIR very well
organizirani i v nikakav sluchai ne
organized and in no.ADJ way not
biaha ot samo 5-6 dushi.
were.DIR by just 5-6 people
'The attacks were very well organized and definitely not by just 5-6 people, *as I saw*.'

[3] Some languages have evidentials with a meaning much wider than just witnessing, such as the Best Possible Grounds (BPG) marker *mi* in Cuzco Quechua (Faller, 2002), which can be used even when one has reportative or inferential information, as long as it is the best possible kind of information one could have about the respective scope situation.

[4] This example also supports Korotkova's 2020b observation that evidentials have a *de se* component: one needs to be aware of one's own experience of the ev event in order to use an ev marker of the respective base.

[5] Source: here.

B: [A] preuvelichava malko pri.uslovie.che
A exaggerate.3SG a.bit given.that
sigurno ne e prisustval tam
probably not be.3SG been there
'[A] is exaggerating a bit, given that he probably wasn't there...'

C: Abe ti da ne si bil tam che
VOC you SBJ not be.2SG been there that
mnogo gi znaesh neshtata, vse.edno
a.lot them know things.DEF as.if
si bil s RPG-to
be.2SG been with RPG-DEF
'Were you(=A) there, that you know how it was, as if you were there with the RPG?'

The example confirms a note made by (Aikhenvald, 2004): "using a wrong evidential is one way of telling a lie" (p. 20).[6] In this case, commenters B and C are calling out the lie on A.

But a direct marker doesn't have to be a lie to be brought to the foreground in conversation. An interlocutor can pick up on the evidential base if it is surprising, such as in (4) (constructed), where B asks for confirmation of the evidential base with a rising declarative, or even take the marker itself as evidence that the speaker was present at the scope situation, as in (5) (natural):

(4) B (assuming that A wasn't at the party):
Chul li si neshto za partito?
hear.PP Q REFL something about party.DEF
Have you heard anything about the party?

A: Mina i Zlati se tseluva-**ha**.
Mina and Zlati REFL kiss.3PL-PST.DIR
'Mina and Zlati were kissing, *I saw*.'

B: Chakaj, chakaj, ti si BIL na partito?
wait wait you are be.PP at party.DEF
'Wait, wait, you WERE at the party?'

(5) A: Nejkov ne **beshe** tam.
Nejkov not was.DIR there
'Mr. Nejkov wasn't there, *I saw*.'

B: Znachi ti si bil tam.
so you be.3SG be.PP there
'So you were there.' source: here[7]

To sum up, this part demonstrated how the direct evidential base can be challenged — when it is impossible or implausible that the speaker was present at the scope situation (and aware about it, see fn. [4]) — and more broadly, how the base can be promoted to a question under discussion.

[6] An example of a sentence with a wrong ev marker is given in Aikhenvald (2004):98, (3.45), but not in conversation.

[7] This example is from 1922, but the judgement is equivalent today. See Kutsarov (1994) for an overview of the history of linguistic works describing evidentials in Bulgarian.

2.2 Different evidentials = different challengeability profiles

The received view is that various evidential bases have a uniform behavior with regards to conversational challenging: they all resist direct denials. But do they all behave uniformly with respect to indirect denials? This part demonstrates that different evidential bases — and even subtypes of bases, in the case of reportatives — have different challengeability profiles.

The previous section showed how to challenge the direct evidential marker; here the reportative is in the spotlight. Willett (1988) distinguishes between SECONDHAND and THIRDHAND reports encoded by evidential markers crosslinguistically:[8] in secondhand reports, the source of the report is identified, while in thirdhand, it's not known; rumors are such reports.[9]

According to Willett (1988), in some languages, there are different markers for the two types of report ev bases, in some, only one is represented, and in others, one marker is used for both readings. Bulgarian is of the latter type. This part shows that the Bulgarian reportative evidential behaves differently with regards to different types of reports: a third-hand reportative base cannot be challenged even indirectly (for reasons different from those predicted in the literature, discussed in §3) and the secondhand reportative base can be challenged under circumstances different from those relevant for the direct evidential.

2.2.1 Secondhand reports and reputation

In order to check whether and how secondhand reportative evidentials can be challenged, one needs to know whether and how what they represent — secondhand reports — can be challenged, highlighting that it is not the content of the claim that needs to be questioned, but the mere existence of a claim with such content.

Since reports are based on what people have said, they cannot be contested on the basis of some objective impossibility as with the direct ev base:

[8] Willett (1988) also considers folklore as part of reportatives, but fiction is beyond the scope of the present paper.

[9] A basic division between known and unknown sources is also used by Aikhenvald (2004), where the former (here: secondhand) is called quotative. There may be other uses of these labels (see AnderBois, 2019a: fn. [2]) and other languages with multiple reportatives manifesting other properties: for example, in Yucatec Maya, there is a reportative marker that allows both types of reports discussed here, and a quotative, which marks direct quotation (AnderBois, 2019a,b).

to do that, one has to be able to show that such a report could not have existed, but in order to show that, one has to know all the things the author ever said, which is in turn objectively impossible.

Implausibility, however, translates well in secondhand reports as the unlikelihood that that particular author would have said the respective report given their prior public commitments, i.e. their reputation.

To illustrate this, let's look at two famous people who publicize their opinion on climate change and stand by it with reliable consistency over time: Donald Trump, who denies global warming, and Greta Thunberg, the environmental activist. Their consistency allows their audience to build expectations of what they are likely to say and not say. In (6), B cannot felicitously reply with 'He can't have said that'. But if A said something implausible, as in (7), B could felicitously challenge the report.

(6) A: Trump tweeted that there is no global warming.
B: #He can't have said that!

(7) A: Trump tweeted that he will fight global warming.
B: He can't have said that!... (Are you sure you were looking at the *real* Trump's profile, or a fake profile? Was there a checkmark by the name?)

If we replace Trump with Greta Thunberg, the judgments are reversed, as in (8), (9), demonstrating that the challenge is indeed dependent on the source's reputation.

(8) A: Greta Thunberg tweeted that there is no global warming
B: She can't have said that!

(9) A: Greta Thunberg tweeted that she will fight global warming
B:#She can't have said that! (etc.)

For evidential markers that can represent secondhand reports, the findings above translate into a prediction that they can be challenged under the same conditions as the respective reports can. The next examples show that this is indeed borne out.

The first example, (10), illustrates this with an appositive that identifies the source.[10] Comparing the infelicitous challenge in (10) with the felicitous one in (11), and the felicity parallel with the respective non-evidential reports in (6) and (7), reveals that challengeability is a function not of an intrinsic property of the category of evidentiality as a whole, but of the source's reputation, just as it is with non-evidential secondhand reports.

(10) A: Spored Trump nyama-**lo** globalno
according.to Trump has.no-REP global
zatoplyane.
warming
'According to Trump, there is no global warming.'
B:#Ne, ne može da e kazal tova!
no not may SBJ is said.PP this
'No, he could not have said this!'

(11) A: Spored Trump globalnoto zatoplyane
according.to Trump global warming
bilo realen problem.
be.REP real problem
'According to Trump, global warming is a real problem.'
B: Ne, ne može da e kazal tova! [OK]
no not may SBJ is said.PP this
'No, he could not have said this!'

The next example uses the property of evidential anaphoricity, which strongly (if not exclusively, for Bulgarian at least) favors the secondhand interpretation, as first discussed by Murray (2010) for Cheyenne and confirmed for Bulgarian by Koev (2016). The generic form of anaphoric sequences is schematized in (12) after Murray (2010): (5.19): a reportative marker in the second independent sentence refers to the attitude holder introduced in the first one. Examples (13)-(14) show that such utterances are challengeable under the same conditions as the non-evidential reports in (6)-(7) and the evidential ones with an oblique source in (10)-(11).

(12) I spoke with Dale. Annie won REP.
REP=what D. said is that A. won

(13) A: Trump pak tweetva aktivno. Nyama-**lo**
Trump again tweets actively has.no-REP
globalno zatoplyane.
global warming
'Trump is actively tweeting again. [*he says*] There is no global warming. '
B: = B in (10), i.e. infelicitous

[10] The reportative evidential here has the so-called CONCORD reading, where its interpretation is 'vacuous', as Schwager (2010) describes it (hence it is missing from the translation). The term is due to Schenner (2010a,b) on German and Turkish, see also Schwager (2010) on Tagalog and German, and Bary and Maier (2021) on Ancient Greek.

(14) A: Trump pak tweetva aktivno. Globalnoto
Trump again tweets actively global.DEF
zatoplyane **bilo** realen problem.
warming is.REP real problem
'Trump is actively tweeting again. [*he says*] Global warming is a real problem. '
B: = B in (11), i.e. felicitous

To summarize, this part has demonstrated that secondhand reports can be challenged on the basis of reputation. The next part shows that the thirdhand ev base differs: it cannot be challenged.

2.2.2 Why thirdhand reports cannot be challenged

The successful challenge cases until now were based on impossibility or implausibility that the evidential event happened, or the reputation of the source. With rumors, one cannot appeal to reputation because the source is by definition unknown. One cannot appeal to impossibility or implausibility because any rumor could in principle exist: one can never rightly object with 'Nobody (ever) said that!' because for any claim there could have been someone who said it — it is objectively impossible to prove that there wasn't, regardless of how the rumor was formally encoded: lexically, as in (15), or grammatically via evidentiality, (16).

(15) A: Mina reportedly kissed Zlati.
B: #Nobody ever said that!
B′: #You didn't hear that!

(16) A: Mina tseluna-**la** Zlati.
Mina kissed-REP Zlati
'Mina *reportedly* kissed Zlati.'
B: #Nikoi ne e kazal tova!
nobody not is said.PP this
'Nobody ever said that!'
B′: #Ne si chula tova!
not be.2SG heard.PP this
'You didn't hear that!'

Therefore, it is indeed impossible to challenge a thirdhand evidential, but this is simply because it is impossible to challenge the ev event it stands for: a rumor. It is not a function of the formal properties of evidentiality as a grammatical category, but simply the nature of rumors in particular.

3 Discussion of findings

This section explores some of the theoretical and methodological implications of the empirical findings reported in this paper.

3.1 Evidentials are indirectly challengeable

The major empirical finding presented in this paper is the first evidence that evidentials are indeed indirectly challengeable, i.e. nothing about the grammatical category prevents that. The theoretical consequences include novel support for: (i) the NAI status of evidentials; (ii) a presuppositional analysis of evidentials over alternative NAI accounts.

3.1.1 NAI

This paper opened with the observation that since Izvorski (1997), the literature has focused on whether evidentials are directly challengeable, and has taken the fact that they aren't as evidence that they are NAI content.

But Korotkova (2016a, 2020a) point out that not being directly challengeable does not entail being NAI: a linguistic expression may be not directly challengeable also if it is simply not challengeable at all. For example, subjective content like pain reports is AI and yet not challengeable, because the speaker has privileged access to their own sensations:

(17) A: I have a splitting headache.
B: #No, you don't.
Korotkova (2016a): (9)

The data presented in §2.1 and §2.2.1 show that the direct and the secondhand reportative ev bases do not represent subjective content, but events in the world (e.g. the evidential events of being a participant in the scope situation, or reading someone's tweets) — and more than one person could have the same kind of evidential access to those events (observing the same scope situation or reading the same tweets). Thirdhand reportatives initially look like they confirm the prediction of the subjective hypothesis that evidentials are not challengeable in any way, but, as discussed in §2.2.2, the reason is not subjectivity, but the low bar for rumor quality: any rumors about anything could in principle exist. Therefore, evidentials as a category are not inherently subjective in the same way that first-person pain reports are.

Section §2 provides novel evidence that evidentials are NAI by showing: (i) that they are indirectly challengeable; (ii) how responses that target the evidential base — even when they accept it — affect the QUD (Simons et al., 2010; Beaver et al., 2017): they change it. For example, in the heritage speaker

data, (2), the QUD is what the boy was like as a baby, but his mom changes the QUD to what he remembers. In the terrorist example, (3), the QUD is how many attackers there were and whether the attacks were organized or spontaneous; the responders change the QUD to whether the commenter who used the direct evidential was a witness or not. In the party example, (4), the question is about what happened at the party, but upon hearing the unexpected evidential, the responder changes the QUD to whether the commenter who used the direct evidential was at the party. If evidentials were AI meaning, they shouldn't change the QUD because they would be part of the QUD.

The aforementioned examples show that evidentials can be used to change the QUD. The next example demonstrates that they cannot be used as AI content by replicating the at-issueness test offered by Bary and Maier (2021):

(18) A: What makes you think that Mary is ill?
B: (i) #Allegedly, she has the flu.
(ii) #Ze **schijnt** griep te hebben.
she seems flu to have
'She has the flu, *reportedly*' [Dutch]
(iii) John told me that she has the flu.

The idea is that if an evidential marker is not interpreted at issue, it cannot be a felicitous answer to an explicit question about how the speaker came to know about the scope situation. Like the Dutch reportative *schijnt*, in Bulgarian, too, neither the reportative, nor the direct evidential allow this:

(19) A: Kak nauchi (vchera), che vali?
how lean.DIR yesterday that rains
'How did you find out that it was raining?'
B: #Valja-**lo**. cf. B′: ✓**Kazaha**
rain-REP
mi.
told.3PL.DIR me
'It was raining, *reportedly*.' | 'I was told.'
B″: #Vale-**she** cf. B‴: ✓ **Vidiah**
rain.DIR saw.1SG.DIR
'It was raining, *I saw*.' | 'I saw.'

Thus, evidentials in Bulgarian can only be used to change the QUD and not to address an already established QUD. This explains why challenging them changes the QUD in the data in §2.

3.1.2 Presupposition

In addition to providing novel evidence that evidentials are NAI, the findings in this paper also inform what type of NAI content they are, supporting a presuppositional account and partially the sincerity account (Faller, 2002), and ruling out alternative hypotheses, such as NAI assertion (Murray, 2010).[11]

Izvorski's account can be generalized as:[12]

(20) the speaker has evidence of type x for the scope situation
where x is a variable for the type of evidence: direct, reportative, inferential, etc.

Such an account predicts that an ev base could be challenged indirectly, similarly to presuppositions (von Fintel, 2004). The data introduced in Section §2 demonstrate that this prediction is borne out, providing novel evidence for the presupposition hypothesis in addition to direct denials, reproduced from Izvorski (1997) in (1), avoiding the reservations about them as a diagnostic for (N)AIness discussed in §3.1.1.

Another parallel between presuppositions and the ways in which evidentials are challenged is (im)plausibility. Potts (2013) points out that a presupposition can be denied accommodation by an interlocutor on the basis of being implausible:

(21) My {giraffe/sister} destroyed my homework.

The less plausible presupposition is much easier to be refused accommodation. The present paper showed that plausibility is an important factor in evidential challenges as well. This parallel also explains why evidentials seem to be generally easily accommodated and why specific conditions of implausibility need to be in place in order for a challenge to become a felicitous conversational move.

All presuppositional accounts of evidentiality to date are also modal accounts, most notably Izvorski (1997) and Matthewson et al. (2007). However, it need not be so, as illustrated by the following account of the direct evidential that is not modal but is presuppositional:

(22) assertion: p
presupposition: the speaker (consciously) participated in the scope situation s such that s

[11] The conventional implicature hypothesis (Koev, 2016) is not discussed here, see Murray (2010): §3.7, §5.4.3 for arguments against it that are independent of challengeability.

[12] Izvorski's account focuses on the indirect base, this is a generalized formulation that extrapolates the idea to other evidential bases; it is adapted to the terminology used here.

`exemplifies`[13] p

Therefore, the findings reported in this paper support the presuppositional hypothesis without informing or subscribing to the modal one.

Now let's look at why two other NAI accounts are less preferred than the presuppositional one.

Faller (2002) encodes evidentials in the sincerity conditions of an utterance. These include the condition that the speaker believes what they say (for assertion), and have evidence for it. In a sense, we can regard evidentials as simply specifying what kind of evidence. Similarly to the remark in Aikhenvald (2004) about evidential lies, this hypothesis correctly predicts that challenging an evidential can felicitously occur and it would amount to challenging the sincerity condition for that type of evidence, as for example in the terrorist attack example, (3).

However, Faller's account entails that insincerity is not just sufficient but also a necessary condition for a challenge to be felicitous. This incorrectly excludes examples like (2) (the heritage speaker, who is sincere) and (4) (the party), where the challenger is signalling her defeated expectations and asks for confirmation. Thus, while Faller's account captures some of the data, it undergenerages felicitous data, while the presuppositional account predicts all the data examined here.

Murray (2010) proposes that evidentials in Cheyenne constitute a new type of NAI content, NAI assertion: a non-negotiable update that directly restricts the common ground. This hypothesis seems to make similar predictions about challengeability as the subjectivity account: that the evidential content cannot be contested at all; therefore, it would not generate any of the novel data presented in this paper, except perhaps correctly ruling out the infelicitous AI uses of evidentials in (19). The NAI assertion hypothesis is therefore untenable for Bulgarian.

3.2 Different ev bases have different challengeability profiles

The second novel finding reported in this paper is that different ev bases do not have uniform behavior within the same language and therefore there is no one size fits all test informative of it; they are challenged under different conditions and in order to verify whether and how they can be challenged, scenarios need to be tailored to each respective base's properties. This paper has introduced the methodology to do this for the direct and second-hand reportative bases (and the lack of such for the thirdhand reportative). From the results emerge some broader methodological implications: When testing for direct challengeability, some works provide examples for just one evidential base, assuming that the behavior of others is analogous. Let's reconsider the example this paper opened with, (1), repeated schematically here:

(23) A: p-REP
B: That's not true. {$\neg p$/*You didn't *hear* p}

Based on the findings in this paper, it is now clear that this test does not demonstrate what it aims to (that evidentials are NAI), because the infelicity does not arise from the nature of evidentiality as a whole, but as a property of rumors in particular.

But this test has previously been taken as informative, and has been replicated over and over, as illustrated below, and with the same results, which is now unsurprising given that the results are not driven by a grammatical property.

(24) A: Ines-qa qaynunchay ñaña-n-ta-**s**
Ines-TOP yesterday sister-ACC-REP
watuku-sqa.
visit-PST2
'Inés visited her sister yesterday, *I'm told*.'

B: Mana-n chiqaq-chu. #Mana-n chay-ta
not-BPG true-NEG not-BPG this-ACC
willa-rqa-sunki-chu.
tell-PST1-3S2O-NEG
'That's not true. #You were not told this.'
[Cuzco Quechua], Faller (2002): (160-1)

(25) A: Méave'ho'eno é-hestahe-**sestse**
Lame Deer 3-be.from-REP.3SG
Mókéé'e.
Mókéé'e
'Mókéé'e is from Lame Deer, *I hear*.'

B: É-sáa-ne-hétóméto-hane-∅.
3-neg-AN-be.true-MODB-DIR
'That's not true...

B′: ✓É-sáa-hestahe-he-∅ M-o
3-neg-be.from-MODA-DIR
...She's not from L.D.'

B″: #Né-sáa-ne-néstó-he-∅
3-neg-AN-hear.B-MODA-DIR
...#You didn't hear that.'

B‴: #Hovánee'e
nobody
é-sáa-ne-hé-he-∅
3-neg-AN-say.-MODB-DIR

[13] In the sense of Kratzer (2002, 2012).

...#Nobody said that.'
Murray (2010): 51, (3.5)

Such examples are often the only ones provided to demonstrate direct non-challengeability. For example, Murray (2010) describes four types of evidentials in Cheyenne, of which the reportative has both a secondhand and a thirdhand function, and yet only one example is provided, the thirdhand reportative (25). While Murray (2010) asserts that the results are the same for the other ev bases, the methodology for testing that is not provided — and this matters because, as the present paper has shown, each base comes with its idiosyncrasies, which have effects not only in indirect denials, but also in direct ones, as discussed here for the reportative.

To sum up, the findings in this paper have methodological implications not only for indirect challenge tests, but also for direct ones, showing that the specifics of each ev base need to be taken into account.

4 Conclusion

While there has been a lot of interest in the literature in whether evidential markers are directly challengeable, this paper provides the first empirical investigation into the question of whether they are indirectly challengeable and demonstrates how this diagnostic differentiates various theoretical hypotheses on evidentiality.

It emerged also that the direct ev base and the secondhand reportative one are challengeable much like presuppositions, while the thirdhand reportative base is not challengeable at all, but for reasons that have nothing to do with the nature of the category of evidentiality as a whole, contrary to what has been previously assumed.

The empirical evidence lays out a methodological blueprint for testing indirect challengeability that can be used for other languages and extended to other ev bases, and has implications for existing tests for direct challenges.

The findings strongly support a presuppositional account of evidentiality (not necessarily a modal one), mildly support a sincerity conditions-based account, and falsify subjective and NAI-assertion accounts with regards to Bulgarian.

Acknowledgments

I thank three anonymous reviewers, Magdalena Kaufmann, and Scott AnderBois for comments on this paper. All errors are my own.

References

Alexandra Aikhenvald. 2004. *Evidentiality*. Oxford University Press.

Scott AnderBois. 2019a. At-issueness in direct quotation: the case of Mayan quotatives. In *Proceedings of Semantics And Linguistics Theory (SALT) 29*, pages 371–391.

Scott AnderBois. 2019b. Reportatives and quotatives in Mayan languages. In *Proceedings of Form and Analysis in Mayan Linguistics (FAMLi) 5*.

Corien Bary and Emar Maier. 2021. The landscape of speech reporting. *Semantics and pragmatics*, 14:8.

David I. Beaver, Craige Roberts, Mandy Simons, and Judith Tonhauser. 2017. Questions under discussion: Where information structure meets projective content. *Annual Review of Linguistics*, 3(1):265–284.

Martina Faller. 2002. *Semantics and pragmatics of evidentials in Cuzco Quechua*. Ph.D. thesis, Stanford University.

Kai von Fintel. 2004. Would you believe it? The King of France is back! Presuppositions and truth-value intuitions. In M. Reimer and A. Bezuidenhout, editors, *Descriptions and Beyond*, 315–341. OUP.

Roumyana Izvorski. 1997. The present perfect as an epistemic modal. *Semantics and Linguistic Theory*, 7:222–239.

Todor Koev. 2016. Evidentiality, learning events and spatiotemporal distance: The view from Bulgarian. *Journal of Semantics*, 34(1):1–41.

Natalia Korotkova. 2016a. Disagreement with evidentials: A call for subjectivity. In *Jersem: The 20th workshop on the semantics and pragmatics of dialogue*, 65–75.

Natalia Korotkova. 2016b. *Heterogeneity and uniformity in the evidential domain*. Ph.D. thesis, UCLA.

Natasha Korotkova. 2020a. Evidential meaning and (not-) at-issueness. *Semantics and Pragmatics*, 13:4.

Natasha Korotkova. 2020b. The subjective heart of evidentiality. Presented at GLOW 43. Available at https://osf.io/qkjht/.

Kepa Korta and Larraitz Zubeldia. 2014. The contribution of evidentials to utterance content: Evidence from the Basque reportative particle *omen*. *Language*, 90(2):389–423.

Angelika Kratzer. 2002. Facts: Particulars or information units? *Linguistics and Philosophy*, 25(5):655–670.

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

Ivan Kutsarov. 1994. *Edno ekzotichno naklonenie v balgarskija ezik*. Sofia University Press.

Lisa Matthewson, Henry Davis, and Hotze Rullmann. 2007. Evidentials as epistemic modals: Evidence from St'át'imcets. *Linguistic Variation Yearbook*, 7.1:201–254.

Eric McCready and Nicholas Asher. 2006. Modal subordination in Japanese: Dynamics and evidentiality. *University of Pennsylvania Working Papers in Linguistics*, 12(1):20.

Sarah E. Murray. 2010. *Evidentiality and the Structure of Speech Acts*. Ph.D. thesis, Rutgers.

Christopher Potts. 2013. Presupposition and implicature. *The Handbook of Contemporary Semantic Theory, 2nd edition. Oxford: Wiley-Blackwell*.

Mathias Schenner. 2010a. Embedded evidentials in German. In Gabriele Diewald and Elena Smirnova, editors, *Linguistic realization of evidentiality in European languages*, 157–185. Walter de Gruyter.

Mathias Schenner. 2010b. Evidentials in complex sentences: Foundational issues and data from Turkish and German. In Tyler Peterson and Uli Sauerland, editors, *Evidence from Evidentials*, volume 28 of *The University of British Columbia Working Papers in Linguistics*, 183–220. University of British Columbia.

Magdalena Schwager. 2010. On what has been said in Tagalog. In Tyler Peterson and Uli Sauerland, editors, *Evidence from evidentials*, volume 28 of *The University of British Columbia Working Papers in Linguistics*, 221–246. University of British Columbia.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. *Semantics and Linguistic Theory*, 20:309.

Anastasia Smirnova. 2013. Evidentiality in Bulgarian: Temporality, epistemic modality, and information source. *Journal of Semantics*, 30(4):479–532.

Thomas Willett. 1988. A cross-linguistic survey of the grammaticization of evidentiality. *Studies in Language*, 12(1):51–97.

Construction Coordination in First and Second Language Acquisition

Arabella J. Sinclair
University of Amsterdam
Amsterdam, Netherlands
a.j.sinclair@uva.nl

Raquel Fernández
University of Amsterdam
Amsterdam, Netherlands
raquel.fernandez@uva.nl

Abstract

Repetition of linguistic forms is a pervasive coordination mechanism in interactive language use. In this paper, we investigate patterns of cross-participant repetition in dialogues where participants have different levels of linguistic ability. Achieving a better understanding of these patterns can not only shed light on how humans coordinate in conversation, but may also contribute to developing more natural and effective dialogue agents in education contexts related to language learning. Our approach is novel in several respects: We focus on multi-word constructions at the lexical and morphosyntactic level, consider both first and second acquisition dialogue, and contrast these setups with adult native conversation. The results of our study show that language acquisition scenarios are characterised by richer inventories of shared constructions but lower usage rates than fluent adult dialogues, and that shared construction use evolves as the linguistic ability of the learners increases, arguably leading to a process of routinisation.

1 Introduction

Interacting through conversation, although arguably the most intuitive form of language use, requires complex interpersonal coordination. Part of such coordination is realised by the tendency of interlocutors to repeat each other's linguistic forms. Indeed, dialogue partners have been shown to align their behaviour at a range of different levels, from phonetic features, lexical choice and syntactic structures to body posture, eye-gaze or gestures (Brennan and Clark, 1996; Pardo, 2006; Reitter et al., 2011; Holler and Wilkin, 2011; Rasenberg et al., 2020). In this study, we investigate patterns of cross-participant repetition of lexical and structural constructions present in two language acquisition settings. We compare dialogues between young children and their caregivers (L1) with learners practicing English as a second language with a tutor (L2), contrasting these to adult native dialogue.

Language acquisition dialogues are particularly interesting scenarios to study alignment since the language choices made by both speakers are not solely for communicative or social purposes, but play a key role in the process of language learning. Therefore, a better understanding of alignment patterns in these scenarios can contribute to developing more natural and effective dialogue agents in education contexts (Litman and Silliman, 2004; Graesser et al., 2005; Steinhauser et al., 2011; Katz et al., 2011; Sinclair et al., 2019b). Beyond education, linguistic alignment has been shown to lead to increased naturalness and task success in dialogue systems (Lopes et al., 2015; Hu et al., 2016) and has been incorporated into chatbots and dialogue assistants (Hoegen et al., 2019; Gao et al., 2019).

In the present study, we adopt a usage-based perspective to language acquisition and investigate multi-word constructions in the sense of Construction Grammar (Goldberg, 2006; Tomasello, 2003; Bybee, 2010). According to this tradition, constructions are form-function units acquired through interaction, where a form is a particular configuration of structural and/or lexical elements. Constructions have been shown to play a role in both first and second language acquisition (Diessel, 2013; Ellis, 2013). In this paper, we focus on cross-participant alignment of constructions, i.e., multi-word expressions at the lexical and morphosyntactic level that are used by both participants within a given dialogue. We call such expressions *shared constructions*. Examples are shown in Table 2, Section 4.

Using data from four different dialogue corpora, we extract the inventory of shared lexical and morphosyntactic constructions within a dialogue and compute several usage measures for these constructions. Our results demonstrate that shared constructions are an important aspect of interaction and

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

reveal interesting contrasts. We find that language acquisition scenarios, particularly regarding L2, are characterised by richer inventories of shared constructions but lower usage rates than fluent adult dialogues. However, over the course of learning, usage rates significantly increase, arguably due to a process of routinisation. With higher linguistic ability, shared constructions become more complex, are more frequently introduced by the learner, and their cross-speaker repetition is less affected by local mechanisms possibly related to priming.

2 Alignment in L1 and L2 Learning

First and second language acquisition have key differences, for example regarding the mental and social maturity of the learner and the absence vs. presence of self-awareness regarding the learning process (Cook, 1973). In addition, in adult L2 acquisition the learner already has full knowledge of their first language, which conditions how a second language will be learned (Cook, 2010). Yet L1 and L2 acquisition also share important features: They involve similar learning stages, with particular structures being acquired in a relatively fixed order (McDonough et al., 2013), and the use of formulaic speech is present in both learning processes (O'Donnell et al., 2013). These similarities and differences are likely to influence the patterns of cross-speaker construction repetition exhibited in these two scenarios. With this study, we aim to gain understanding of these patterns, contrasting them to those present in adult native dialogue.

Several previous studies have analysed alignment and repetition processes in first and second language acquisition dialogue, but to our knowledge these two settings have not been compared directly. In the context of L1 acquisition dialogue, it has been shown that there is cross-speaker coordination at lexical and syntactic levels and that this occurs at higher rates in adjacent turns (Dale and Spivey, 2005, 2006; Fernández and Grimm, 2014; Misiek et al., 2020). Clark and Bernicot (2008) argue that cross-participant lexical repetition in child-adult dialogue is typically used to draw attention to the partner's utterances and to add the repeated information to the common ground. While more recently, Denby and Yurovsky (2019) found that parental alignment at the level of syntax and function words is also present and is a strong predictor of vocabulary development in young children. In L2 acquisition dialogue, repetition patterns have also been shown to occur and to serve several functions, such as testing newly learned words, clarifying, or indicating understanding and misunderstanding (Allwood and Ahlsén, 1986; Broeder, 1992; Costa et al., 2008). Furthermore, alignment at the lexical level and coordination in terms of dialogue act usage between tutors and L2 learners has been shown to increase with language ability (Sinclair et al., 2018, 2019a).

In this paper we focus on multi-word lexical and morphosyntactic shared constructions. We consider both L1 and L2 acquisition dialogue, compare these two setups, and contrast them with adult native conversation.

3 Data

Child-adult dialogue (L1) We use a set of dialogues from the CHILDES Database (MacWhinney, 2000). In line with previous work (Chouinard and Clark, 2003; Dale and Spivey, 2005, 2006; Fernández and Grimm, 2014), we draw longitudinal data from the following three English child-adult corpora involving three different young children in relatively early stages of first-language acquisition: Abe (age range of the child 2;5–5;0) from the Kuczaj corpus, Sarah (age range 2;6–5;1) from the Brown corpus, and Naomi (age range 1;11–4;9) from the Sachs corpus. The dialogues are between a caregiver and a child who are interacting in free play. Since our focus is on multi-word constructions, we selected all dialogue transcripts from each of these three corpora where the child utterances have a minimum mean length of 2 words.

Student-tutor dialogue (L2) We use a set of dialogues from the Talkbank Database, specifically from the Barcelona English Language Corpus (BELC) (Muñoz, 2006). The BELC dialogues involve an English language tutor and a high school student (ranging in age from 11 to 18 years old) whose native language is Spanish or/and Catalan (students may be bilingual). The tutor conducts an interview in English about daily life aspects. The interviews are semi-guided, but learner-initiated topics are occasionally present since the goal is to favour natural interaction. The dialogues were gathered at four time points: after 200 hours, 416 hours, 726 hours, and 826 hours of English-language instruction (level 1, 2, 3, and 4, respectively).

Dialogue between adult native speakers As control group, we use two different corpora of adult

	L1		L2		MapTask		Switchboard	
total # dialogues	379		118		128		1155	
utterances / dialogue	388±220		132±48		162±83		192±80	
tokens / dialogue	1527±753		685±245		1182±639		1618±664	
	Adult	Child	Tutor	Student	Giver	Follower	A	B
utterance length	4.4±0.7	3.8±1.4	6.0±0.7	4.2±1.5	10.2±2.1	4.4±1.3	8.6±2.3	8.6±2.4
% utterances / dialogue	0.46±0.1	0.54±0.1	0.59±0.0	0.41±0.0	0.51±0.0	0.49±0.0	0.51±0.1	0.49±0.1
type-token ratio	0.4±0.1	0.3±0.08	0.3±0.1	0.4±0.1	0.4±0.1	0.2±0.1	0.3±0.1	0.3±0.1
vocabulary overlap	0.4±0.1	0.4±0.14	0.3±0.1	0.5±0.1	0.6±0.1	0.4±0.1	0.4±0.1	0.5±0.1

Table 1: Corpus statistics. Except *total number of dialogues*, we report mean and standard deviation per dialogue. For Switchboard, A is the participant who speaks first in each dialogue. Type-token ratio indicates the level of lexical diversity per dialogue. Vocabulary overlap refers to the percentage of word types used by a participant that are also used by the dialogue partner within a dialogue.

native dialogue: MapTask (Anderson et al., 1991), a corpus of task-oriented dialogue where the participants have different roles, and Switchboard (Jurafsky et al., 1997) where there is no role difference. The MapTask dialogues consist of one participant (the instruction giver) directing the other (instruction follower) to navigate to a point on a map. In the Switchboard dialogues both participants were asked to make conversation over the phone about one of a pre-specified range of daily life topics.

Table 1 summarises the corpora used in our analysis. All dialogue corpora are freely available, distributed tokenised and with part-of-speech tags.

4 Shared Constructions: Extraction and Key Properties

We focus on multi-word constructions used by both dialogue participants within a conversation. In our approach, constructions consist of at least two contiguous non-punctuation tokens at the utterance level. A construction becomes *shared* within a dialogue once both participants have used it, that is, once it has appeared in at least one utterance per dialogue participant. We consider lexical constructions (i.e., sequences of words, such as 'go to the' or 'how old are you') as well as morphosyntactic patterns (i.e., sequences of part-of-speech tags). Reusing a morphosyntactic pattern may or may not involve repeating some or all of its lexical realisation. For example, '`PREP N`' could be realised lexically as 'at home' or 'for friends', and '`CONJ PRO V`' as 'if you have' or 'if I have'. Table 2 contains examples of the types of shared constructions in the L1 and L2 corpora we examine in this study.

The automatic extraction of shared constructions per dialogue is an instance of the longest common subsequence problem (Hirschberg, 1977; Bergroth et al., 2000), which can be solved in linear time given the total number of tokens in a dialogue. For each dialogue in our corpora, we extract the inventories of shared lexical and morphosyntactic constructions using the method proposed by Duplessis et al. (2017a; 2017b).[1] A total of 29 (out of 1155) Switchboard dialogues do not contain shared constructions of at least length 2, thus are excluded.

For each of the two dialogue-specific inventories of shared constructions (lexical and morphosyntactic), we compute the following measures:

- *Relative inventory size:* The number of shared construction types normalised by the length of the dialogue in tokens. This measure indicates how large the set of shared constructions

T—144: is it a **big bedroom** or a small bedroom?
S—145: **big bedroom**.
T—146: a **big bedroom** okay .

(a) L2: BELC

A—550: it had lollipops **in it**
C—551: what's **in it**
A—552: it doesn't open it just a whosejigger
...
C—556: a lollipops is **in it**

(b) L1: CHILDES

Table 2: Example dialogue excerpts from BELC (T: tutor, S: student) and CHILDES (A: adult, C: child). Underlined expressions indicate shared morphosyntactic constructions for the patterns '`ADJ N`' in BELC and '`DET N`' and '`PREP PRO`' in CHILDES. Expressions in bold indicate shared **lexical** constructions.

[1] The original code by Duplessis et al. (2017b) is available at https://github.com/GuillaumeDD/dialig. We adapt it to extract sequences of POS tags besides surface text and to constrain the minimum sequence length to two tokens.

is taking into account dialogue length. We take this to capture the relative importance of the use of these shared constructions as a conversational mechanism.

- *Construction length:* Average length in tokens of the shared constructions in a dialogue.
- *Usage rate:* Proportion of utterances in the dialogue which contain a shared construction.

The plots in Figure 1 summarise the properties of the shared construction inventories found in the corpora under investigation. Regarding relative inventory size (left), we find that the learner dialogues have richer inventories of shared constructions than the fluent adult dialogues, i.e., they have more shared construction types per number of tokens in a conversation. This suggests that reuse of constructions across speakers is an important feature of this type of interaction. In particular, L2 dialogues have the richest construction inventories, both at the lexical, and morphosyntactic level. All differences between corpora are statistically significant (Welch's independent t-test, $p < 0.01$).[2]

As for shared construction length (middle), we observe that lexical shared constructions in L1 dialogues are significantly shorter (2.18 words on average), while their length is very similar across the other three corpora (around 2.5 words on average). More pronounced differences can be observed with respect to morphosyntactic shared constructions, which are significantly longer in the fluent adult dialogues (above 3 tokens on average; for example '`PRO V ADV V`'), reflecting the higher linguistic proficiency level of the speakers in these dialogues.

Finally, regarding usage rate (right), overall there is a higher proportion of utterances containing shared constructions in the adult fluent corpora than in the learner dialogues. Thus, while there are fewer shared construction types in the fluent dialogues these constructions arguably correspond to very common collocations in English and are therefore present in a higher proportion of utterances. Focusing on the learner dialogues, we see a clear contrast: The proportion of utterances per dialogue that contain shared constructions is significantly higher in L1 than in L2, in particular regarding morphosyntactic constructions (0.31 vs. 0.58, $p < 0.001$). We attribute the high usage rate in L1 to the high degree of repetition present in this type of dialogue (Bannard and Lieven, 2009).

Overall, these results suggest that in L2 dialogues, establishing shared constructions is a prominent conversational mechanism but does not necessarily involve frequent use of such constructions. For L1, repetition is the norm. In the following section we analyse in more detail how these patterns may relate to the language learning activity inherent to both L1 and L2 acquisition dialogue.

5 The Dynamics of Shared Constructions in Learner Dialogues

We now investigate in more detail how shared constructions are established and exploited in L1 and L2 acquisition setups. We address two aspects: differences in role (adult vs. child and tutor vs. student), and changes over the course of learning.

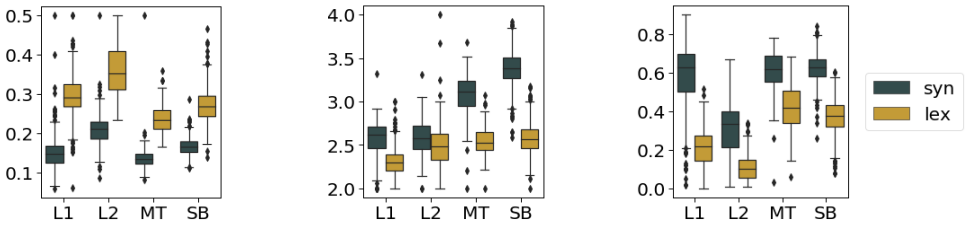
5.1 Differences across types of dialogue participant

We compute two additional measures for each of the dialogue-specific inventories of constructions:

- *Construction initiator:* Percentage of shared constructions introduced by each dialogue participant. Concretely, we use *initiator* to designate the dialogue participant who first uses a construction that will become shared, and *re-user* for their dialogue partner. Naturally, after the first two usages of a construction, establishing it as shared, both participants may repeat the construction further.
- *Usage rate per participant:* Proportion of an individual speaker's utterances which contain a usage of a shared construction.

We firstly compare the differences in which speaker acts as *initiator* or *re-user* of shared constructions in a dialogue. While initiating a construction takes work (as the speaker needs to draw from their own linguistic knowledge without the scaffolding provided by the partner's usage), shared constructions are only established when repeated by the dialogue partner. Our hypothesis is that the significance of reusing a construction initiated by the partner varies depending on the relative roles of the initiator and re-user. In the language acquisition dialogues, a reuse by the learner serves to uptake and practice constructions introduced by the adult or tutor, while a reuse by the more proficient speaker serves to acknowledge and ratify a construction initiated by

[2] Unless stated otherwise, all significance values reported in Sections 4 and 5 use Welch's independent t-test computed with the Python package SCIPY, TTEST_IND, version 1.3.3.



(a) Relative inventory size (b) Construction length (c) Usage rate

Figure 1: Key properties of shared lexical (lex) and morphosyntactic (syn) constructions per corpora: CHILDES (L1), BELC (L2), MapTask (MT), and Switchboard (SB).

the learner, occasionally to both ratify and correct (Clark and Bernicot, 2008; Chouinard and Clark, 2003). Thus both directions have potential to facilitate language learning.

Figure 2a shows our results regarding construction initiator. The first aspect worth highlighting is that while there are differences across participant types in the L1 and L2 acquisition dialogues, these are not extreme – thus confirming that the two directions mentioned above are both at play. The differences across participants in the learner dialogues are in fact less pronounced than in MapTask, where the asymmetric task-related roles of the participants lead to more striking differences regarding construction initiation: In this case, the instruction giver has a strong tendency to initiate and the instruction follower to reuse for both lexical and morphosyntactic constructions.

In contrast, the learner dialogues exhibit more nuanced patterns. We find that in L1 the child is more likely to introduce constructions that will be repeated verbatim at the lexical level by the adult (53% vs. 45% average initiation by the adult and the child, respectively, $p < 0.001$), while the adult is more likely to introduce constructions that will be taken up at the morphosyntactic level by the child (52% adult vs. 47% child average initiation, $p < 0.001$). L2 acquisition dialogues show the same tendency regarding morphosyntactic constructions, with the tutor being more likely to introduce constructions that will be reused by the learner at the morphosyntactic level (58% vs. 42% average initiation by the tutor and the student, respectively, $p < 0.001$). In L2 there is however no difference regarding percentage of initiator and re-user roles for lexical shared constructions.

We interpret these results as an indication that in L1 acquisition reuse of lexical constructions is slightly more likely to constitute a ratification by the adult than an uptake by the child. While in both L1 and L2 acquisition, the reuse of morphosyntactic constructions is more likely to be the result of uptake by the less proficient speaker than a confirmation strategy by the adult or the tutor.

Finally, no significant differences are observed between speakers in the Switchboard corpus (not shown in Figure 2), where participants exhibit neither the asymmetry of task-related role (MapTask) nor language ability (CHILDES & BELC). Thus, patterns of initiation and reuse of constructions appear to be tightly connected to the presence of asymmetries between dialogue participants.[3]

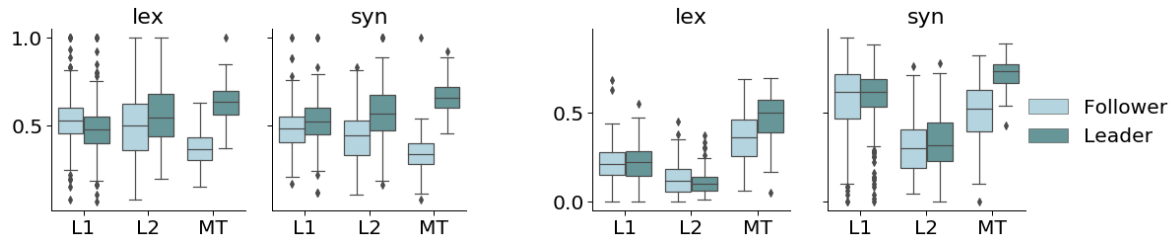
Turning our attention to usage rate per participant (Figure 2b), differences across participant types are minor in the learner dialogues: only L2 speakers show significant differences at the lexical level, with students showing a higher proportion of utterances containing shared lexical constructions than their tutors (0.12 vs. 0.10, $p < 0.05$). Again we observe clear contrasts in MapTask, with no significant differences in Switchboard.

5.2 Changes over the course of learning

Next, we investigate the dynamics of shared construction use over the language learning process regarding size of construction inventories, construction length, usage rate, and initiation. For space reasons, Figure 3 displays some key results only for the L2 acquisition dialogues.

Regarding the relative size of the inventories of shared constructions, a weak positive correlation with child age shows that there is a mild increase in L1 acquisition at both lexical and morphosyntactic levels (Spearman's $r = 0.2, p < 0.001$). We do not observe any significant changes over the ability lev-

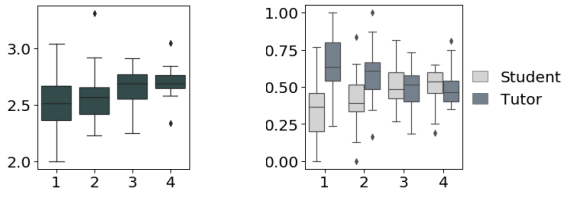
[3]Other kinds of asymmetry may also have an impact. E.g., alignment patterns (at levels other than constructions) have been shown to be influenced by social power (Danescu-Niculescu-Mizil et al., 2012; Noble and Fernández, 2015).



(a) Construction initiator (b) Usage rate

Figure 2: Trends per participant type across the three asymmetric corpora. *Follower* indicates child, student or instruction follower; *Leader* indicates adult, tutor or instruction giver in the L1, L2 and MT corpora, respectively.

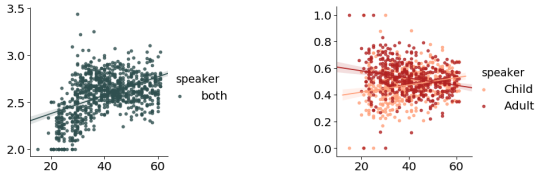
els in L2 acquisition dialogues. As for shared construction length (Figure 3a), we find a significant increase in the length of shared morphosyntactic constructions in both types of setups ($r = 0.46$ in L1 and $r = 0.31$ in L2, $p < 0.001$), while the length of lexical constructions does not significantly change over time. Regarding usage rate,there is a clear increase for both lexical ($r = 0.34$ in L1 and $r = 0.48$ in L2, $p < 0.001$) and morphosyntactic shared constructions ($r = 0.41$ in L1 and $r = 0.62$ in L2, $p < 0.001$). Finally, concerning shared construction initiation (Figure 3b), while there are no significant differences across level regarding the initiation of shared lexical constructions, we find that both L1 and L2 learners are able to introduce a higher proportion of morphosyntactic constructions with increased ability level ($r = 0.26$ in L1 and $r = 0.31$ in L2, $p < 0.001$). In BELC in particular, by level 4, speakers show equal likelihood of introducing shared morphosyntactic constructions.



(a) Length (b) Initiator

Figure 3: Trends on morphosyntactic shared construction use across student ability levels in BELC.

In summary, over the course of learning, morphosyntactic shared constructions become more complex and learners are progressively more able to introduce them. Moreover, both lexical and morphosyntactic shared constructions are used more frequently (higher proportion of utterances) in the dialogues as language learning advances. We interpret this as indication of increased ability leading



(a) Length (b) Initiator

Figure 4: Trends on morphosyntactic shared construction use across age in months in Childes.

to greater likelihood of routinisation. We discuss this further in the next section, where we explore local patterns of construction repetition.

6 Effect of Locality on Cross-Speaker Construction Repetition

We now analyse the extent to which cross-speaker construction repetition is local, i.e. influenced by distance in utterances between usages. Concretely, we test whether the likelihood of speaker B repeating an expression used by their dialogue partner A decreases as the number of utterances from A's use of the expression increases. A similar kind of analysis has been carried out on fluent adult dialogue for single words and syntactic rules (Reitter et al., 2006; Howes et al., 2010; Reitter et al., 2011; Healey et al., 2014). Here our aim is to shed light on the importance of local dynamics on multi-word lexical and morphosyntactic construction reuse patterns in language acquisition dialogue.

A negative effect of distance (i.e., a higher proportion of construction repetition at short distance) may have two main causes: (1) it may be due to priming effects, since priming is assumed to be strongest immediately after a representation has been activated and then decay with distance from the prime (Reitter et al., 2006, 2011); (2) it may be due to other functions of repetition, such as ac-

knowledging, elaborating or clarifying, which take place locally in dialogue structure (Clark, 1996). In contrast, routinisation mechanisms, whereby formulaic constructions are established as part of the repertoire of expressions of a speaker, are expected to be less affected by proximity than priming and grounding moves (Du Bois, 2014; Pickering and Garrod, 2005). All these mechanisms have been shown to influence, and play a role in language learning (Broeder, 1992; Wood, 2002; Chouinard and Clark, 2003; Huttenlocher et al., 2004; Clark and Bernicot, 2008; Costa et al., 2008; Gerard et al., 2010)

We expect that in the fluent adult dialogues the kind of multi-word constructions we focus on in this study will be more indicative of routinisation (in the sense of frequent collocations) than priming or grounding moves. While in the learner dialogues we hypothesise priming and grounding will be more prominent, particularly at lower levels of linguistic ability, while routinisation will develop further as learning progresses and constructions become more established in the learners' own repertoires. Therefore, we expect that repetition of constructions at shorter distance will be stronger in the language acquisition scenarios than in fluent adult dialogues, and that the negative effect of distance will become weaker over the course of learning.

6.1 Methods

We model distance in terms of utterances, considering a window of 25 utterances after the use of a construction in the shared inventory. Given that a participant has used construction e in utterance u_t, for each utterance $u_{t'}$ by the other participant (where $t < t' \leq 25$) we record whether e is used and the distance $d = t' - t$ from u_t. We extract this information for each construction in the inventory of shared constructions per dialogue. This allows us to compute a *cross-speaker construction repetition proportion* (xCRP) value for each distance $d \leq 25$, defined as the number of times a construction is repeated by the other participant over the total number of opportunities available for cross-speaker repetition, at a given distance.

Distance effects are obviously dependent on the temporal order of utterances in the dialogue. To control for chance effects, we create a scrambled version of each dialogue, maintaining the turn-taking relationship but shuffling utterance order.

6.2 Results

Figure 5 shows xCRP per distance value up to a distance of 25 utterances between repetitions (x axis shows log-transformation of this value) for the original dialogues and the shuffled control dialogues. As can be observed in the plots, there is a significant locality effect of xCRP in the original dialogues that is not present in the control dialogues. We fit General Linear Models (GLM) to the original dialogues per corpus in order to investigate the effect of distance on shared construction use and its interaction with participant type and ability level.[4]

We fit a GLM with (log-transformed) distance as predictor and xCRP as dependent variable. In the learner corpora and MapTask, the effect of distance is significant for both shared construction types: The probability of repeating a construction is highest in adjacent turns (distance 0) and then decreases progressively as distance from the use of the expression increases. The effect is stronger in the L1 and L2 corpora. In Switchboard, there is no distance effect regarding shared lexical constructions and a significant effect in the opposite direction regarding morphosyntactic constructions, i.e., the probability of repeating a morphosyntactic expression by the dialogue partner is lowest in adjacent turns. This confirms similar results regarding structural divergence in adjacent turns in non-task oriented fluent adult dialogue (Healey et al., 2014).

The plots in Figure 5 also show the distance effect broken down per participant type. To check whether there are significant differences between participants, we fit a second set of GLMs with distance and participant type as predictors. We find a significant interaction between distance and participant type in the learner corpora for shared lexical constructions, with adult and tutor showing stronger effect size than children and students. This difference between speakers is more pronounced in L2 than L1. As for MapTask, while there is a significant effect of participant type on xCRP at the morphosyntactic level, there is no significant interaction between distance and participant type. No differences are observed in Switchboard.

In order to test our hypothesis that the distance effect will change with ability level, we fit a third set of GLMs with distance and ability level as pre-

[4] We use a Binomial link function, to capture, of the opportunities for repetition, the repetition proportion (xCRP), a value in the interval of $[0, 1]$. Python's SCIPY GLM package version 1.3.3. is used. Full output of the models can be found in the appendix.

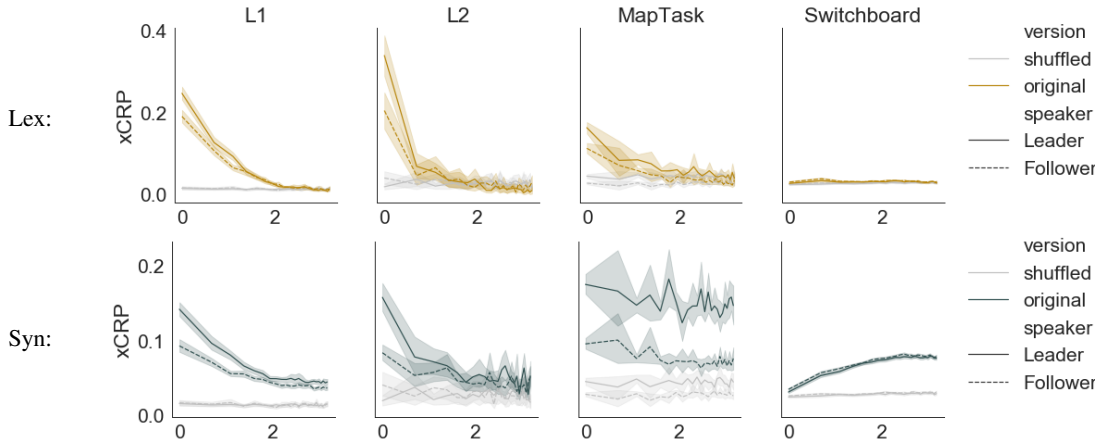


Figure 5: Distance effects on cross-speaker construction repetition proportion (xCRP) by speaker role. Log-scaled distance in utterances in the x axis. The plots distinguish between speaker roles: Leader (solid lines) quantifies situations where the adult, tutor, instruction giver, or first speaker repeats a construction used by the child, learner, instruction follower, or second speaker, respectively per corpora. And vice versa for Follower (dotted lines).

dictors.[5] We find a significant effect of level and a significant interaction of level and distance for both shared lexical and morphosyntactic constructions in L1 and for lexical constructions in L2. In particular, both xCRP and the effect of distance on xCRP decrease as child and student ability increases, with a substantially stronger effect size in L2. However, there is no effect of level nor interaction between level and distance regarding use of shared morphosyntactic constructions in L2.

In sum, our results provide evidence of a strong local effect on shared construction repetition in learner dialogues, with smaller effects in task-oriented and no or opposite effects in conversational fluent adult dialogue. The locality effect becomes weaker with ability level. This is in line with our hypothesis that priming and grounding moves may be more prominent in terms of shared construction use in L1 and L2 acquisition dialogue, while routinisation (which should be less affected by locality effects) develops as learning progresses. Surprisingly, however, we do not find a weakening of the distance effect with increased ability level regarding shared morphosyntactic constructions in L2. Thus, while the use of this type of shared construction certainly changes over time, we do not see a significant decrease in the importance of locality.

[5] As in Section 5, in CHILDES level corresponds to the age of the child in months, while in BELC it is captured by the instruction level; in both cases the level predictor is numerical.

7 Conclusion

We have investigated cross-speaker repetition of multi-word constructions at the lexical and morphosyntactic level in both L1 and L2 acquisition setups, contrasting them with adult native dialogue. Our results demonstrate that shared constructions form an important aspect of dialogue, both learner and fluent. We show that language acquisition scenarios are characterised by richer inventories of shared constructions, and that their use evolves as learner linguistic ability increases. In language learning setups, particularly in L2 learning, such constructions are used at lower rates overall, but with higher local repetition than in fluent adult dialogues. This trend decreases with learner ability. We interpret this change as an increase in construction routinisation, which we take to be present in fluent adult dialogue: Constructions become more established as part of the learner's own repertoire, thus requiring less reliance on the interlocutor's language use and less local confirmation by the dialogue partner.

Besides contributing to a better understanding of alignment patterns in language learning scenarios, our empirical results are relevant for the development of more natural and effective tutoring dialogue agents. For example, monitoring the level of learner ability in terms of degree of routinisation could help make decisions on the need to increase or decrease the amount of support provided by the tutoring agent.

Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Jens Allwood and Elisabeth Ahlsén. 1986. Lexical convergence and language acquisition. In *Papers from the Ninth Scandinavian Conference of Linguistics, University of Stockholm: Dept of Linguistics*, pages 15–26.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Colin Bannard and Elena Lieven. 2009. Repetition and reuse in child language learning. *Formulaic language*, 2:299–321.

Lasse Bergroth, Harri Hakonen, and Timo Raita. 2000. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE.

S. Brennan and H. Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.

Peter Broeder. 1992. Learning to repeat to interact: learner's repetitions in the language acquisition process of adults. *Journal of Intercultural Studies*, 13(2):19–35.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Michelle M Chouinard and Eve V Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–670.

E. V. Clark and J. Bernicot. 2008. Repetition as ratification: How parents and children place information in common ground. *Journal of Child Language*, 35(2):349–371.

H. Clark. 1996. *Using language*. CUP.

Vivian J Cook. 1973. The comparison of language development in native children and foreign adults. *IRAL-International Review of Applied Linguistics in Language Teaching*, 11(1-4):13–28.

Vivian J Cook. 2010. The relationship between first and second language acquisition revisited. *The Continuum companion to second language acquisition*, pages 137–157.

Albert Costa, Martin J Pickering, and Antonella Sorace. 2008. Alignment in second language dialogue. *Language and cognitive processes*, 23(4):528–556.

R. Dale and M.J. Spivey. 2005. Categorical recurrence analysis of child language. In *Proc. CogSci*, pages 530–535.

R. Dale and M.J. Spivey. 2006. Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3):391–430.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Joseph Denby and Dan Yurovsky. 2019. Parents' linguistic alignment predicts children's language development. In *Proceedings of CogSci*, pages 1627–1632.

Holger Diessel. 2013. Construction grammar and first language acquisition. *The Oxford handbook of construction grammar*, pages 347–364.

John W Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410.

Guillaume Dubuisson Duplessis, Franck Charras, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset. 2017a. Utterance retrieval based on recurrent surface text patterns. In *European Conference on Information Retrieval*, pages 199–211. Springer.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017b. Automatic measures to characterise verbal alignment in human-agent interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*.

Nick Ellis. 2013. Construction grammar and second language acquisition. In *The Oxford handbook of construction grammar*. Oxford University Press New York.

Raquel Fernández and Robert Grimm. 2014. Quantifying categorical and conceptual convergence in child-adult dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. *arXiv preprint arXiv:1909.05361*.

J. Gerard, F. Keller, and T. Palpanas. 2010. Corpus evidence for age effects on priming in child language. In *Proc. CogSci*, pages 1559–1564.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612–618.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6).

Daniel S Hirschberg. 1977. Algorithms for the longest common subsequence problem. *Journal of the ACM (JACM)*, 24(4):664–675.

Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 111–118.

Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35(2):133–153.

Christine Howes, Patrcik GT Healey, and Matthew Purver. 2010. Tracking lexical and syntactic alignment in conversation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. 2016. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 151–164. Springer.

J. Huttenlocher, M. Vasilyeva, and P. Shimpi. 2004. Syntactic priming in young children. *Journal of Memory and Language*, 50(2):182–195.

D Jurafsky, E Shriberg, and D Biasca. 1997. Switchboard dialog act corpus. *International Computer Science Inst. Berkeley CA, Tech. Rep*.

Sandra Katz, Pamela Jordan, and Diane Litman. 2011. Rimac: A natural-language dialogue system that engages students in deep reasoning dialogues about physics. *Society for Research on Educational Effectiveness*.

Diane Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

B. MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*, 3 edition. Lawrence Erlbaum Associates.

Darlene McDonough et al. 2013. Similarities and differences between adult and child learners as participants in the natural learning process. *Psychology*, 4(03):345.

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.

Carmen Muñoz. 2006. *Age and the rate of foreign language learning*, volume 19. Multilingual Matters.

Bill Noble and Raquel Fernández. 2015. Centre stage: How social network position shapes linguistic coordination. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 29–38.

Matthew Brook O'Donnell, Ute Römer, and Nick C Ellis. 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, 18(1):83–108.

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Martin J Pickering and Simon Garrod. 2005. Establishing and using routines during dialogue: Implications for psychology and linguistics. *Twenty-first century psycholinguistics: Four cornerstones*, pages 85–101.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: an integrative framework. *PsyArXiv, 24*.

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 121–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Reitter, Frank Keller, and Johanna D Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive science*, 35(4):587–637.

Arabella Sinclair, Rafael Ferreira, Adam Lopez, CG Lucas, and Dragan Gasevic. 2019a. I wanna talk like you: Speaker adaptation to dialogue style in l2 practice conversation. In *Proceedings of Artificial Intelligence in Education - 20th International Conference*.

Arabella Sinclair, Adam Lopez, Christopher Lucas, and Dragan Gasevic. 2018. Does ability affect alignment in second language tutorial dialogue? In *19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2018)*.

Arabella Sinclair, Kate McCurdy, Christopher G Lucas, Adam Lopez, and Dragan Gaševic. 2019b. Tutorbot corpus: Evidence of human-agent verbal alignment in second language learner dialogues. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, pages 414–419. ERIC.

Natalie B Steinhauser, Gwendolyn E Campbell, Leanne S Taylor, Simon Caine, Charlie Scott, Myroslava O Dzikovska, and Johanna D Moore. 2011. Talk like an electrician: Student dialogue mimicking behavior in an intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 361–368. Springer.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

David Wood. 2002. Formulaic language acquisition and production: Implications for teaching. *TESL Canada Journal*, pages 01–15.

A Effect of locality on cross-speaker construction repetition

In this section we provide the full outputs from the models described in the main paper. The variables mentioned are the following:

- *xCRP*: the dependent variable for all models, namely the *cross-speaker construction repetition proportion*.

V *version*: indicates whether the dialogues are the original version, or a scrambled baseline where the order of the utterances are randomly reindexed, maintaining the turn taking order of the speakers. The variables are either ORIG or BASE, for original or shuffled baseline.

D *ln_dist*: log-transformed distance in utterances.

S *speaker*: indicates which interlocutor utters the shared construction, S1 represents the Lead speaker i.e. the Adult, Tutor or instruction follower, and S1 represents the Follower speaker: Child, Student or instruction follower. In Switchboard where the speakers have equal roles, S1 is whichever speaker speaks first in the dialogue.

L *level*: indicates the Child and Students' relative competence which is measured by either the child's age in months, or one of 4 ability level brackets for the L2 student.

A.1 Baseline *(V)*

For the baseline *V* indicates whether the dialogue is the scrambled baseline or in the original order.

Lexical - $xCRP \sim D * V$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.5383	0.263	-13.446	0.000	-4.054	-3.023
V[T.orig]	2.2902	0.326	7.029	0.000	1.652	2.929
D	-0.0299	0.110	-0.273	0.785	-0.245	0.185
D:V[T.orig]	-1.0875	0.157	-6.923	0.000	-1.395	-0.780

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-4.2741	0.185	-23.154	0.000	-4.636	-3.912
V[T.orig]	2.9675	0.209	14.176	0.000	2.557	3.378
D	-0.0103	0.076	-0.134	0.893	-0.160	0.140
D:V[T.orig]	-1.0880	0.095	-11.397	0.000	-1.275	-0.901

MapTask	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.4445	0.236	-14.624	0.000	-3.906	-2.983
V[T.orig]	1.3673	0.319	4.283	0.000	0.742	1.993
D	0.0293	0.097	0.302	0.762	-0.161	0.219
D:V[T.orig]	-0.4681	0.138	-3.385	0.001	-0.739	-0.197

Switchboard	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.6511	0.081	-45.146	0.000	-3.810	-3.493
V[T.orig]	0.2225	0.134	1.661	0.097	-0.040	0.485
D	0.0510	0.033	1.543	0.123	-0.014	0.116
D:V[T.orig]	-0.0532	0.055	-0.964	0.335	-0.161	0.055

Morphosyntactic - $xCRP \sim D * V$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.5383	0.263	-13.446	0.000	-4.054	-3.023
V[T.orig]	1.2272	0.354	3.465	0.001	0.533	1.921
D	-0.0299	0.110	-0.273	0.785	-0.245	0.185
D:V[T.orig]	-0.2669	0.151	-1.762	0.078	-0.564	0.030

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-4.2741	0.185	-23.154	0.000	-4.636	-3.912
V[T.orig]	2.0452	0.219	9.359	0.000	1.617	2.474
D	-0.0103	0.076	-0.134	0.893	-0.160	0.140
D:V[T.orig]	-0.3417	0.093	-3.691	0.000	-0.523	-0.160

MapTask	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.4445	0.236	-14.624	0.000	-3.906	-2.983
V[T.orig]	1.4421	0.302	4.778	0.000	0.851	2.034
D	0.0293	0.097	0.302	0.762	-0.161	0.219
D:V[T.orig]	-0.1352	0.125	-1.078	0.281	-0.381	0.111

Switchboard	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.6511	0.081	-45.146	0.000	-3.810	-3.493
V[T.orig]	0.6338	0.115	5.531	0.000	0.409	0.858
D	0.0510	0.033	1.543	0.123	-0.014	0.116
D:V[T.orig]	0.1432	0.046	3.089	0.002	0.052	0.234

A.2 Distance *(D)*

Models are only fitted on the original version of the data.

Lexical - $xCRP \sim D$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-1.2481	0.192	-6.495	0.000	-1.625	-0.871
D	-1.1174	0.113	-9.925	0.000	-1.338	-0.897

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-1.3066	0.099	-13.237	0.000	-1.500	-1.113
D	-1.0982	0.057	-19.202	0.000	-1.210	-0.986

MapTask	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.0772	0.215	-9.641	0.000	-2.499	-1.655
D	-0.4388	0.099	-4.448	0.000	-0.632	-0.245

Switchboard	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.4286	0.107	-32.085	0.000	-3.638	-3.219
D	-0.0021	0.044	-0.049	0.961	-0.089	0.084

Morphosyntactic - $xCRP \sim D$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.3112	0.237	-9.752	0.000	-2.776	-1.847
D	-0.2968	0.105	-2.838	0.005	-0.502	-0.092

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.2289	0.117	-19.054	0.000	-2.458	-2.000
D	-0.3520	0.052	-6.738	0.000	-0.454	-0.250

MapTask	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.0024	0.189	-10.613	0.000	-2.372	-1.633
D	-0.1059	0.080	-1.331	0.183	-0.262	0.050

Switchboard	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.0174	0.081	-37.165	0.000	-3.176	-2.858
D	0.1943	0.033	5.976	0.000	0.131	0.258

A.3 Speaker Role *(S)*

Models are only fitted on the original version of the data.

Lexical - $xCRP \sim D * S$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-1.7343	0.206	-8.428	0.000	-2.138	-1.331
S[T.S2]	0.6978	0.271	2.574	0.010	0.167	1.229
D	-0.8325	0.108	-7.733	0.000	-1.044	-0.622
D:S[T.S2]	-0.5032	0.157	-3.215	0.001	-0.810	-0.196

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-1.5105	0.098	-15.347	0.000	-1.703	-1.318
S[T.S2]	0.3424	0.134	2.554	0.011	0.080	0.605
D	-0.9767	0.054	-18.026	0.000	-1.083	-0.870
D:S[T.S2]	-0.1742	0.076	-2.289	0.022	-0.323	-0.025

MapTask	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.2595	0.222	-10.186	0.000	-2.694	-1.825
S[T.S2]	0.3952	0.293	1.350	0.177	-0.179	0.969
D	-0.4413	0.101	-4.372	0.000	-0.639	-0.243
D:S[T.S2]	0.0148	0.132	0.112	0.911	-0.245	0.274

Switchboard	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-3.3461	0.101	-33.120	0.000	-3.544	-3.148
S[T.S2]	-0.0862	0.145	-0.593	0.553	-0.371	0.199
D	-0.0232	0.042	-0.553	0.580	-0.105	0.059
D:S[T.S2]	0.0177	0.060	0.294	0.769	-0.100	0.136

Morphosyntactic - $xCRP \sim D * S$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.5268	0.235	-10.775	0.000	-2.986	-2.067
S[T.S2]	0.4649	0.309	1.505	0.132	-0.140	1.070
D	-0.2386	0.102	-2.347	0.019	-0.438	-0.039
D:S[T.S2]	-0.1212	0.135	-0.894	0.371	-0.387	0.144

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.4103	0.118	-20.504	0.000	-2.641	-2.180
S[T.S2]	0.4160	0.155	2.676	0.007	0.111	0.721
D	-0.3010	0.052	-5.818	0.000	-0.402	-0.200
D:S[T.S2]	-0.0891	0.069	-1.292	0.197	-0.224	0.046

MapTask	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.3478	0.201	-11.666	0.000	-2.742	-1.953
S[T.S2]	0.7373	0.253	2.910	0.004	0.241	1.234
D	-0.0864	0.084	-1.023	0.306	-0.252	0.079
D:S[T.S2]	0.0217	0.106	0.205	0.838	-0.186	0.229

Switchboard	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-2.9305	0.077	-38.228	0.000	-3.081	-2.780
S[T.S2]	-0.0784	0.110	-0.714	0.475	-0.294	0.137
D	0.1758	0.031	5.716	0.000	0.116	0.236
D:S[T.S2]	0.0190	0.044	0.432	0.666	-0.067	0.105

A.4 Level *(L)*

Models are only fitted on the original version of the data.

Lexical - $xCRP \sim D * L$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-0.3086	0.324	-0.953	0.341	-0.943	0.326
D	-1.6166	0.194	-8.331	0.000	-1.997	-1.236
L	-0.4839	0.151	-3.194	0.001	-0.781	-0.187
D:L	0.2457	0.083	2.961	0.003	0.083	0.408

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	0.5368	0.265	2.025	0.043	0.017	1.056
D	-1.7210	0.153	-11.263	0.000	-2.020	-1.421
L	-0.0478	0.007	-6.968	0.000	-0.061	-0.034
D:L	0.0166	0.004	4.342	0.000	0.009	0.024

Morphosyntactic - $xCRP \sim D * L$

BELC	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-1.8653	0.382	-4.886	0.000	-2.614	-1.117
D	-0.5268	0.171	-3.082	0.002	-0.862	-0.192
L	-0.1771	0.167	-1.059	0.290	-0.505	0.151
D:L	0.0994	0.073	1.360	0.174	-0.044	0.243

Childes	coef	stderr	z	P>|z|	[0.025	0.975]
Intercept	-0.9576	0.301	-3.182	0.001	-1.547	-0.368
D	-0.7814	0.136	-5.763	0.000	-1.047	-0.516
L	-0.0312	0.008	-4.080	0.000	-0.046	-0.016
D:L	0.0109	0.003	3.202	0.001	0.004	0.018

Poster Abstracts

Annotating Events and Entities in Dialogue

Tatiana Anikina, Ivana Kruijff-Korbayová
DFKI / Saarland Informatics Campus, Saarbrücken, Germany
tatiana.anikinal@dfki.de ivana.kruijff@dfki.de

1 Introduction

We present the EveEnti (Event and Entity) annotation framework for events and entities in dialogue that we use to annotate several dialogues in German from the emergency response domain (Willms et al., 2019).

Events and entities are crucial for natural language understanding but the research on modeling them in dialogue has been limited due to the lack of annotated resources. There exist several corpora but they are mostly specialized and cover only a subset of possible annotations. For instance, the relations between events were annotated with respect to their temporal structure (Minard et al., 2016), causal dependencies (Mirza et al., 2014) or coreference links (Lee et al., 2012). Some work has been done towards the unification of different annotations (Mostafazadeh et al., 2016; O'Gorman et al., 2016). However, these corpora are based on standard narrative texts and do not include dialogues. Moreover, none of them have complete annotations for events, entities and discourse relations at the same time. For instance, the ARRAU corpus (Poesio et al., 2018) has rich entity annotations and includes dialogues but does not provide detailed information about events. On the other hand, the RED corpus (O'Gorman et al., 2016) has good coverage of event annotations but does not include fine-grained entity types and rhetorical relations at the document level.

Motivated by the fact that there is no unified framework for annotating events and entities in dialogue we decided to develop an annotation scheme and tool that will capture various semantic aspects and also maintain the relations between the annotation layers. We tried to use the established annotation standards and guidelines as much as possible to keep compatibility with other corpora but had to introduce some adjustments as well.

2 EveEnti Annotation Framework

Our dataset currently consists of the dialogues in German collected during several disaster response training sessions (Kruijff-Korbayova et al., 2015; Willms et al., 2019). These dialogues represent team communication between the team leader and several operators who remotely operate robots in order to explore some area, find hazardous materials, locate fire, damage or victims. In total, there are 2398 transcribed dialogue turns in our corpus. Additionally, we have 818 dialogue turns in English that come from the same domain and we are planning to add these data in the next round of annotation.

The EveEnti annotation scheme was designed to capture all events and entities in a dialogue and annotate the relations between them in such a way that allows to analyze different layers of annotation both independently and in combination. All events and entities in a document can be seen as nodes in a graph that are annotated with various semantic features and the relations between them are edges. The resulting graph can represent the unfolding discourse where events have temporal order and rhetorical relations and entities have thematic roles with respect to their corresponding events as well as semantic types and coreference information. Our goal was to find a flexible annotation scheme that is independent of any other/preceding levels of processing, in particular parsing. The scheme should be easy to apply to different languages and genres of text and it does not require from annotators to write complicated logical formulas or graph representations as is the case with Discourse Representation Theory (Kamp and Reyle, 1993) and Abstract Meaning Representation (Banarescu et al., 2013) annotations.

We developed an annotation tool that shows dialogue threads and turns as rows in a table. Each turn

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

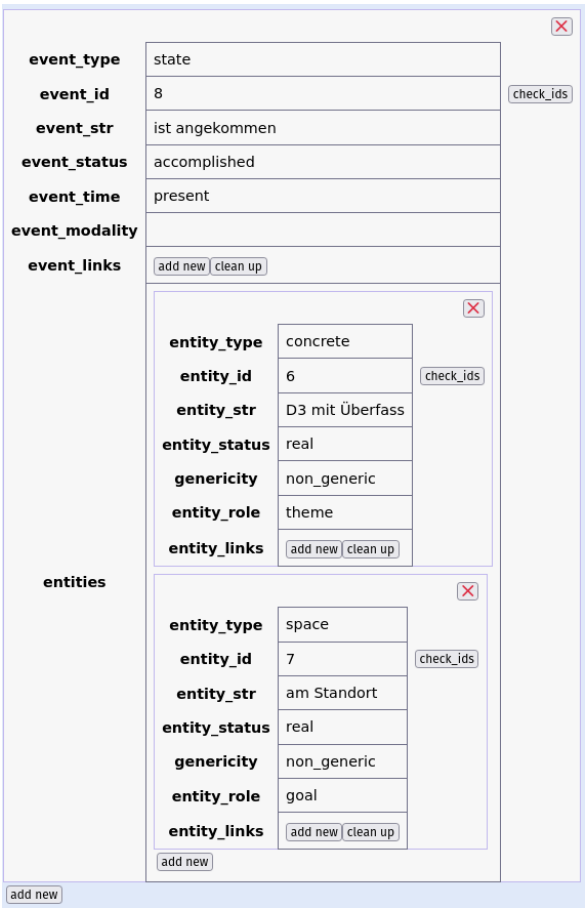


Figure 1: Event and entity annotations for *"D3 mit Überfass ist am Standort angekommen"* (D3 with a barrel arrived to the position)

has columns with a unique id, speaker, addressee, turn text, one column for the event and entity annotations and another column for annotating relations between the turns.

We annotate each predicate in a dialogue turn that can be described as a state, action, process or habitual as an event. Several events can be annotated for each turn and adding a new event simply means creating a box with the following fields: event type, id, string, status, time, modality and links (see Figure 1). We distinguish between accomplished and non-accomplished events ("status"), past/present/future events ("time") and annotate negation, necessity and possibility ("modality"). In the field "event_links" annotators can link events to each other using their ids and annotate the relations such as e.g., cause or condition defined in the ISO DR-Core (ISO 24617-8) scheme developed in (Prasad and Bunt, 2015). We added an "argument" relation to account for the nested cases when an event has another event as an argument, e.g., *"nicht funktioniert"* (doesn't work) in *"ich glaube, dass es nicht funktioniert"* (I believe that it doesn't work). To annotate temporal order of linked events, we use the TimeML-inspired relations that were proposed in (Mostafazadeh et al., 2016) and distinguish between the following four categories: before, overlaps, contains and identity.

Additionally, each event can have a list of entities associated with it. Each entity has its own box that includes annotations for its semantic type, id, string, status, genericity and semantic role. Semantic types include eight common categories: abstract, concrete, animate, person, organization, time, date and space. Entity status defines whether it is a real or assumed (imagined) entity. We annotate thematic role with respect to the associated event, e.g., agent or theme. Entities can also have links to each other that are either captured via coreference chains with the same ids or annotated in the "entity_links" field. We use the following relations for non-standard anaphora: set/member, part/whole and bridging. These relations were also used in the RED annotation scheme (O'Gorman et al., 2016).

Because we want to model relations at both levels: individual events and entities and complete dialogue turns, we include a separate column "function" and assign to each turn the most likely communicative function based on the following categories: call, call response, feedback positive or negative, question, answer, request, request response and task inform. These categories represent a simplified version of the functions proposed in the ISO standard (Bunt et al., 2012, 2020). Moreover, EveEnti has a separate column for annotating dependencies between the paired turns such as question/answer or request/request_response.

3 Conclusion

We present the first comprehensive framework for event and entity annotation in dialogue. It allows us to annotate events and entities jointly and use a variety of annotation layers that include semantic type, role and status for entities, temporal order and modality for events as well as coreference chains and rhetorical relations. To our knowledge, none of the existing corpora provides all of these annotations at the same time. While annotating events and entities we focus on dialogue and also annotate communicative functions and relations between the turns. Although the annotation process is ongoing we are planning to present our inter-annotator agreement results at the time of the conference.

Acknowledgments

The authors have been supported by the German Ministry of Education and Research (BMBF) through project CORA4NLP (grant Nr. 01IW20010). We would like to thank K. J. Christian, N. Skachkova and H. Düe for their help with annotations and A. González-Palomo for technical support and help in setting up the annotation tool.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 178–186.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Kôiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 430–437.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Chengyu Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 549–558.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic - Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in linguistics and philosophy*. Springer.

Ivana Kruijff-Korbayova, Francis Colas, Mario Gianni, Fiora Pirri, Joachim Greeff, Koen Hindriks, Mark Neerincx, Petter Ogren, Tomáš Svoboda, and Rainer Worst. 2015. Tradr project: Long-term human-robot teaming for robot assisted disaster response. *KI - Künstliche Intelligenz*, 29.

Heeyoung Lee, Marta Recasens, Angel X. Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 489–500.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James F. Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events, EVENTS@HLT-NAACL 2016, San Diego, California, USA, June 17, 2016*, pages 51–61.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Christian Willms, Constantin Houy, Jana-Rebecca Rehse, Peter Fettke, and Ivana Kruijff-Korbayová. 2019. Team communication processing and process analytics for supporting robot-assisted emergency response. In *IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2019, Würzburg, Germany, September 2-4, 2019*, pages 216–221.

What do you mean?
Eliciting enthymemes in text-based dialogue

Ebba Axelsson Nord, Vladislav Maraev, Ellen Breitholtz and Christine Howes
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
ebbaxelsson@hotmail.com; vladislav.maraev@gu.se;
ellen.breitholtz@ling.gu.se; christine.howes@gu.se

Abstract

We report a proof of concept text-based chat study which inserts spoof questions in a real time conversation in order to elicit participants reasoning. The results show that different questions are more or less effective at eliciting explicit enthymematic reasons, which is a technique that could be employed to augment conversational artificial intelligence systems to improve their reasoning abilities.

1 Introduction

Reasoning is crucial for humans and for dialogue agents, and much research has been dedicated to enabling computers to reason from premises to conclusions. Nevertheless, how people interactively reason in natural dialogue is still poorly understood, since much reasoning in human dialogue is enthymematic, i e it relies on non-logical common sense principles of reasoning (Breitholtz, 2020). No existing artificial intelligence system is able to make use of this type of reasoning, which people find so natural.

Besides having little knowledge about reasoning in natural dialogue for building better dialogue agents, we also lack adequate data resources, especially given that important reasoning often takes place "behind the scenes", and can't be extracted directly from dialogue transcripts. This work can be viewed as a step towards collecting data for dialogue agents which are capable of reasoning.

Specifically we present a proof of concept chat tool study to investigate whether people can be prompted to provide their reasoning in an unobtrusive way. As noted by (Schlöder et al., 2016), one way of probing enthymematic reasoning is through the use of questions like "why?" or "what do you mean?".

2 Method

2.1 DiET chat tool

The Dialogue Experimental Toolkit (DiET) chat tool (Healey et al., 2003) is a text-based chat interface into which interventions, such as adding fake turns, can be introduced into a dialogue in real time, thus causing a minimum of disruption to the 'flow' of the conversation. For this experiment we used the new mobile version of DiET, which runs through the messenger app Telegram.[1]

2.2 Task

The subjects discussed the balloon task – a moral dilemma known to elicit dialogues containing extended reasoning sequences. Participants are instructed to reach agreement on which of four passengers should be thrown out of a hot air balloon that will otherwise crash, killing all the passengers, if one is not sacrificed. The four passengers are:

Mr Tom Harris – the balloon pilot who is the only passenger with any balloon flying experience
Mrs Susie Harris – Tom's wife, a primary school teacher who is 7 months pregnant with their second child
Dr Robert Lewis – a cancer research scientist, who believes he is on the brink of discovering a cure for most common types of cancer
Miss Heather Sloan – a 14 year old musical prodigy who is considered to be a "twenty-first century Mozart"

2.3 Procedure

The 32 participants, from the student population at the University of Gothenburg, were instructed via zoom on how to access the experiment using the Telegram app. Once logged in, they were told to discuss the task until they got a message from the server to stop, in order to ensure sufficient turns. The manipulation consisted of introducing 'spoof' turns into the dialogue, which appeared to the recipient to have originated from their dialogue partner.

[1] https://dialoguetoolkit.github.io/chattool/

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

The spoof turns, consisting of questions such as 'why?' and 'what do you mean?' (see table 1) were generated pseudo randomly from a limited list of probes by the DiET-server and triggered by particular words or phrases in the preceding turn (e.g., "the doctor", "the pregnant woman").

2.4 Annotation

The dialogues were examined to establish whether or not the spoof question was responded to and if so, whether the response was a direct response (1), an indirect response (if there was an intervening contribution before the response, as in (2)) or a clarification request (3).

(1) **1:** I bet the father would not want the child nor their partner to die
2: **why?** *[artificial turn]*
1: Paternal instincts and all that

(2) **1:** But is it morally acceptable to throw out the girl if the pilot is needed?
2: **what do you mean?** *[artificial turn]*
2: The child you mean?
1: yeah

(3) **2:** Doctor can have botanical knowledge or whatever
1: **why?** *[artificial turn]*
2: Why which part

3 Results and discussion

With the exception of one pair, who were discarded from further analysis, the debriefing showed that none of the subjects were aware of any experimental manipulations. Each pair was exposed to an average of 6.4 spoof questions generated by the server.

Spoof question	NR	DR	IR	CR	Total
???	4	6	1	0	11
how so?	4	6	3	0	13
what?	4	6	4	0	14
what do you mean?	5	16	14	2	37
why?	4	14	6	3	27
Total	21	48	28	5	102

Table 1: Type of response given by probe. NR/DR/IR – no/direct/indirect response. CR – clarification request.

A greater proportion of "???", "how so?" and "what?" questions received no response compared to "why?" and "what do you mean?" (32% vs 14% $\chi^2_1 = 4.47; p = 0.03$). One possible explanation for this finding is that in the former cases the participants may think that the ongoing dialogue shows that the question is already resolved so they do not feel obliged to answer. This could also be affected by the text based medium in which participants can type simultaneously with turns both interleaved and persistent (Healey et al., 2018).

Qualitative analysis of the data by the first author suggests that "why" and "how so" spoof questions are more capable of addressing reasoning (as in (1)) with the other three cues more open ended and available to interpretations ranging from semantic or orthographic ambiguities to reasoning. Nevertheless, each of the probes, even the more open ended ones, did produce some responses which, combined with a previous utterance, constitute enthymemes, showing that the interpretation of a question is not fixed. An example of this can be seen in (4).

(4) **1:** I think pregnant women are not supposed to fly actually
2: **what?** *[artificial turn]*
1: There are safety regulations at least during the 7th month of pregnancy

This study shows that our approach is a useful method for eliciting enthymemes to collect resources for common sense reasoning in spoken dialogue systems. This is useful in task-oriented domains to argue about the decisions taken by the system, as well as in chit-chat dialogues – especially ones which are concerned with controversial topics or current issues, such as the climate crisis.

In the process of dialogue systems development enthymeme eliciatation can be a part of data collection based on a *distilling dialogue* process (Larsson et al., 2000; Jönsson and Dahlbäck, 2000). After enthymeme elicitation, we plan to take the following steps to collect enthymematic resources for a dialogue system:

1. dependency parsing and pattern-based extraction of enthymeme candidates based on their surface structure
2. annotation, whether or not the extracted structure is an enthymeme and annotation of the premise(s) and the consequence(s) of it.
3. enthymeme classification (for example, keywords like 'since' can relate to a time frame)
4. enthymeme parsing, that will lead to a semantic representation of an enthymeme

Extracted enthymemes can then be clustered to induce more general principles of reasoning, such as the Aristotelian topos of 'the more and the less'. The gist of this topos is that a small thing is contained in a large thing – for example, if you can build a castle you can build a cottage, or if you can run a marathon then you can run a half marathon.

Acknowledgments

This work was supported by the Incremental Reasoning in Dialogue (IncReD) project funded by the Swedish Research Council (VR) (2016-01162). Maraev, Breitholtz and Howes were additionally supported by VR grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP).

References

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

Patrick G. T. Healey, Gregory Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science*, 10(2):367–388.

Patrick G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Gregory Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, Massachusetts.

Arne Jönsson and Nils Dahlbäck. 2000. Distilling dialogues-a method using natural dialogue corpora for dialogue systems development. In *6th Applied Natural Language Processing Conference, Seattle, Washington, 2000*, pages 44–51. Association for Computational Linguistics Stroudsburg.

Staffan Larsson, Lena Santamarta, and Arne Jönsson. 2000. Using the process of distilling dialogues to understand dialogue systems. In *Proceedings of ICSLP 2000*, pages 374–377. Citeseer.

Julian Schlöder, Ellen Breitholtz, and Raquel Fernández. 2016. Why? In *Proceedings of JerSem*, pages 5–14.

The Deictic Nature of Speech Act Reference

Friederike Buch
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
Schützenstr. 18
10117 Berlin, Germany
buch@leibniz-zas.de

Motivation. I conducted two experiments to access the ontological status of speech acts in discourse in German. If speech acts are part of the utterance situation, deixis to speech acts should be possible, but anaphora, which is restricted to entities introduced in the discourse, should not.

Experiment 1. In the first, exploratory online experiment (30 participants, 2×12 items), it was tested whether pronominal reference to speech acts could be elicited from native speakers in German. They were presented with short contexts with two interlocutors and were asked to complete the last utterance whose beginning either involved the demonstrative *das* or the personal pronoun *es*. The results were annotated with respect to the referent of the pronoun (SPA_ILL = illocutionary act, EVT = eventuality, PROP = proposition), but I will focus on speech acts and events here. The example in (1) shows a *das* variant and a continuation that indicates a speech act reference. The relative frequencies of some of the referent types intended by the participants for each of the two pronouns are given in Fig. (1).

(1) *Niklas hat gerade sein Abitur mit*
Niklas has just his high-school-diploma with
Bravur bestanden. Seine große Schwester
flying-colors passed his older sister
Lara freut sich für ihn.
Lara is-happy refl for him
'Niklas hast just passed his high-school diploma with flying colors. His older sister Lara is happy for him.'

Lara sagt zu Niklas: „Herzlichen
Lara says to Niklas cordial
Glückwunsch!"
congratulation
'Lara says to Niklas, "Congratulations!"'

Was könnte Niklas darauf sagen?
What could Niklas there-on say
'What could Niklas answer to this?'

*Niklas erwidert: „Das ist . . . **lieb von dir,***
Niklas replies that is nice of you
***danke.**"* [SPA_ILL]
thanks
'Niklas replies, "That is **nice, thank you.**"'

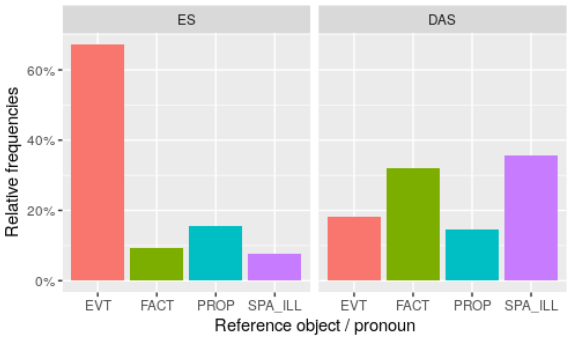


Figure 1: Naturalness of pronouns

This pilot study was not primarily designed to test a hypothesis, but rather to elicit speech act references from speakers and to learn which of the pronouns *es* and *das* are suitable expressions for this purpose. Another open question concerned the types of referents the participants would make the pronouns refer to.

The central hypotheses in Exp. 1 concern the comparison between illocutionary acts as representatives of situational events on the one hand and events that are introduced linguistically on the other. The personal pronoun *es* cannot refer to situational events, so *es* should prefer reference to linguistically introduced events. *Das* on the other hand should prefer reference to speech acts, since eventualities that have recently been introduced linguistically are salient and should therefore rather be referred to by personal pronouns than demonstratives. Additionally, independently from the individual predictions of the two pronouns, there is an overall effect of pronoun choice on the choice between SPA_ILL and EVT.

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

I fit two logistic regression models in R on the data from Exp. 1. One test compared the frequencies of speech acts with the frequencies of other eventualities depending on the given pronoun, and the other one compared propositions with facts. The first model revealed that, given the pronoun *es* and regarding only the SPA_ILL and EVT cases, EVT was strongly preferred over SPA_ILL with a proportion of 89.6 % ($p = 5.26 * 10^{-6}, SE = 0.472$). Given the pronoun *das*, SPA_ILL was chosen in 66.2 % of the time ($p = 0.00618, SE = 0.246$). There was a difference between *es* and *das* in the choice of referent ($p = 1.13 * 10^{-7}, SE = 0.533$).

Experiment 2. The pilot study did not disentangle two factors in the answering strategies of the participants: a) which pronoun can refer to what kinds of entities, and b) what entity the participant intuitively wants to make a statement about, given the context. This entanglement may lead to situations were the given pronoun is dispreferred for the favored referent.

The first experiment tested the naturalness of pronominal speech act reference. In the second experiment (99 participants, 18 items), the preferred choice for either personal or demonstrative pronouns was investigated, dependent on a given type of referent. To test which pronoun is preferred for reference to what kind of entity, we have to reverse the study design of Exp. 1. The referent must be given, and a choice of pronoun must be offered.

The 12 items from Exp. 1 were reused in such a way that two items each were accompanied with a continuation that favored reference to one of the six referent types, including illocutionary acts (SPA_ILL) and non-speech-act eventualities (EVT). The continuations were inspired by the participant continuations from Exp. 1 to ensure naturalness. The participants were given the choice between the demonstrative *das* and the personal pronoun *es* as the first position in the sentences.

(2) [. . .] *Niklas erwidert: „{ Es / Das } ist lieb*
Niklas replies it that is nice
von dir, danke.“ [SPA_ILL]
of you thanks
‘Niklas replies, “{ It / that } is nice of you, thank you.”’

As the predictor variable in Exp. 2 is a multilevel categorical variable, we can formulate a general hypothesis of significance, as well as predictions for the direction of influence for each of the levels, i.e. for the different types of referents.

I assume a preference for personal pronouns for reference to eventualities. Eventualities are present as available referents in the discourse, provided that they have been introduced linguistically. These entities should be salient and ready for uptake by, preferably, personal pronouns. Anaphora with a demonstrative pronoun is possible if the referent is not salient. Additionally, there should be an overall effect of referent type on pronoun choice.

Personal pronouns were used very rarely to refer to speech acts (96% demonstratives), while verbally introduced events were referred to by both demonstrative (44%) and personal pronouns. The relative frequencies of the two pronouns, for some of the given referents, are displayed in Fig. (2).

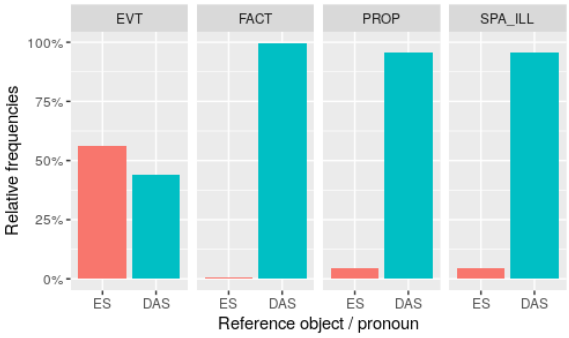


Figure 2: Preference for pronouns

Binomial logistic regressions tested the effect of each level of the predictor (type of referent) on the choice of pronoun. I again focus here on speech acts and events. There was no significant effect for eventualities ($p = 0.106$), but for speech acts ($p = 1.22 * 10^{-17}$), which showed a strong preference for the demonstrative *das*. Additionally, I used the logistic models to test the significance of the difference in the influence on pronoun choice for the level pair EVT/SPA_ILL ($p = 1.54 * 10^{-17}$). Speech act reference had a much higher proportion of demonstratives than event reference. Finally, I performed an χ^2 test as an omnibus test on the predictor, which proves significant with $p = 2.2 * 10^{-16}$.

Conclusion. Reference to speech acts is deictic, as German native speakers a) do not choose speech acts as referents of personal pronouns, but only demonstrative pronouns, and b) choose demonstratives in order to refer to speech acts. I will additionally present on the findings regarding other referent types such as facts and propositions, which are not discussed in this abstract.

References

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

Peter Bosch, Graham Katz, and Carla Umbach. 2007. The non-subject bias of german demonstrative pronouns. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Anaphors in text: cognitive, formal and applied approaches to anaphoric reference*, pages 145–164.

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

Werner Frey. 2012. On two types of adverbial clauses allowing root-phenomena. In Lobke Aelbrecht, Liliane Haegeman, and Rachel Nye, editors, *Main Clause Phenomena. New Horizons*. John Benjamin Publishing Company.

Bart Geurts and Emar Maier. 2003. Layered drt. *Semantics archive*.

Liliane Haegeman. 2010. The internal syntax of adverbial clauses. *Lingua*, 120(3):628–648.

Michael Hegarty. 2006. Type shifting of entities in discourse. In K. von Heusinger and K. Turner, editors, *Where Semantics Meets Pragmatics*, volume 16 of *Current Research in the Semantics/Pragmatics Interface*, pages 111–128. Elsevier.

Michael Hegarty, Jeanette K. Gundel, and Kaja Borthen. 2001. Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27(2–3):163–186.

Julie Hunter, Nicholas Asher, and Alex Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11.

Keir Moulton, Elizabeth Bogal-Allbritten, and Junko Shimoyama. 2020. Things we embed. Handout at BCGL 13.

Sandra A. Thompson and William C. Mann. 1987. Rhetorical structure theory: A theory of text organization. *ISI Reprint Series (ISI/RS-87-190)*.

Michael Tomasello, Malinda Carpenter, Joseph Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28:675–735.

Identity Models for Role-Play Dialogue Characters

Patricia Chaffey
Institute for Creative Technologies
University of Southern California
chaffey@ict.usc.edu

David Traum
Institute for Creative Technologies
University of Southern California
traum@ict.usc.edu

Abstract

An ongoing challenge in dialogue systems is maintaining consistent personalities and attributes throughout a conversation. To this end, our ongoing work aims to address this obstacle by implementing an 'identity model' which references personality traits from the Big 5 model and personal attributes that appear frequently in conversation. This effort builds off of previous work conducted using a Wizard-of-Oz (woz) system, which set the dialogue agents in a wildfire search-and-rescue scenario. The data gathered in this study helps identify what attributes are most important for the specific context of our system, but the approach being taken for the dialogue agents may be favorable for more general approaches to the development of robust dialogue agents.

1 Introduction

Dialogue agents are capable of communicating a variety of topics, depending on their purpose and intended scope, and in recent years, there has been a focus on increasing their likeness in communication to how humans interact with one another. A related challenge to this focus is the difficulty in presenting a persistent personality and character attributes throughout an interaction. As such, an area of research within this domain seeks to incorporate elements of personality and character attributes as a means of developing more unique and approachable characters for people to converse with. To this end, we are developing an approach for modelling identities, and working towards implementing these modelled characters in a previously developed framework relating to a wildfire search and rescue simulation developed by Chaffey et al. (2019).

One area that we are currently expanding upon is the development of what we refer to as 'identity models'. These models track personality scores as well as character attributes (e.g. a character's age, name, occupation, etc.). A character's personality and/or their attributes can be adjusted very quickly, either by changing the personality scores or by modifying the specific attributes.

2 Related Research

Models of personality and character attributes are not new concepts in the field of natural language processing. Some approaches have utilized large corpuses with aims to model consistent behavior in their chat agents (Li et al., 2016; Mazaré et al., 2018; Zhang et al., 2018). Others, such as Mairesse and Walker (2007), looked to demonstrate varying degrees of a personality trait across dialogue options. A challenge that is consistently brought up is the lack of persistence in personality and character attributes in conversations, which can be jarring to users.

Fillwock and Traum (2018) conducted an analysis on how personal attributes are shared in conversational settings. Their work identified a set of common topics that arise frequently in conversations, which provides a general set of characteristics that a dialogue agent should be able to contend with. In addition, a study conducted by Mitsuda et al. (2017) found that how a system responded to inclusions of personal attributes could be satisfactory or unsatisfactory depending on the nature of the attributes. Namely, expressing understanding towards attributes that were considered permanent was better received than attempting to refer to temporary states, which at times could feel unnatural.

There is also a wealth of information on Big 5 traits themselves, and how they relate to other facets of personality. Roccas et al. (2002) connect motivational values to traits in the Big 5, which include extraversion, agreeableness, openness to experience, conscientiousness, and neuroticism. The motivational values described by Roccas et al. (2002) help elaborate how the Big 5 traits come into play in an individual's personality.

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

3 The Identity Model

At this stage, we currently utilize the Big 5 personality model as discussed and expanded on by Roccas et al. (2002). Agents receive an array of five scores, indicating how closely they align (or don't align) with a trait from the Big 5 model. For example, an agent scoring higher in the neuroticism category may demonstrate more anxiety, and would select dialogue reflecting that anxiety. How high or low the agent is scored in other categories will influence whether or not they select dialogue reflecting their responsibility (or lack thereof), their introversion/extraversion, whether or not they are imaginative, and more. To go into more detail on the flow of the system, we utilize the sequential model via Keras API with 3 layers to rank all possible character responses to user input. Once we have this list of possible responses, we leverage the identity model to select responses that a.) fit the character's attributes (e.g. characters wit an age attribute of under 16 would not suggest they drive themselves to safety), and b.) best match their personality scores (i.e. responses are tagged with associated Big 5 traits and scores, and can be used to select responses more in line with a character's personality scores).

In addition, Fillwock and Traum (2018)'s identified common attributes are used to develop the sustained information that an agent will reference when sharing details about themselves with the person interacting with them. Currently, the personality model is used to determine how a statement is phrased, and the list of attributes assigned to a character determines which features they share with the user.

Excerpts from two transcripts have been included below. The first one demonstrates an individual who has a higher score in conscientiousness, as well as a higher score in agreeableness. As can be seen below, this individual does not argue or question too much, and readily agrees with what the user is suggesting.

USER: Hi Ross, nice to meet you. We're doing a sweep through the town to let people know a fairly large fire is headed in this direction, about a day or two out. we're asking residents to evacuate ahead of time

RESIDENT: Wow. I'd heard about it on the news, but I didn't really think it'd make it this far

USER: Yeah, unfortauntely it's looking like it will. So we need you, and anyone with you, to gather up the essential items and head south

RESIDENT: How long do you think we have to get everything together?

USER: I would say maybe a day or so, but I wouldn't recommend lingering

RESIDENT: Yeah, okay, I hear you. I'll grab what I need. Shouldn't take too long to go through the apartment

USER: That sounds great. Do you need any assistance evacuating?

RESIDENT: I think I'm all set, actually

This second excerpt is of an individual with a lower score in agreeableness, which leads them to resist the operator's suggestion they evacuate.

RESIDENT: I can't leave my condo! I've lived here for 3 years

USER: i understand how important your home is to you, but we're worried about this fire. we recoomend you gather all of your most important items and get out of the area until the danger is past

RESIDENT: I hear what you're saying, but...I don't know, leaving my home is so hard

USER: I sympathize. We're going to do everything we can to control the fire, but it will help us a lot if we know you're safe

RESIDENT: If you're well-read like me, you don't just give up on the place you love. Surely there's something I can do to keep my home safe

USER: Here's what I recommend: take a little bit of time to clear out the yard of any dead plants or debris. Then gather important things, and head out.

RESIDENT: Okay, I'll try to gather the most important things

4 Future Work and Conclusions

Efforts towards automating agents capable of displaying distinct personalities and maintaining their own persistent set of attributes are ongoing. The data collected in previous experiments with the wildfire simulation is useful for honing in on the attributes necessary for the specific scenario our system currently supports, but we aim to expand our dialogue agent's capabilities using the observations discussed by Fillwock and Traum (2018). This, alongside with the added attention to modifiable personalities, currently sets us on a promising path for more robust dialogue agents capable of demonstrating a distinctive personality while maintaining persistent attributes in a conversation. Another avenue we intend to explore is applying our model to the users themselves during conversation, by taking note of when the user shares personal information.

Acknowledgments

This research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF- 20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of the Army Research Office or the U.S. Government.

References

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 *Proceedings of the 9th Language and Technology Conference*, Poznań, Poland. LTC.

Sarah Fillwock and David Traum. 2018. Identification of personal information shared in chat-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents.

Koh Mitsuda, Ryuichiro Higashinaka, and Junji Tomita. 2017. Investigating the effect of conveying understanding results in chat-oriented dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 389–394, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801.

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? *CoRR*, abs/1801.07243.

Can rule-based chatbots outperform Neural models without pre-training in Small Data Situations?: A Preliminary Comparison of AIML and Seq2Seq

Md Mabrur Husan Dihyat, Julian Hough[1]
[1] Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
{m.dihyat@se17.qmul.ac.uk, j.hough@qmul.ac.uk}

Abstract

This paper addresses the limitations of rule-based and end-to-end neural chatbots with little training data. We compare an AIML-based chatbot and a Seq2Seq chatbot developed on a small, high quality dataset of 300 turns of IT Service queries and responses. The chatbots were then evaluated using ROUGE automated evaluation metrics as well as task completion rate by human judges. While the Seq2Seq model could generalize quite well to new inputs, the rule-based AIML chatbot was found to ensure better task completion rates as well as higher ROUGE scores. The findings suggest that rule-based chatbots are still a useful tool with little resource available, though more needs to be done to confirm their limitations.

1 Introduction

While much previous research has concentrated on developing and testing advanced chatbot systems within rule-based and deep learning based paradigms, there have been few studies that objectively compare these two types of system using the same data. Moreover, with the advance of neural models with large parameter spaces, it is not clear how these models scale downwards when big data is not available.

The purpose of this paper is a preliminary investigation into building, evaluating and comparing two common examples of these two chatbot paradigms in a realistic real-word application and resource scenario: an IT services chatbot at Queen Mary University of London which responds to queries relating to computing issues people may have at the institution, where only a small amount of example dialogue data is available to develop on, without the use of pre-training for a neural system.

2 Methodology

Data A corpus of 400 turns, or 200 query-response pairs, was gathered from the Queen Mary University of London IT Services chat transcripts collected between 2018-2020 between students and human support staff. While creating and cleaning the responses, it was found that 49% of the user queries had detailed relevant answers in the FAQ pages of the Queen Mary IT Services webpages. Thus, where possible, the FAQ answer was used as the response, replacing the original human agent response. 150 query-response pairs were used for the development of both the rule-based and neural chatbots, with the remaining 50 heldout for testing.

Rule-based chatbot We use Artificial Intelligence Markup Language (**AIML**) (Wallace, 2003), an XML-based markup language, to create our rule-based chatbot. AIML uses pattern matching techniques to formulate answers from queries. Each AIML file consists of <category> tags which are the basic unit of knowledge in AIML, containing an input question, an output answer and an optional text. Inside each category, the question is stored in the <pattern> tag while the corresponding answer to the question is stored in the <template> tag, which is the text the chatbot will reply with. The pattern language consists of words, spaces and wildcard symbols such as: "_" and "*". Wildcards are used to replace strings in AIML (words or sentences). The wildcard "_" is given the highest priority, which means that categories containing this wildcard are analysed first (Mikic et al., 2009).

Based on the user intents found while analysing the dataset, a total of ten AIML files were created where each file addresses a specific issue. Out of these ten files, eight of them were created from scratch for the domain (e.g. *Login Issues* and others in Fig. 2) while two were imported from the A.L.I.C.E chatbot system (Wallace, 2009), sourced

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

Task	ROUGE-1			ROUGE-L		
	Average precision score	Average recall score	Average f-measure	Average precision score	Average recall score	Average f-measure
Greetings	0.62	0.69	0.68	0.62	0.69	0.68
Login Issues	0.50	0.48	0.48	0.49	0.48	0.48
MYHR Issues	0.49	0.53	0.47	0.46	0.49	0.45
Password Issues	0.61	0.54	0.55	0.59	0.53	0.54
Password Requirement	0.60	0.50	0.54	0.57	0.50	0.51
Agresso Issues	0.67	0.64	0.65	0.66	0.62	0.64
Address	0.54	0.54	0.548	0.54	0.54	0.548

Task	ROUGE-1			ROUGE-L		
	Average precision score	Average recall score	Average f-measure	Average precision score	Average recall score	Average f-measure
Greetings	0.79	0.73	0.75	0.79	0.73	0.75
Login Issues	0.33	0.23	0.26	0.32	0.22	0.25
MYHR Issues	0.47	0.35	0.39	0.30	0.19	0.23
Password Issues	0.21	0.27	0.23	0.15	0.18	0.17
Password Requirement	0.64	0.52	0.47	0.60	0.54	0.53
Agresso Issues	0.89	0.87	0.88	0.87	0.86	0.86
Address	0.554	0.45	0.49	0.554	0.45	0.49

Figure 1: ROUGE scores of the AIML chatbot (left) and Seq2Seq chatbot (right)

from the website *Kaggle* (Bhatia, 2020). There were on average 12 categories per file, each one designed to closely match and cover all the relevant training set queries, and the response templates for IT issues were the strings from the corresponding responses in the cleaned dataset.

Due to its popularity and relative simplicity, the `Python-aiml` (Stratton, 2003) library was used to build the AIML engine for the chatbot.

Seq2Seq neural chatbot For our neural model, we train an LSTM (long short-term memory) Sequence-to-Sequence (**Seq2Seq**) (Sutskever et al., 2014) model on the 150 query-response pairs in the training data, with the responses identical to the AIML templates. The chatbot was developed in Python using the Tensorflow and Keras libraries (Panchal, 2020).

The Seq2Seq model has one input layer x, a vector of length 40 (the maximum query length) and a decoder input layer, y, a vector of length 141 (the maximum response length). The encoder model has three more layers after the input layers: An Embedding layer (of size 200), an LSTM layer, and the Dense layer of dimension (141, 535), where 535 is the vocabulary size. The Seq2Seq model has a total of 963,135 trainable parameters.

Evaluation We evaluate the success of the responses to the 50 test set queries automatically using a **ROUGE-1** and **ROUGE-L** (Lin, 2004) comparison to the ground truth response (precision, recall and F-1 measure) and also measure **human-judged task completion success**. For the task completion evaluation, both authors judged separately whether the outputs for the models constituted successful outputs or not, based on their knowledge of the IT problem in question, and the authors agreed on the judgement of success on 97 of the 100 responses from the two systems.

TASKS	AIML CHATBOT	SEQ2SEQ CHATBOT
1. GREETINGS	5/7 = 71.4%	5/7 = 71.4%
2. LOGIN ISSUES	3/7 = 42.8%	1/7 = 14.3%
3. MYHR ISSUES	5/11 = 45.5%	3/11 = 27.3%
4. PASSWORD ISSUES	5/8 = 62.5%	1/8 = 12.5%
5. PASSWORD REQUIREMENT	4/6 = 66.7%	4/6 = 66.7%
6. AGRESSO ISSUES	4/6 = 66.7%	5/6 = 83.3%
7. ADDRESS	3/5 = 60%	2/5 = 40%
OVERALL	29/50 = 58%	21/50 = 42%

Figure 2: Task completion rates.

3 Results and Discussion

Using the first author's judgements as the ground truth, as can be seen in Fig. 2, overall the AIML chatbot was found to be approximately 16% more proficient in handling user queries than the Seq2Seq model (58% vs 42% task completion). The AIML chatbot particularly outperformed the Seq2Seq model on *Login Issues*, *Password Issues*, *MyHR Issues*, and *Address* queries.

In terms of automatic metrics, the mean ROUGE-1 and ROUGE-L scores for both chatbots' responses is shown in Fig. 1. As can be seen, in most problem types the AIML bot outperforms the Seq2Seq model across the metrics in each category, with two exceptions (*Greetings* and *Agresso Issues*). The Seq2Seq model shows some generalization with novel input sequences which are similar, but not identical, to those in its training data in less flexible categories: in more varied input categories, AIML is more robust.

This preliminary investigation suggest that with a small amount of data, both in terms of task success and output quality, it is still safer to use a rule-based chatbot with AIML than relying on generalization from an end-to-end neural model. No pre-training of the Seq2Seq model was employed, so future work will involve testing its effect on performance systematically from this baseline.

References

Ruchi Bhatia. 2020. AIML files from Kaggle.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fernando A Mikic, Juan C Burguillo, Martín Llamas, Daniel A Rodríguez, and Eduardo Rodríguez. 2009. Charlie: An aiml-based chatterbot which works as an interface among ines and humans. In *2009 EAEEIE Annual Conference*, pages 1–6. IEEE.

Shubham Panchal. 2020. Creating a chatbot from scratch using keras and tensorflow.

Cort Stratton. 2003. PyAIML – The Python AIML Interpreter.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS*, Montreal, CA.

Richard Wallace. 2003. The elements of AIML style. *Alice AI Foundation*, 139.

Richard Wallace. 2009. The anatomy of ALICE. In *Parsing the Turing Test*, pages 181–210. Springer.

From local hesitations to global impressions of a speaker's feeling of knowing

Tanvi Dinkar
LTCI, Télécom Paris
IP Paris, France
tanvi.dinkar@telecom-paris.fr

Beatrice Biancardi
LTCI, Télécom Paris
IP Paris, France
beatrice.biancardi@telecom-paris.fr

Chloé Clavel
LTCI, Télécom Paris
IP Paris, France
chloe.clavel@telecom-paris.fr

1 Introduction

The aim of this work is to empirically study on a real-life dataset, whether the utterance level use of fillers can help in understanding/ interpreting the perception of the speaker that was formed by a listener. According to Brennan and Williams (1995), the listener's interpretation of the speaker's utterance includes estimates about the speaker's commitment to/ expressed confidence in what they are saying. Flavell (1979) termed these processes (of the speaker) as **metacognitive** ones, that is cognition about cognitive phenomena, or more simply "thinking about thinking". Research has linked fillers to the listener's assessment of a speaker's metacognitive state (Brennan and Williams, 1995). However, these results may not apply to spontaneous speech datasets collected in real-life contexts, or non-QA datasets. Additionally, the focus of analysis tends to be on utterances as if they occur in isolation, rather than part of an overall discourse. Thus existing studies do not focus on the connection between the hierarchical levels of discourse; i.e. how a speaker's local use of fillers could lead to a listener's overall (global) impression of the speaker. In this work, we study how does a speaker's use of fillers relate to the message from the speaker, and consequently, whether this relates to a listener's perception of the speaker. We do so by studying the POM dataset, a corpus of publicly available English monologue movie reviews (Park et al., 2014). Annotators (listeners) were asked to label the reviews for attributes such as "confidence"; without explicitly being told to pay attention to the speaker's use of fillers.

RQ1: (Local effect of fillers): How does a speaker's use of fillers relate to the message from the speaker? **H1**: Fillers are more likely to occur before the introduction of new and upcoming information in the review.

RQ2: (Global effect of fillers): How does the speaker's use of fillers relate to a listener's perception of the speaker? **H2**: the speaker's use of fillers preceding new information in the message contributes to the listener's perception of the speaker's confidence.

2 Methodology

To investigate **H1**, we consider the speaker's mention of entities related to the movie, that we extract from metadata files[1]. These entities could be categorised into actor, director or title of the movie. We inspect for each transcript, the distribution of filler positions, in relation to the automatically annotated entities in the discourse (denoted by Ent). We split these entities into Ent_new; i.e. entities newly introduced in the discourse, to indicate new information, and Ent_old to indicate entities already introduced in the discourse. Then, we check whether the distributions of filler positions (by its token position in the transcript) are significantly different compared to the distributions of 1. Ent_new and 2. Ent_old positions (by its first token's position), by utilising a Kruskal-Wallis H test with Benjamini-Hochberg correction. We then estimate the effect size by computing Cliff's Delta δ. Lastly, we compare the δ distributions of the two experiments, i.e. fillers with Ent_new versus fillers with Ent_old using a Wilcoxon signed-rank test.

To investigate **H2**, we take the mean of the three confidence labels provided by the three annotators. We then consider reviews that are categorised as high-confidence (HC; ratings ≥ 6, n=130) and low-confidence (LC; ratings ≤ 3, n=116). To calculate the percentage of fillers preceding new information (denoted by a new entity), we count the number of fillers in the review that occur before (but not

[1]The complete code and processed data will be made available online for reproducibility here https://github.com/tdinkar/fillers_in_POM.git

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

after) an Ent_new, constrained to a maximum distance of 1 token in between the filler and Ent_new. We normalise by dividing this count by the total number of fillers used in the review. From this, we obtain the percentage of fillers that occur before an Ent_new versus the percentage of fillers used in any other context that is not Ent_new. We then sum these two values for all HC and LC reviews, to get a cumulative percentage. We compute Odds Ratios (ORs) in order to investigate whether the use of fillers around new entities is associated with confidence, where the odds denote the outcome of HC or LC, given the occurrence of fillers before new entities, compared to the occurrence of fillers that do not occur before new entities. We expect that the more fillers are used in the context of preceding new entities, the greater the odds of HC.

3 Results and Discussion

H1: By Kruskal-Wallis H test the distributions of filler positions compared to 1. Ent_new and 2. Ent_old positions are significantly different for $\approx 15-20\%$ of the reviews (where $p \leq .05$). However, after utilising the Benjamini-Hochberg procedure for multiple testing correction, the distributions using this method do not significantly differ. This test is calculated using the sum of the ranks of each distribution. Given that the average review length is short ($\approx$ 256 tokens), and considering the close average median of fillers, Ent_new and Ent_old on reflection, this test may not capture nuances of the positional effects of fillers. However, by computing δ to estimate effect sizes, we found that for most reviews, fillers do occur before Ent_new (median = -0.30 , $SD = 0.41$), but not before Ent_old (median = 0.20, $SD = 0.37$), where the distributions of the δ values significantly differ ($Z = 27578.0$, $p < .05$ using Wilcoxon signed rank test). Majority of the reviews (565) have fillers occurring before Ent_new (sum of "large" to "small" δ sizes)[2], compared to 163 reviews that had negligible effect size, and 139 reviews that had positive effect size (reviews that had fillers occurring after the introduction of new entities). We see the opposite δ effect sizes for Ent_old, where most of the reviews have fillers occurring after entities already introduced in the discourse, but not before. Fillers occurring after Ent_old is entirely plausible given that new entities can occur throughout the review, and not just at the start of one. Given the large group with negligible effect size (247 reviews) for Ent_old, this does show that speakers may sometimes use fillers when repeating entities already introduced into the discourse. Compared to Dinkar et al. (2020), our findings suggest that there is more nuance to the way speakers utilise fillers (and indeed, our methodology is agnostic to sentence boundaries) in spontaneous speech. Therefore, regarding H1, **stylistically speakers do tend to use fillers when introducing a new entity rather than one already introduced** (whether intentionally or not remains an open question), and the positions of fillers w.r.t Ent_new significantly differ from positions of fillers w.r.t Ent_old.

H2: The results of the test show $OR = 0.72$ ($p < .001$, $95\%CI$ $0.6 - 0.8$). While $OR < 1$ (indicating that **the presence of fillers occurring before new entities gives a higher odds of LC), the presence of the stimulus on the outcome is small.** Interestingly, these findings are the opposite of what was expected (the speaker's use of fillers preceding new information contributes to the listener's perception of confidence; i.e. the more fillers are used in this way, the greater the odds of HC). According to the results, fillers occurring before new entities do not have a great effect on the odds of the HC (28% lower given the presence of new entities) rating that the listener gives the speaker. Inspecting the average rate of fillers in the review, it is clear that the use of fillers differs between HC and LC rated speakers (median filler rate of 0.026 and 0.045 respectively, with $U = 3873.0$ and $p < .05$ by Mann-Whitney U test). These results do not necessarily contradict Brennan and Williams (1995), i.e. there could be impressions formed by the listener about the speaker's expressed confidence based on fillers in spontaneous speech (as found in Dinkar et al. (2020)). However, these results would suggest that the effect may not be from fillers used in the context of introducing new entities. This is an interesting finding; as fillers in these contexts could still have a metacognitive function; i.e. the listener is drawn to the mind of the speaker and (their difficulties in) formulating a new referent as discussed in (Barr and Seyfeddinipur, 2010). But, it may be expected by the listener and and thus not necessarily contribute to the listener's perception of the speaker's expressed confidence.

[2]The magnitude of Cliff's Delta δ can be interpreted by using the thresholds from Romano et al. (2006), i.e. $|\delta| < 0.147$ "negligible", $|\delta| < 0.33$ "small", $|\delta| < 0.474$ "medium", and otherwise "large".

References

Dale J Barr and Mandana Seyfeddinipur. 2010. The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, 25(4):441–455.

Susan E Brennan and Maurice Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34(3):383–398.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7985–7993, Online. Association for Computational Linguistics.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI 2014, page 50–57, New York, NY, USA. Association for Computing Machinery.

Jeanine Romano, Jeffrey D Kromrey, Jesse Coraggio, and Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the nsse and other surveys? In *annual meeting of the Florida Association of Institutional Research*, volume 177.

Exploring the Personality of Virtual Tutors in Conversational Foreign Language Practice

Johanna Dobbriner **Cathy Ennis** **Robert Ross**
School of Computer Science
Technological University Dublin
{johanna.dobbriner,cathy.ennis,robert.ross}@tudublin.ie

Abstract

Fluid interaction between virtual agents and humans requires the understanding of many issues of conversational pragmatics. One such issue is the interaction between communication strategy and personality. As a step towards developing models of personality driven pragmatics policies, in this paper, we present our initial experiment to explore differences in user interaction with two contrasting avatar personalities. Each user saw a single personality in a video-call setting and gave feedback on the interaction. Our expectations, that a more extroverted outgoing positive personality would be a more successful tutor, were only partially confirmed. While this personality did induce longer conversations in the participants, we found that interactions with both were enjoyed and that user perception of them differed less than intended.

1 Introduction

When learning a foreign language, Computer Assisted Language Learning (CALL) systems and virtual tutors can be a viable alternative for one-to-one conversational practice. Several of these systems have already been tried – embedded in video games, on their own or as part of a larger CALL system (Dalton and Devitt, 2016; Cheng et al., 2017; Wang et al., 2017; Collins et al., 2019; Divekar* et al., 2021). For these CALL applications, the automated tutor is usually embodied, for example as a virtual avatar with a dialogue system for conversing with the student.

There are several possible avenues to increase user engagement with a virtual avatar, one of which could be tailoring specific personality traits of the virtual tutor to the student. We see this tailoring as a manifestation of pragmatics modelling and choice as ultimately the conversational policy must be tuned to and react to the conversational traits of the user. As a first step towards investigating this hypothesis, we designed a Wizard-of-Oz (Kelley, 1984) pilot experiment to see whether there are any observable variations in the interaction and feedback when students are confronted with two opposing tutor personalities.

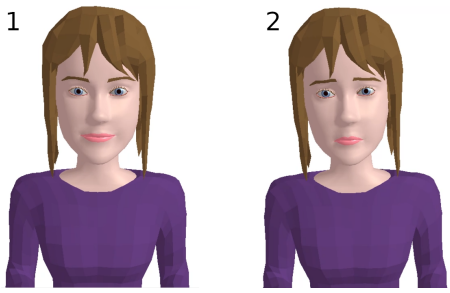


Figure 1: Avatar expressing different personalities – P1 (left) and P2 (right)

2 Experiment Design

For our experiment, each participant provided some demographic information before talking to one of the two virtual avatars for a few minutes in English followed by rating the avatar personality and giving feedback on the conversation itself.

To implement this study, we built upon the expressive avatar from Sloan et al. (2020), animated with an Irish English, female voice [1] to embody our virtual tutor. Our two avatar personalities varied along three of the OCEAN model's (Goldberg, 1990) five dimensions – extroversion, openness and agreeableness and were expressed through dialogue scripts where the avatar exhibited specific personality traits, posture, facial expression and speech characteristics (see Figure 1). Personality 1 (P1) represents the higher end of the 3 dimensions, being open, friendly and sociable, while personality 2 (P2) exhibits low values along the 3 dimensions,

[1] https://www.cereproc.com/en/node/1155

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

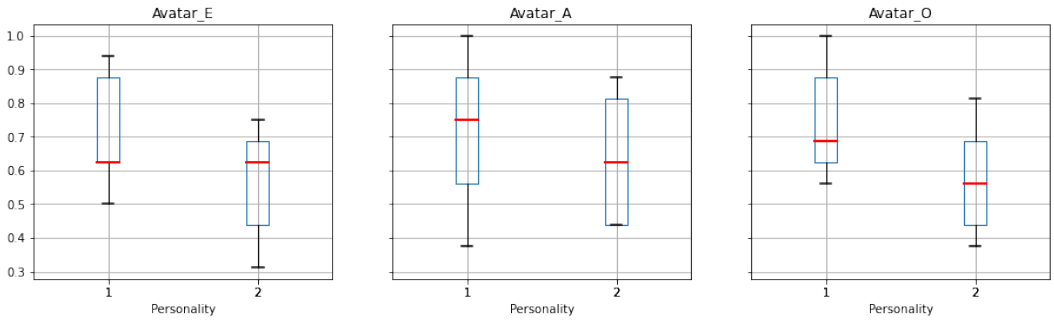


Figure 2: Boxplots of Avatar personality scores between avatar personalities for Extroversion (Avatar_E), Agreeableness (Avatar_A) and Openness (Avatar_O)

behaving in a more closed off, curt and distant manner.

The participants of this experiment were adult English learners, 18 years or older, who had been learning English between 7 and 22 years ($M = 12.5$, $SD = 4.33$). 18 participants completed the study, 44% male and 56% female, aged 18 - 45 years. At 83%, the majority of participants spoke German as their native language with another 11% Italian and 5.6% (one person) Chinese. The total duration of the participant's interaction with the avatar varied significantly between avatar personalities with P1 interactions lasting an average of 12.15 minutes ($SD = 1.81$) and P2 conversations being markedly shorter at 7.55 minutes on average ($SD = 1.54$). With a sample size as low as this, any statistics computed on the collected data cannot be very robust and all results are to be taken as indicative.

3 Results

We expected P1 to be more pleasant and enjoyable to converse with, which we expected in turn would show itself in positive user feedback and high scores on an avatar personality survey, with low personality scores and fewer participants enjoying the interaction for P2. Additionally, we hypothesised that P1's conversation style would encourage participants to talk more than P2.

Our results as shown in Figure 2, only partly support this hypothesis: While P1 generally achieved higher scores, the box plot for Agreeableness shows much overlap between personalities. For Extroversion, the median is the same in both groups. The clearest distinction is found for Openness. A T-test at a significance threshold of 0.05 confirms these results, with Openness ($T(18) = 2.166$, $p = 0.046$) showing the only significant difference between P1 and P2, whereas Extroversion ($T(18) = 1.816$, $p = 0.088$) is close to significant and may prove distinct with more participants. Agreeableness ($T(18) = 1.099$, $p = 0.288$) was not perceived as significantly different, likely due to the avatar unintentionally interrupting the participants occasionally due to connectivity issues and human error. The speaking time ratio (human vs. avatar) showed the expected significant difference ($T(18) = 2.525$, $p = 0.022$), so P1 appears to encourage students to talk more than P2. In terms of user feedback, all participants noted they enjoyed the interaction regardless of avatar personality.

4 Conclusions and Future Work

In conclusion, this initial study opens the door to a multitude of interesting research directions that bear further exploration.

In this pilot study we found that only one out of three personality dimensions, Openness, was perceived significantly differently between both groups. Extroversion came close to the significance threshold and Agreeableness was widely distributed with relatively high scores in both groups. However, the participant's speaking time relative to the avatar was significantly higher in the personality designed to be more pleasant overall, thus matching our expectations.

Further research might focus on making the personality differences more apparent, building a chatbot for each personality, adding more personalities exploring measures to automatically adapt to the student.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland Centre for Research Training in

Digitally-Enhanced Reality (D-REAL) under Grant number [18/CRT/6224]. 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

Alan Cheng, Lei Yang, and Erik Andersen. 2017. Teaching language and culture with a virtual reality game. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 541–549.

Naoise Collins, Brian Vaughan, Charlie Cullen, and Keith Gardner. 2019. Gaeltechvr: Measuring the impact of an immersive virtual environment to promote situated identity in irish language learning. *Journal For Virtual Worlds Research*, 12(3).

Gene Dalton and Ann Devitt. 2016. Gaeilge gaming: Assessing how games can help children to learn irish. *International Journal of Game-Based Learning (IJGBL)*, 6(4):22–38.

Rahul R. Divekar*, Jaimie Drozdal*, Samuel Chabot*, Yalun Zhou, Hui Su, Yue Chen, Houming Zhu, James A. Hendler, and Jonas Braasch. 2021. Foreign language acquisition via artificial intelligence and extended reality: design and evaluation. *Computer Assisted Language Learning*, 0(0):1–29.

Lewis R Goldberg. 1990. An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

John Sloan, Daniel Maguire, and Julie Carson-Berndsen. 2020. Emotional response language education for mobile devices. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA. Association for Computing Machinery.

Yifei Wang, Stephen Petrina, and Francis Feng. 2017. VILLAGE - virtual immersive language learning and gaming environment: Immersion and presence. *Br. J. Educ. Technol.*, 48(2):431–450.

Getting from A to B: Exploring Floor State Transitions in Conversation

Emer Gilmartin
ADAPT Centre, Trinity College Dublin
Ireland
gilmare@tcd.ie

Marcin Włodarczak
Stockholm University
Sweden
wlodarczak@ling.su.se

1 Introduction

In this abstract we describe ongoing analysis of multiparty spoken interaction, where participants start and stop speaking, taking and relinquishing the floor and locally arranging turn change and retention (Sacks et al., 1974). We consider conversation in terms of stretches of speech and silence, using only timing information for analysis. We have devised a method of labelling to capture *floor state transitions* - sequences of speech and silence involved in transitions from a stretch of one party speech (speech in the clear) by one speaker to the next stretch of one party speech by the same speaker (within speaker transition - WST) or another speaker (between speaker transition - BST). To approximate turn changes and retention, we impose left and right hand side minimum duration thresholds on the single party speech in the clear bordering the transitions. We dub the intervals of speech, silence and overlap between the single-speaker stretches *intervening* intervals. We have been analysing patterns of intervening intervals in multiparty talk, concentrating on 3-party interaction in Estonian, Swedish, and English. Below we briefly explain the labelling scheme and summarize results to date in this work.

2 Labelling Scheme

We define the 'floor state' at any point of a conversation as the totality of participants speaking, and represent interaction as a series of labels for intervals of varying where a particular floor state prevails (see Figure 1). For example, an interval where A and B are speaking in overlap is labelled AB, C speaking alone is labelled C, and general silence X. In n-party speech, there are 2^n possible floor states, so 3-party speech could include any of the 8 labels: X, A, B, C, AB, AC, BC, ABC. We define a transition as the set of intervening intervals between two stretches of single party speech. We impose left and right hand 1-second minimum thresholds on the single party speech, generating *1Sp1-1Sp1* transitions, which can be BST or WST.

Figure 1 shows a stretch of talk with three instances of two-party overlap (AB, AC, AC), an instance a three-party overlap (ABC), and three intervals of solo speech (A, B, C). We can define a five-interval BST from A to C comprising **AB_ABC_AB_B_BC**. Note that if the right hand one-second threshold were not applied, the example would be classified as involving two transitions (from A to B and from B to C), even though the short stretch of solo speech by B is unlikely to be a claim for turn possession.

We process segmentation data from spoken interaction with a Python script using TextGridTools (Buschmeier and Włodarczak, 2013) to create floor state and transition labels and extract the number and identity of participants speaking during the transition. All code and annotation are available at https://zenodo.org/record/4923246.

3 Summary of Results to Date

We have used these labels to explore floor state transitions in corpora of three-party spontaneous conversations in Estonian (Lippus et al.), Swedish (Włodarczak and Heldner, 2017), and the TEAMS corpus of collaborative conversational games in English (Litman et al., 2016), and also in casual multiparty talk in English.

In all cases, the majority of transitions involve more than one intervening interval to complete and the vast bulk of transitions involve odd numbers of intervening intervals (Gilmartin et al., 2019, 2020; Włodarczak and Gilmartin, to appear). The scarcity of even numbers of intervening intervals follows from the rarity of smooth switches and instances of simultaneous onset or offset of speech. We have

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

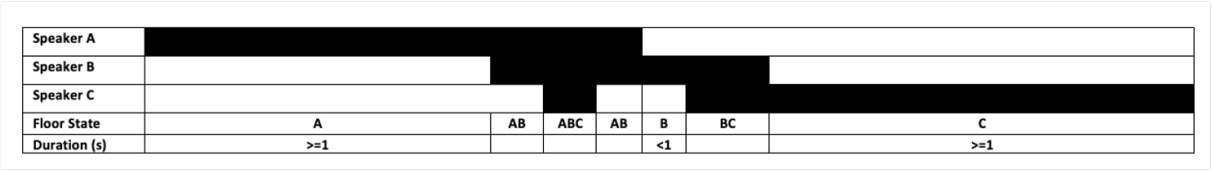


Figure 1: An example of a between-speaker transition.

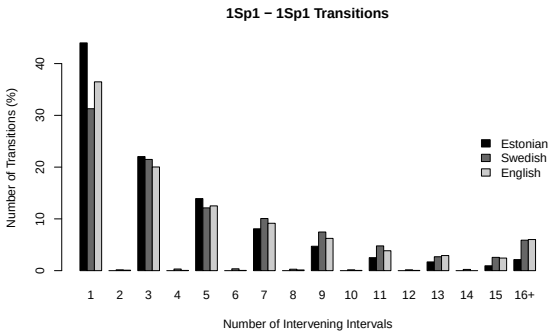


Figure 2: Number of floor state intervals between single-speaker intervals of 1 second or more in duration - Estonian, Swedish, and English 3-party conversation

found that one-interval transitions are the largest class, and the frequency of transitions decreases with increasing numbers of intervening intervals, as shown in Figure 2. We have also found that WSTs account for more of the one-interval transitions, particularly around silence, perhaps due to breathing pauses.

In terms of speaker participation in transitions, one-interval transitions are silence or overlap,with 0 or 2 speakers involved. With more intervals, transitions can have more participants, with participation by all three speakers more likely in BST than WST. In the Estonian, Swedish, and Teams data, silence was present in over 90% of transitions, with overlap appearing in 53%. With more intervals, transitions contain more complex combinations of speech and silence, and all of these features become more likely. Silence occurs in the vast bulk of transitions of 3 or more intervals, as does solo speech. The incidence of overlap increases with increasing number of intervals, and is more common throughout in BST than WST. (Wlodarczak and Gilmartin, to appear)

We found that silence accounts for a large share of the duration of 3-intervals WSTs and BSTs, and remains the lead component in terms of duration but decreases with increasing numbers of intervening intervals, while the duration from overlap increases (Wlodarczak and Gilmartin, to appear).

The distribution of the most common transition sequences across the datasets are similar. In all cases, the most common sequences overall were A_X_A (within speaker silence) and A_X_B (between speaker silence). Interestingly, for WST, both A_X_A_X_A and A_X_B_X_A were more common than 1-interval overlap (A_A:B_A), while the second most common BST was 1-interval overlap (A_A:B_B).

Almost 60% of all transitions are 1- or 3-interval. For all languages, the most common 5-interval transitions were less frequent than the fifth most frequent 3-intervals, except for English WSTs, where the most common 5-interval transition was marginally more frequent than the fifth most frequent 3-interval WST. We therefore further explored the 3-interval transitions to understand the most common transitions in the data. The first and second most common 3-interval WST and BST sequences for were the same in all three languages analysed. All of the top five 3-interval within speaker transitions across the three languages are accounted for by six transition labels, while the top five between speaker transitions are covered by seven transition labels. The categories also show great similarity in their percentage frequencies across the three datasets.

4 Ongoing Work

Our explorations have shown interesting results on the composition of within and between speaker transitions in multiparty talk, with similarities in how these occur across the languages we have analysed. We are expanding the analysis to other corpora, including dyadic speech, to see how well our findings generalize. We intend to create an inventory of transitions most commonly found across a large number of corpora, and will perform detailed phonetic analysis of the more common sequences. This will add to our understanding of how spoken interaction works, as well as inform design of more appropriate spoken dialog technology in applications requiring human like behaviour.

5 Acknowledgements

This work was conducted with the support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant Number 13/RC/2106. The work was also funded by Swedish Research Council project 2019-02932 *Prosodic functions of voice quality dynamics* to Marcin Włodarczak.

References

Hendrik Buschmeier and Marcin Włodarczak. 2013. TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013)*, volume 65 of *Studientexte zur Sprachkommunikation*, pages 152–157, Dresden. TUDpress.

E. Gilmartin, Kätlin Aare, M. O'Reilly, and M. Wlodarczak. 2020. Between and within speaker transitions in multiparty conversation. In *Speech Prosody*, pages 799–803.

Emer Gilmartin, Mingzhi Yu, and Diane Litman. 2019. Comparing speech, silence and overlap dynamics in a task-based game and casual conversation. In *Proceedings of ICPhS 2019*, pages 3408–3412.

Pärtel Lippus, Tuuli Tuisk, Nele Salvestre, and Pire Tiras. Phonetic corpus of Estonian spontaneous speech.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

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

Marcin Włodarczak and Mattias Heldner. 2017. Respiratory constraints in verbal and non-verbal communication. *Frontiers in Psychology*, 8:708.

Martin Wlodarczak and Emer Gilmartin. to appear. Speaker transition patterns in three-party conversation:
evidence from english, estonian and swedish. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. ISCA.

Annotating Low-Confidence Questions Improves Classifier Performance

Stephanie Hernandez* and **Ron Artstein**
USC Institute of Creative Technologies
12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
st3phy831@gmail.com artstein@ict.usc.edu

*Now at Hartnell College and CSU Monterey Bay

Abstract

This paper compares methods to select data for annotation in order to improve a classifier used in a question-answering dialogue system. With a classifier trained on 1,500 questions, adding 300 training questions on which the classifier is least confident results in consistently improved performance, whereas adding 300 arbitrarily selected training questions does not yield consistent improvement, and sometimes even degrades performance. The paper uses a new method for comparative evaluation of classifiers for dialogue, which scores each classifier based on the number of appropriate responses retrieved.

1 Introduction

Statistically trained dialogue systems can often be improved by adding annotated training data; when the system is deployed with real users, it often collects more interaction data than can be annotated, so prioritization of the data for annotation is required. This paper presents an experiment on selecting data for annotation using a dialogue system's internal confidence measure: it prioritizes annotation of those utterances for which the system is the least confident about how to react. This can be considered a form of active learning (Settles, 2010). Adding these utterances as training data (with appropriate annotations) improves the system's performance, whereas adding a comparable number of utterances that were arbitrarily selected does not improve performance to the same extent.

We use data from the Digital Survivor of Sexual Assault (Artstein et al., 2019), which is a system based on NPCEditor, a classifier trained on linked questions and answers (Leuski and Traum, 2011). For each new question, the classifier provides a score for every available answer, reflecting how well the classifier thinks it answers the question, and then returns the answers whose confidence exceeds a threshold (the list may be empty if all answers are below threshold). This experiment uses these scores to identify low-confidence questions to prioritize for annotation.

2 Method

2.1 Materials

For the baseline system we chose a very limited dataset, with 1,542 questions and only 1,517 links between questions and answers. Starting with an impoverished baseline allows room for measurable improvement with the addition of a small number of questions and links, whereas on a better trained baseline, the impact of additional training data is expected to be smaller. Also, the small baseline system left us with many questions that were already annotated and available for the experiment. The additional training data were taken from four datasets of questions annotated with links to appropriate answers (the four datasets were labeled "Alpha", "Beta", "Beta2" and "Windows", reflecting the development stage at which the questions were collected; see Artstein et al. 2019). All systems were tested using a fixed test set of 399 questions linked to appropriate answers.

2.2 Procedure

From each of the four annotated datasets, three sets of 300 questions (with corresponding links) were extracted: The "Duplicates" set simply selected 300 arbitrary questions, possibly including duplicate questions (that is, instances where the same exact question was asked by different users, though possibly annotated with different links). The "No-Duplicates" set also selected 300 arbitrary questions, ensuring that the 300 questions are all distinct from one another. Selection of questions for both the "Duplicates" and "No-Duplicates" sets was done through custom python scripts. A

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

third set of 300 questions was extracted by giving a full dataset of questions to the baseline classifier; the 300 questions for which the classifier returned the lowest confidence were chosen as the "Active Learning" dataset. Overall, we extracted 12 sets of 300 questions annotated with links (three from each of the four datasets).

Each of the 12 sets of annotated questions was added (separately) to the baseline classifier, and the resulting classifier was retrained; this resulted in a total of 13 classifiers (including the baseline). Each of the 13 classifiers was then run on the test set, returning an output of ranked responses for each question. A custom python script was then used to tabulate the outputs, pairing each question with the top-ranked three responses from each classifier (or fewer responses, if the classifier returned fewer than 3).

2.3 Evaluation

Traditional measures such as precision and recall are not well-suited for comparing the performance of ranked lists, because of the way responses are used in a dialogue system: the most common action of the system is to choose the top-ranked response, less commonly it chooses the second, then the third and so on. For this experiment, we chose to compare classifiers by the number of appropriate responses retrieved: for each of the test questions, a point was given to the classifier (or classifiers) with the highest number of appropriate responses within the top three. The total score of a classifier therefore reflects the number of questions for which it retrieved the highest number of appropriate responses, compared to the other classifiers.

3 Results

For 54 of the test questions, all classifiers gave the same output; these questions are excluded from further analysis. Figure 1 shows the classifier scores on the remaining 345 questions. These results show that the Active Learning classifiers score consistently above the baseline ($t(3)=8.8$, $p=0.003$); they also score signifantly higher than the Duplicates classifiers (Mann-Whitney $U=16$, $p=0.03$), and the difference from the No Duplicates classifiers approaches significance ($U=15$, $p=0.06$). No other differences were significant. Interestingly enough, many of the Duplicates and No Duplicates classifiers scored below the baseline, though these differences were not statistically significant.

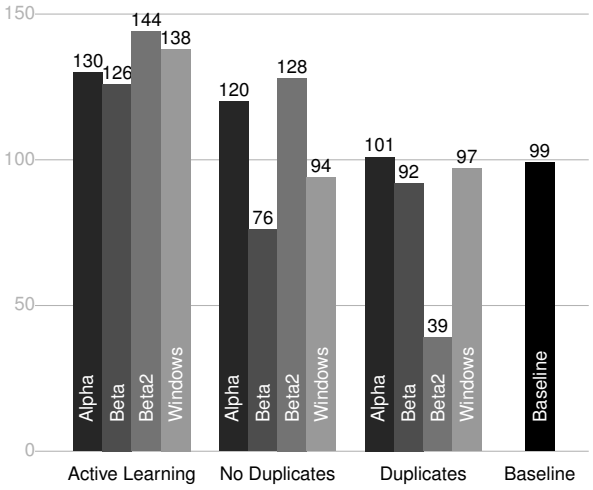


Figure 1: Classifier performance (points)

4 Discussion

The experiment shows that adding a small amount of low-confidence questions as training data can consistently improve the performance of a classifier, whereas adding the same amount of arbitrary questions does not lead to consistent improvement; this suggests that the classifier's confidence is a useful measure for prioritizing annotation. One limitation of this experiment is the impoverished baseline classifier, which reflects the very earliest stages of dialogue system development; at this stage, systems are usually not widely deployed and development budgets are still relatively large, so it is common to annotate all the available data anyway. It remains to be seen whether this method is useful at more mature stages of development, when the amount of available data exceeds the capacity for annotation. Another interesting observation is that some cases of added training data resulted in lower performance: this suggests that perhaps annotating all the available data is not the best approach, and that careful curation of data to be annotated needs to be explored.

Acknowledgments

The first author was supported by NSF award 1852583 "REU Site: Research in Interactive Virtual Experiences" (PI: Ron Artstein). The second author was 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

Ron Artstein, Carla Gordon, Usman Sohail, Chirag Merchant, Andrew Jones, Julia Campbell, Matthew Trimmer, Jeffrey Bevington, COL Christopher Engen, and David Traum. 2019. Digital survivor of sexual assault. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 417–425, Marina del Rey, California. ACM.

Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Burr Settles. 2010. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Situated UMR for Multimodal Interactions

Kenneth Lai[1], Richard Brutti[1], Lucia Donatelli[2], James Pustejovsky[1]
[1]Department of Computer Science
Brandeis University, USA
[2]Department of Language Science and Technology
Saarland University, Germany
{klai12,brutti,jamesp}@brandeis.edu
donatelli@coli.uni-saarland.de

Abstract

We discuss the requirements on a meaning representation language for annotating both verbal and non-verbal communicative acts in multimodal interactions, as it impacts both the deployment of annotation efforts and the development of corpora reflecting these phenomena. We argue that Uniform Meaning Representation (UMR) can be naturally extended to capture multiple communicative channels in discourse, including both language and gesture, while also encoding the annotation of referential grounding in situated contexts.

1 Introduction and Related Work

As multimodal interactive systems become more common and sophisticated, there are increasing expectations that they will approximate interactions with other humans. Human-computer interaction (HCI) and human-robot interaction (HRI) involve communicating intentions, goals, and attitudes through multiple modalities beyond language, including gesture, gaze, and situational awareness. With this interest comes a need for capturing and representing the data that encodes these different modalities during such interactions.

Any representation suitable to this task should, at a minimum, both accommodate the structure and content of the different modalities, as well as facilitate alignment and binding across them. However, it is also important to distinguish between alignment across channels in a multimodal dialogue (language, gesture, gaze), and the situated grounding of an expression to the local environment, be it objects in a situated context, an image, or a formal registration in a database. Therefore, such a meaning representation should also have the basic facility for situated grounding; i.e., explicit mention of object and situational state in context.

Presently, there are few meaning representation languages for situated (dialogue) interactions, that are both adequately expressive of the content and compact enough for corpus development. There have been several annotation efforts utilizing Abstract Meaning Representation (AMR) (Banarescu et al., 2013). Advantages of AMR include its relative simplicity, ease of annotation, and available corpora. AMR has been expanded and applied to multi-sentence settings (O'Gorman et al., 2018), and to task-oriented dialogues (Bonial et al., 2020).

More recently an extension of AMR, Uniform Meaning Representation (UMR) has been developed to be scalable, accomodate cross-linguistic diversity, and support lexical and logical inference (Van Gysel et al., 2021). To this end, UMR incorporates aspect, scope, temporal and modal dependencies, as well as inter-sentential coreference.

We argue that we can combine multimodal elements in a single representation for alignment and grounded meaning. Specifically, we believe that an enriched version of UMR, which we call *Situated UMR (SUMR)*, is an ideal representational format to this end. This allows for an immediate referencing for deictic, pronominal, and underspecified expressions, as well as a spatial "registration" for objects in the discourse (and common ground).

A variety of corpora exist that seek to capture language and dialogue in a situated environment. However, existing annotation schemes often fall short of capturing true situated *meaning*, instead annotating distinct channels separately with little guidance as to how these channels interact to create emergent meaning (Krishnaswamy and Pustejovsky, 2019). The SCOUT corpus is an example of a situated, but *unimodal*, dataset (Bonial et al., 2020). It introduces Dialogue-AMR to extend and enrich AMR in support of HRI, in a navigation setting. The EGGNOG corpus (Wang et al., 2017) is comprised of video of two participants working on shared tasks. It is annotated with the annotator-inferred *intent* of the gestures, as well as their morphology

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

(physical description). However, the annotation is a label with no grounding. Objects introduced in one intent are not available in the next; objects are *identified* but not referenced or registered.

2 Common Ground in Situated UMR

For the present discussion, we focus on the semantics of integrated multimodal expressions in the context of task-oriented dialogues. We assume the model presented in Pustejovsky and Krishnaswamy (2021) and Krishnaswamy and Pustejovsky (2021), where a *common ground structure* (CGS) integrates both intermodal expressions in the discourse and the situational anchoring to objects perceived and referenced in the context. An agent's communicative act, C_a, is a tuple of expressions from the diverse modalities involved (e.g., speech S, gesture G). The CGS embeds C_a within a monad identifying: **A** the communicating agents; **B**, the salient shared belief space; **P**, the objects and relations that are jointly perceived in the environment; and $\mathcal{E}$, the agents' joint embedding space. Here we focus on a communication in relation to the perceived context.

Consider Figure 1, where a multimodal command aligns the linguistic utterance, "*That move there*" with an ACTION-RESULT gesture sequence of "*Point Action Point*". This example illustrates two kinds of gestures: (a) establishing a reference; and (b) depicting an action-object pair (Kendon, 2004; Lascarides and Stone, 2009).[1]

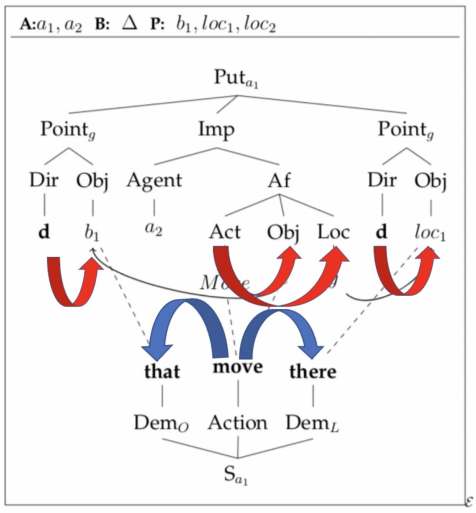


Figure 1: Intermodal alignment between linguistic and gesture dependency structures

Given these assumptions, we introduce a multimodal extension of UMR we call *Situated UMR (SUMR)*, that allows for the representation of both multiple channels of communication, as well as the perceptual (object and situational) awareness present to an agent in the common ground.

```
(a) (c / cgs
    :agent (a / agent)
    :agent (a2 / agent)
    :perception (b / block)
    :perception (l / location)
    :perception (l2 / location))

(b) (s1c2 / command-00
    :ARG0 a1
    :ARG1 (c3 / communicative-act
          :gesture (g / gesture-unit
                    :op1 (d / deixis
                          :DIR (v / vector)
                          :OBJ b)
                    :op2 (a3 / action
                          :ACT (m / move-01)
                          :OBJ (i / implicit-role
                                :op1 "moved")
                          :LOC (i2 / implicit-role
                                :op1 "destination"))
                    :op3 (d2 / deixis
                          :DIR (v2 / vector))
                          :OBJ l))
          :speech (m2 / move-01
                   :mode imperative
                   :ARG0 (i3 / implicit-role
                          :op1 "mover")
                   :ARG1 (t / that)
                   :ARG2 (t2 / there))
    :ARG2 a2)

(c) (s1 / sentence
    :coref ((a2 :same-entity i3)
            (b :same-entity i)
            (b :same-entity t)
            (l :same-entity i2)
            (l :same-entity t2)))
```

Figure 2: Example SUMR corresponding to the communicative act in Figure 1

The example SUMR in Figure 2 has three parts. First, in (a), the agents and perceived objects are listed in the CGS (in the example, **B** and $\mathcal{E}$ are omitted for brevity, but can be included). For each communicative act, we have a sentence-level UMR representation (b), with the gesture and speech modalities labeled. We assume the dialogue act annotation from Bonial et al. (2020). As with sentences in text and discourse, gestural expressions can also be sequenced; and, as in multi-sentence AMR, their corresponding individual representations can capture the object coreference inherent in the discourse (O'Gorman et al., 2018). This is captured in the document-level representation (c). Details of the alignment of the speech and gesture expressions are beyond the scope of this poster.

As a platform for multimodal situated dialogue annotation, we believe that SUMR has some attractive properties. It is adequately expressive at both utterance and dialogue levels, while easily accommodating the dependency structures inherent in gestural expressions. Further, the native reentrancy facilitates both the linking between modalities and situational grounding to contextual bindings.

[1] We adopt the gesture grammar developed in (Pustejovsky and Krishnaswamy, 2021).

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-amr: Abstract meaning representation for dialogue. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 684–695.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Nikhil Krishnaswamy and James Pustejovsky. 2019. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51.

Nikhil Krishnaswamy and James Pustejovsky. 2021. The role of embodiment and simulation in evaluating hci: Experiments and evaluation. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior*, pages 220–232, Cham. Springer International Publishing.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, page ffp004.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

J Pustejovsky and N Krishnaswamy. 2021. Embodied human computer interaction. *Künstliche Intelligenz*.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, pages 1–18.

Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2017. EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *To appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*.

Challenges for the Conversational Entity Dialog Model

Wolfgang Maier and **Stefan Ultes**
Mercedes-Benz Research & Development
Sindelfingen, Germany
{wolfgang.mw.maier,stefan.ultes}@daimler.com

Abstract

The conversational entity dialogue model (CEDM) (Ultes et al., 2018) offers an intuitive way of modeling task-oriented dialogues in a statistical spoken dialogue system around objects and relations instead of task domains. We address several open challenges around the CEDM and possible extensions of the model.

1 Introduction and Motivation

Research in statistical spoken dialog systems (SDS) (Young et al., 2013) has produced successful systems for task-oriented human-machine dialogues (Lison, 2011; Wang et al., 2014; Budzianowski et al., 2017, among others). In such systems, dialogues are generally modeled using a *multi-domain dialogue model* (MDDM), i.e., they are modeled around single or multiple task domains.

The MDDM focuses on task domains. Therefore, it is hard to describe particular objects of the same type, or to address relations between them. As an example, consider the following example, where two objects (restaurant and hotel) are requested which share an attribute (area).

user i am looking for a restaurant and a hotel in the same area

The first to provide a principled approach for intuitively modeling objects and their relations in the context of SDS have been Ultes et al. (2018). In their *conversational entity dialogue model* (CEDM), dialogues are modeled around objects and the relations between instead of domains.

The CEDM leaves open a number of challenges, which we want to address in this paper. In the following section, we briefly describe the CEDM, and then discuss challenges and possible extensions in Sec. 3 and 4. We close by presenting perspectives for future work.

2 Conversational entity dialog model

The CEDM (Ultes et al., 2018) defines a *conversational entity* as a virtual conversational entity that exists in the context of the current conversation and that is either a conversational object or a conversational relation. A conversational *object* is a conversational entity with a certain type together with a set of attributes which may or may not map to a real-world entity. A conversational *relation* is a conversational entity which connects objects or attributes of objects. Object instances reside in a conversational world that can be derived from the user input, or be predefined.

Dialogues using the MDDM can be modeled using the CEDM, by treating a domain as a conversational object of a specific type, and the slots as the attributes of the type (Ultes et al., 2018, Sec. 4.4). As the CEDM additionally allows for the modeling of relations, it is more expressive than the MDDM.

For more details regarding the handling of belief tracking, etc., please refer to Ultes et al. (2018).

3 Hierarchical Extension

The CEDM can model objects with attributes and relate the attributes. However, types in the CEDM are flat. For instance, the type *hotel* is no more related to the type *guesthouse* than it is to the type *lamp*. This makes it challenging to model (or rather talk about) semantic relations between objects such as hypernymy or hyponymy. Similar to relations in the CEDM that relate attributes of object (e.g., *price2price*), one could allow conversational objects that relate types using a WordNet-style (Fellbaum, 1998) relation (e.g., τ_1 *is a* τ_2 for two types τ_1, τ_2). Consider the following example:

user i am looking for a running outfit
sys here is a suggestion for a jacket, shirt, pants, underwear, and shoes
user i want only black *outerwear*

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

Here, *outerwear* would be a new conversational object with attribute *color=black*, such that for all previously objects except the underwear stand in an *is a* relation with it.

Note that this method explicitly models a relation between the actual objects in the conversation, i.e., it is more expressive than that the ontological knowledge from a backend knowledge base alone.

4 Further Challenges

Count The count of objects present in the conversation can be essential.

user	i want to book the tour for tomorrow at 8am
sys	how many people will participate
user	four
sys	please tell me the name of the first person
⋮	
user	did I say four, make that three persons.

In this case, the count of the *person* objects could be a conversational object itself which needs to be linked to the count of the person objects in the conversational world (Ultes et al., 2018).

Additional knowledge In case several objects are present in the conversation, a user can refer to a subset of the objects in various ways. Above, we have described the case of using a hyperonym (*outerwear*). It is easy to find a similar example where no direct hyperonym is involved.

user	i am looking for a running outfit
sys	here is a suggestion for a jacket, shirt, pants, underwear, and shoes
user	please *no pockets*

In this case *no pockets* addresses a subset of the objects in the conversation, probably *jacket* and *pants*. It could be resolved through the backend knowledge base of the SDS. The subset itself could be handled in the same way as proposed in the previous section.

One can think of more sophisticated ways of addressing subsets of objects in the conversation, such as the following example.

user	i am looking for a running outfit
sys	here is a suggestion for a jacket, shirt, pants, underwear, and shoes
user	too boring above the waistline, can you suggest something else

In such cases, it would be harder to determine what the actual subset is, as it depends, e.g., on a user model (cf. *boring*) or on more general world knowledge. However, the approach of addressing the actual subset once it is found can be the same as before.

Relating multiple attributes The CEDM introduces binary relations between the attributes of objects in the conversation. The following example however implies an n-ary *equals* relation for the *color* attributes.

user	i am looking for a running outfit
sys	here is a suggestion for a jacket, shirt, pants, underwear, and shoes
user	can you suggest something with identical colors?

A similar challenge is the selection of the correct subset of objects based on an attribute for which a particular relation holds, such as *color=red* in the following example.

user	i am looking for a running outfit
sys	here is a suggestion for a jacket, shirt, pants, underwear, and shoes
user	can you switch the red items for yellow ones?

More complex relations Ultes et al. (2018) present the *equals* relation, as for instance in the following example, where it holds between the *area* attributes of the restaurant and the hotel objects.

user	i am looking for a hotel and a restaurant *in the same area*

Also, other relations are mentioned such as *less than*. A trivial extension would be to allow for relations on strings such as *startswith*:

user	show my contacts with last name starting with 'Z'

One can also think of other knowledge-based relations, such as *matching* in the following example.

user	i am looking for a running outfit
sys	here is a suggestion for a jacket, shirt, pants, underwear, and shoes
user	i want shirt and pants to have *matching* colors

5 Future Work

In this paper, we have sketched open challenges of the conversational entity dialogue model (Ultes et al., 2018). Currently, we are working on an integration of some of the presented aspects into the PyDial dialogue system (Ultes et al., 2017) with the goal of a proper evaluation.

References

Paweł Budzianowski, Stefan Ultes, Pei-Hao Su, Nikola Mrkšić, Tsung-Hsien Wen, Iñigo Casanueva, Lina M. Rojas-Barahona, and Milica Gašić. 2017. Sub-domain modelling for dialogue management with hierarchical reinforcement learning. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 86–92, Saarbrücken, Germany. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Pierre Lison. 2011. Multi-policy dialogue management. In *Proceedings of the SIGDIAL 2011 Conference*, pages 294–300, Portland, Oregon. Association for Computational Linguistics.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yen-Chen Wu, Steve Young, and Milica Gašić. 2018. Addressing objects and their relations: The conversational entity dialogue model. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017. PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

Zhuoran Wang, Hongliang Chen, Guanchun Wang, Hao Tian, Hua Wu, and Haifeng Wang. 2014. Policy learning for domain selection in an extensible multi-domain spoken dialogue system. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 57–67, Doha, Qatar. Association for Computational Linguistics.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Conflict Search Graph for Common Ground Consistency checks in Dialogue Systems

Maria Di Maro
University of Naples
"Federico II"
maria.dimaro2@unina.it

Antonio Origlia
University of Naples
"Federico II"
antonio.origlia@unina.it

Francesco Cutugno
University of Naples
"Federico II"
cutugno@unina.it

Abstract

In this work, we account for the formalisation of a Conflict Search Graph as a module managing domain knowledge, dialogue state tracking and information consistency in dialogue systems. Insights on its ability to recognise Common Ground Inconsistencies and make them explicit via specific linguistic feedback are also reported.

1 Introduction

The wide success and the current spread of conversational agents are shedding a new light not only on conversation analysis but also on computational pragmatics. In fact, beside the study of dialogue systems architectures, training techniques and materials, many other aspects are important when dealing with conversational agents. Among them one is not to take for granted: *Understanding* (i.e., in terms of words identification, word meaning, and *speaker meaning/intention*). To make Understanding an easier task, messages are, usually, encoded upon the so called *common ground*.

As pointed out by scholars such as Clark (1996), to pursue the aim of succeeding in the joint activity of conversation, the interlocutors need to *ground* what is being communicated. *Grounding* refers to the act of establishing that what we intend to say (or what has been said) can be well understood (or has been well understood) (Clark and Brennan, 1991). To establish a *common ground* (CG), different strategies, such as linguistic or para-linguistic feedback (Traum, 1999), are adopted. From the linguistic point of view, dialogue efficiency can rely on the analysis of communicative feedback, whose relevance was pointed up by Allwood et al. (1992) and which continues to be considered as an important characteristics in dialogue modelling (Buschmeier and Kopp, 2018).

In this report, we consider the specific case of deliberation dialogues, as defined in Walton (1984); Walton and Krabbe (1995) and we investigate how corrective feedback, in this type of dialogue, can be generated upon problems in the Common Ground, namely when inconsistencies in the Common Ground occur (§ 2). Specifically, we propose the use of graph databases as an integrated solution to dialogue state tracking, knowledge representation, and conflict detection as a fundamental building block for dialogue systems with argumentation capabilities (§ 3).

2 Common Ground Inconsistencies

With *Common Ground Inconsistencies* we refer to the incompatibility between the listener belief and the new evidence provided by the speaker. Given a domain D, we define a set of sequential actions A as a number of different a_i. Each a_i is associated with a set of states S_i composed of verifiable pre-conditions s_pre and post-conditions s_post. D is inconsistent when an action a_i exists, associated with its S_i, where either s_pre and/or s_post are incompatible with respect to the S set belonging to another a_j, as they cannot co-exist. When this conflict takes place, an inconsistency occurs. This conflict can depend on i) a s_pre which is incompatible with the rules of the Communal Common Ground (CCG)[1] (i.e., *cut the milk*), ii) the incompatibility of s_pre of the current a with s_post resulting from a preceding a, saved in the set of shared knowledge - the Personal Common Ground (PCG)[2]. Clarification requests can be in this case adopted as a corrective feedback.

[1] The amount of information shared with people that belong to the same community (Clark, 2015)

[2] The amount of information collected over time through communicative exchanges with an interlocutor(Clark, 2015)

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

3 Conflict Search Graph

The Conflict Search Graph allows to represent dialogue history and connect it with domain knowledge, so that CG stability checks and dialogue state tracking can be represented in the form of graph queries. From a formal point of view, dialogue states are defined by extending the concept of D as a sequence of actions. The aim of this module is to have a structured resource where the domain knowledge (part of the CCG) is stored, and whose conflict search module can be used to signal which input does not respect the rules of the CCG and cannot become part of the PCG. In fact, the graph is not just used to represent the domain and its rules: it also supports the automatic process of recognising Common Ground Inconsistencies. In other words, it is used to store the dialogue history so that inconsistencies caused by post-conditions applied by previous actions guide the identification of the potential source of the current inconsistency. *Pre-conditions* of an action describe the configurations of the CG that are compatible with action instancing. On the other hand, *post-conditions* are the graph updates applied after an action has been accepted in the PCG. When a post-condition resulting from a previous action clashes with a pre-condition of the current action an inconsistency occurs and a responsible action can be identified. Whereas the pre-conditions make aware of the possible presence of a conflict, the post-conditions help identify the conflicting action. The consistency checking process guides the adoption of linguistic feedback, such as Clarification Requests.

This module is represented as a (Neo4j-based (Webber, 2012)) graph[3] $D = \langle V, E \rangle$ where V is a set of vertices and E is a set of edges among the vertices in V. Edges are defined as functions between v_1 and v_2 where $v_1, v_2 \in V$. The edge is assumed to be oriented from v_1 to v_2. A *stable* CG is defined as a graph G where a set of stability checks, also based on frames pre-conditions, are all verified. A new candidate action to be included in the CG can be defined as a tuple $X = \langle a_n \rangle, < \bar{N}, \bar{E} >$ containing a new action a_n, a set of named entities $\bar{N}$ and a set of new edges $\bar{E}$. At any given time t, G_t represents the common ground configuration at t. Updating G by accepting X means creating a new graph $G' =< V', E' >$ where $V' = V \cup a_n \cup \bar{N}$ and $E' = E \cup \bar{E}$. G', can be accepted as an updated version of G only if G' is stable, so that:

$$G_{t+1} = G' \; if \; stable(g') \; else \; G$$

To verify that the Conflict Search Graph structure could actually detect inconsistencies to be, consequently, properly signalled, dedicated tests were carried out. For the test, 20 cooking recipes were used, in the form of lists of actions. For instance, the action *melt butter in a pan* was represented in a frame-based structure (Baker et al., 1998), as follows:

$$Apply\_heat \quad Food : butter; Container : pan$$

In each recipe, an erroneous action was inserted. While the command itself is acceptable by the time it is presented, it prevents the acceptability of a command appearing at least five steps later in the recipe, thus raising a conflict. The conflict was found **85%** of the times. In fact, for 3 recipes out of 20 the expected conflict action did not correspond to the one selected by the system. Nevertheless, the system outcomes cannot be considered as proper mistakes, as the system choices have reasonable explanations. For the *Pancakes* recipe, the expected conflict corresponded to *melt butter in a pan*, where *butter* was entirely used because no quantity was specified. The conflict is, therefore, triggered when the action *put the butter in the pan* is received in input, as the butter is no longer available. Nonetheless, the conflict was found at *add milk and butter to the yolks*. In fact, the unavailability of the ingredient can also be caused by the action of adding *butter* to other ingredients. In fact, when the system must select only one conflicting action, the most recent one is chosen. These results proved that the system could analyse pre-conditions rules correctly in a real context of use, even providing alternative views about the potential problems than the ones expected at design time.

4 Conclusions

Promising preliminary results collected in a simulated interaction scenario showed the potentiality of the Conflict Search graph for finding Common Ground Inconsistencies in dialogue. Starting from here, our purpose is to extend the experimentation of such a module in a real interactive scenario and to generalise the application of the graph. Other types of conflicts will also be investigated.

[3]The graph was built using data coming from Wikidata and FrameNet to represent the knowledge domain, part of the CCG; the PCG is, on the other hand, represented by the list of communicated actions. These are incrementally stored in the graph after running consistency checking queries.

References

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

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

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 '18, pages 1213–1221, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Eve V. Clark. 2015. Common ground. In *The Handbook of Language Emergence*, page 328–353. Wiley, Chichester, UK.

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

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 222–233, Washington, DC, USA. American Psychological Association.

David R Traum. 1999. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*, pages 124–131.

Douglas N Walton. 1984. Logical dialogue-games and fallacies.

Douglas N Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.

Jim Webber. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218.