



SEMDIAL 2019

LondonLogue

Proceedings of the 23rd Workshop
on the Semantics and Pragmatics of Dialogue
London, 4-6 September 2019

Christine Howes,
Julian Hough
and Casey Kennington (eds.)



ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

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

LondonLogue Website

<https://semdial2019.github.io/>

LondonLogue Sponsors



LondonLogue Endorsements



**The
Alan Turing
Institute**

Preface

LondonLogue brings the SemDial Workshop on the Semantics and Pragmatics of Dialogue to Queen Mary University of London for the first time, though back to London, as the 12th meeting – LONDIAL – took place at King’s College London in 2008. LondonLogue, 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 30 full paper submissions, 16 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 7 of the remaining submissions, together with 15 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 LondonLogue programme features three keynote presentations by Rose McCabe, Staffan Larsson and Sophie Scott. 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.

LondonLogue has received generous financial support from the Queen Mary University of London’s Institute of Applied Data Science (IADS). We have also been given endorsements by the ACL Special Interest Group SigDial and the Alan Turing Institute.

This year marks the changing of the guard at SemDial, and we would like to offer our utmost thanks and regards to our outgoing presidents, Raquel Fernández and David Schlangen who held the position from 2008-2018. We hope and believe that SemDial will continue to go from strength to strength under the stewardship of its new presidents Ellen Breitholtz and Julian Hough.

We are grateful to Casey Kennington for his stewardship on the challenge of making SemDial proceedings more indexable from this year onwards and to Chris Howes for her leadership and organisation in 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 Cognitive Science Group at Queen Mary University of London who have made LondonLogue possible. Special mentions go to our web-master and proceedings cover designer Janosch Haber and head of production Sophie Skach for bringing quality graphic design and user experience to the event. We would also like to thank Taste QMUL catering, Estates and Facilities and Events and Hospitality QM for their fantastic service on campus. Thanks to everyone who helped with all aspects of the organisation.

Christine Howes, Julian Hough and Casey Kennington

London

September 2019

Programme Committee

Christine Howes (chair)	University of Gothenburg
Julian Hough (chair)	Queen Mary University of London
Casey Kennington (chair)	Boise State University
Ellen Breitholtz	University of Gothenburg
Harry Bunt	Tilburg University
Mathilde Dargnat	Nancy University and ATILF-CNRS
Emilie Destruel	University of Iowa
Simon Dobnik	University of Gothenburg
Raquel Fernandez	University of Amsterdam
Kallirroi Georgila	University of Southern California
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	King's College London
Patrick G. T. Healey	Queen Mary University of London
Julie Hunter	Universitat Pompeu Fabra, Barcelona and Université Paul Sabatier, Toulouse
Amy Isard	University of Edinburgh
Ruth Kempson	Kings College London
Staffan Larsson	University of Gothenburg
Alex Lascarides	University of Edinburgh
Pierre Lison	Norwegian Computing Center
Gregory Mills	University of Groningen, Netherlands
Valeria de Paiva	Samsung Research America and University of Birmingham
Massimo Poesio	Queen Mary University of London
Laurent Prévot	Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309
Matthew Purver	Queen Mary University of London
James Pustejovsky	Computer Science Department, Brandeis University
Hannes Rieser	Bielefeld University
Mehrnoosh Sadrzadeh	University College London
David Schlangen	University of Potsdam
Mandy Simons	Carnegie Mellon University
Matthew Stone	Rutgers University
Grégoire Winterstein	Université du Québec à Montréal

Local Organizing Committee

Julian Hough (chair)	Queen Mary University of London
Janosch Haber (webmaster)	Queen Mary University of London
Sophie Skach (production)	Queen Mary University of London
Leshao Zhang	Queen Mary University of London
Shamila Nasreen	Queen Mary University of London
Ravi Shekhar	Queen Mary University of London
Zico Putra	Queen Mary University of London
Massimo Poesio	Queen Mary University of London
Tom Gurion	Queen Mary University of London
Alexandra Uma	Queen Mary University of London
Arkaitz Zubiaga	Queen Mary University of London
Patrick G.T. Healey	Queen Mary University of London
Frank Foerster	Queen Mary University of London
Mehrnoosh Sadrzadeh	University College London
Andrew Lewis-Smith	Queen Mary University of London
Gijs Wijnolds	Queen Mary University of London
Jorge del Bosque	Queen Mary University of London

Table of Contents

Invited Talks

Questions and Answers in Suicide Risk Assessment: A Conversation Analytic Perspective	2
<i>Rose McCabe</i>	
Meaning as Coordinated Compositional Classification	3
<i>Staffan Larsson</i>	
The Science of Laughter	4
<i>Sophie Scott</i>	

Oral Presentations

Coherence, Symbol Grounding and Interactive Task Learning	6
<i>Mattias Appelgren and Alex Lascarides</i>	
The Devil is in the Detail: A Magnifying Glass for the GuessWhich Visual Dialogue Game	15
<i>Alberto Testoni, Ravi Shekhar, Raquel Fernández and Raffaella Bernardi</i>	
Meet Up! A Corpus of Joint Activity Dialogues in a Visual Environment	25
<i>Nikolai Ilinykh, Sina Zarriß and David Schlangen</i>	
A Sigh of Positivity: An Annotation Scheme for Sighs in Dialogue	35
<i>Christopher Cash and Jonathan Ginzburg</i>	
Posture Shifts in Conversation: An Exploratory Study with Textile Sensors	44
<i>Sophie Skach and Patrick G. T. Healey</i>	
When Objecting to Presupposed Content Comes Easily	54
<i>Alexandra Lorson, Chris Cummins and Hannah Rohde</i>	
Implicatures in Continuation-based Dynamic Semantics	61
<i>Florrie Verity</i>	
A Framework for Annotating Co-working Dialogues in Complex Task Settings	70
<i>Emma Barker and Robert Gaizauskas</i>	
Good call! Grounding in a Directory Enquiries Corpus	79
<i>Christine Howes, Anastasia Bondarenko and Staffan Larsson</i>	
A Corpus Study on Questions, Responses and Misunderstanding Signals in Conversations with Alzheimer's Patients	89
<i>Shamila Nasreen, Matthew Purver and Julian Hough</i>	
How to Reject What in Dialogue	99
<i>Julian Schlöder and Raquel Fernández</i>	
The Status of Main Point Complement Clauses	109
<i>Mandy Simons</i>	
How to Put an Elephant in the Title: Modeling Humorous Incongruity with Topoi	118
<i>Ellen Breitholtz and Vladislav Maraev</i>	

Co-ordination of Head Nods: Asymmetries between Speakers and Listeners	127
<i>Leshao Zhang and Patrick G. T. Healey</i>	
Character Initiative in Dialogue Increases User Engagement and Rapport	136
<i>Usman Sohail, Carla Gordon, Ron Artstein and David Traum</i>	
Modeling Intent, Dialog Policies and Response Adaptation for Goal-Oriented Interactions	146
<i>Saurav Sahay, Shachi H. Kumar, Eda Okur, Haroon Syed and Lama Nachman</i>	
Poster Presentations	
Analysis of Satisfaction and Topics in Repeated Conversation through Days	157
<i>Tsunehiro Arimoto, Hiroaki Sugiyama, Masahiro Mizukami, Hiromi Narimatsu and Ryuichiro Higashinaka</i>	
On Visual Coreference Chains Resolution	159
<i>Simon Dobnik and Sharid Loáiciga</i>	
Rezonator: Visualizing Resonance for Coherence in Dialogue	162
<i>John Dubois</i>	
Within and Between Speaker Transitions in Multiparty Casual Conversation	165
<i>Emer Gilmartin and Carl Vogel</i>	
A Wizard of Oz Data Collection Framework for Internet of Things Dialogues	168
<i>Carla Gordon, Volodymyr Yanov, David Traum and Kallirroi Georgila</i>	
Normativity, Meaning Plasticity, and the Significance of Vector Space Semantics	171
<i>Eleni Gregoromichelaki, Christine Howes, Arash Eshghi, Ruth Kempson, Julian Hough, Mehrnoosh Sadrzadeh, Matthew Purver and Gijs Wijnholds</i>	
Comparing Cross Language Relevance vs Deep Neural Network Approaches to Corpus-based End-to-end Dialogue Systems	174
<i>Seyed Hossein Alavi, Anton Leuski and David Traum</i>	
Collection and Analysis of Meaningful Dialogue by Constructing a Movie Recommendation Dialogue System	177
<i>Takashi Kodama, Ribeka Tanaka and Sadao Kurohashi</i>	
Towards Finding Appropriate Responses to Multi-Intents - SPM: Sequential Prioritisation Model	180
<i>Jakob Landesberger and Ute Ehrlich</i>	
Tense Use in Dialogue	183
<i>Jos Tellings, Martijn van der Klis, Bert Le Bruyn and Henriëtte de Swart</i>	
Shared Gaze toward the Speaker and Grounding Acts in Native and Second Language Conversation	186
<i>Ichiro Umata, Koki Ijuin, Tsuneo Kato and Seiichi Yamamoto</i>	
A Taxonomy of Real-Life Questions and Answers in Dialogue	189
<i>Maxime Amblard, Maria Boritchev, Marta Carletti, Lea Dieudonat and Yi-Ting Tsai</i>	
Pattern Recognition is Not Enough: Representing Language, Action and Perception with Modular Neural Networks	192
<i>Simon Dobnik and John Kelleher</i>	

Investigating Variable Dependencies in Dialogue States	195
<i>Anh Duong Trinh, Robert Ross and John Kelleher</i>	
“What are you laughing at?” Incremental Processing of Laughter in Interaction	198
<i>Arash Eshghi, Vladislav Maraev, Christine Howes, Julian Hough and Chiara Mazzocchi</i>	
Exploring Lattice-based Models of Relevance in Dialogue for Questions and Implicatures	201
<i>Julian Hough and Andrew Lewis-Smith</i>	
Interactive Visual Grounding with Neural Networks	204
<i>José Miguel Cano Santín, Simon Dobnik and Mehdi Ghanimifard</i>	
Bouletic and Deontic Modality and Social Choice	207
<i>Sumiyo Nishiguchi</i>	
Towards a Formal Model of Word Meaning Negotiation	210
<i>Bill Noble, Asad Sayeed and Staffan Larsson</i>	
Towards Multimodal Understanding of Passenger-Vehicle Interactions in Autonomous Vehicles: Intent/Slot Recognition Utilizing Audio-Visual Data	213
<i>Eda Okur, Shachi H. Kumar, Saurav Sahay and Lama Nachman</i>	
Pronominal Ambiguity Resolution in Spanish Child Dialogue: A Corpus Based Developmental Language Acquisition Approach	216
<i>Martha Robinson</i>	
Eye Gaze in Interaction: Towards an Annotation Scheme for Dialogue	219
<i>Vidya Somashekarappa, Christine Howes and Asad Sayeed</i>	

Invited Talks

Questions and Answers in Suicide Risk Assessment: A Conversation Analytic Perspective

Rose McCabe
City, University of London
`rose.mccabe@city.ac.uk`

There are no physical tests or signs of suicide. Professionals assess risk of suicide in face-to-face contacts with people. The U.K. National Confidential Inquiry into Suicide (2016) found that professionals judged immediate risk of suicide at the patient's final appointment before death to be low or not present in 85% of deaths by suicide. A number of studies have found that, prior to death, patients do not communicate suicidal ideation/thoughts, "deny" suicidal ideation and are classified as low risk.

This talk will explore how suicide risk is assessed in professional-patient interaction. Using conversation analysis to investigate question polarity and preference for agreeing responses, it will focus on how questions are designed by professionals and how the design of the question impacts on patient responses. Data come from primary care, accident and emergency departments and secondary mental health care settings.

Firstly, professionals always ask closed yes/no questions when assessing suicide risk. This puts strong constraints on the patient's response to answer with a brief yes or no. Secondly, subtle differences in the wording of the question invite either a yes or a no response. Professionals tend to invite patients to confirm they are not feeling suicidal through the use of negative polarity items and negative declarative questions. This significantly biases patients' responses towards reporting no suicidal ideation.

In cases where patients also completed self-report suicide measures, some patients reported thoughts of ending their lives although this was not elicited in the assessment with the professional. These findings shed some light on patients denying suicidal thoughts before taking their own life. Professionals may use negatively framed questions because of the institutional pressure to assess risk so that it becomes a 'tick box' exercise. Paradoxically, this makes the assessment unreliable. If patients do disclose suicidal thoughts, there can also be an increased workload (e.g. more paperwork if a patient needs to be admitted to hospital). This micro-analysis of questions in institutional interactions reveals subtle features of assessment which have significant consequences for people, where getting it right can be a matter of life and death.

Meaning as Coordinated Compositional Classification

Staffan Larsson
University of Gothenburg
`staffan.larsson@gu.se`

Here are some fundamental questions about linguistic meaning: What is it, and where does it come from? How is word meaning related to utterance meaning? How are the meanings of words and utterances related to the world and our perception of it? We are working towards a formal semantics that aims to provide answers to these and related questions, starting from the notion of situated interaction between agents, i.e., dialogue.

By interacting using language, agents coordinate on the meanings of linguistic expressions. The meanings of many expressions can be modeled as classifiers of real-world information. Expressions can be individual words, but they can also be phrases and sentences whose meanings are composed from the meanings of their constituents. To make formally explicit the notions of coordination, compositionality and classification, and to relate these notions to each other, we use TTR (a type theory with records).

The Science of Laughter

Sophie Scott
University College, London
`sophie.scott@ucl.ac.uk`

In this talk I will address the ways that laughter is used in human interactions. I will explore the evolutionary background to this, and also the neural control of laughter production, and the neural basis of laughter perception. I will demonstrate candidate roles for spontaneous and communicative laughter, and explore the ways that we learn to process and understand these. I will end with a consideration of the roles for conversation and laughter in emotion regulation.

Oral Presentations

Coherence, Symbol Grounding and Interactive Task Learning

Mattias Appelgren

University of Edinburgh

M.R.Appelgren@sms.ed.ac.uk

Alex Lascarides

University of Edinburgh

alex@inf.ed.ac.uk

Abstract

To teach agents through natural language interaction, we need methods for updating the agent’s knowledge, given a teacher’s feedback. But natural language is ambiguous at many levels and so a major challenge is for the agent to disambiguate the intended message, given the signal and the context in which it’s uttered. In this paper we look at how coherence relations can be used to help disambiguate the teachers’ feedback and so contribute to the agent’s reasoning about how to solve their domain-level task. We conduct experiments where the agent must learn to build towers that comply with a set of rules, which the agent starts out ignorant of. It is also unaware of the concepts used to express the rules. We extend a model for learning these tasks which is based on coherence and show experimentally that our extensions can improve how fast the agent learns.

1 Introduction

Many commercial scenarios create planning problems consisting of goal conditions which are complex and vaguely specified. An example is problems created by Standard Operating Procedures (SOPs)—large manuals containing instructions and rules which workers must follow. In companies such as Amazon or Ocado these feature rules such as “make sure the box is properly sealed” or “never put frozen items in the same bag as meat products”.

Building a precise formal representation of such problems which supports inference and planning is a challenging task for two reasons. Firstly, the array of contingencies where SOPs apply may be so extensive that it is untenable for a domain expert to communicate all these possibilities to a software developer; and secondly, the SOPs often change in unforeseen ways (such as in bespoke

manufacturing or large online retail where product lines are highly dynamic), making previously irrelevant concepts become relevant. For example, a company that starts to sell batteries must ensure the labels are put to the left rather than right of the package (this is a SOP in Amazon (Personal Communication)). This spatial relation may not have been part of the original domain specification, but an agent that had to follow this rule would now have to refine their domain model to include it, and learn what the word “left” means.

Since communicating the current SOPs is difficult and they change periodically, it would be useful for the domain expert to be able to teach the agent personally, after the agent has been deployed. A natural way to do so is through a teacher-apprentice interaction where the teacher observes the apprentice attempting to complete the task, reacting when the apprentice performs actions inconsistent with the SOPs. This way of teaching is simpler on the teacher since it is easier to react to a situation than predicting all contingencies in advance. The apprentice, in this situation, must have the capacity to learn the constraints as well as new concepts which were not previously a part of their domain model.

In this paper we tackle a task which is analogous to, but simpler than, SOP compliant packing. Instead of rules referring to weight or fragility (“don’t put heavy things above eggs” or “protect the vase with bubble wrap because it is fragile”), in our task the agent must learn and reason about constraints in a blocks world where colour is a proxy for these concepts (e.g. “put red blocks on blue blocks”). The agent starts out with a domain model with no colour concepts, nor does it have any colour terms within its natural language vocabulary. It must learn from a teacher both the rules that constrain the task, and how to ground the previously unknown colour terms (which pop-

ulate the rules).

This work extends the task and agent in [Appelgren and Lascarides \(2019\)](#) where an agent learns from a simulated teacher’s corrective feedback. They build a graphical model that captures the semantics of correction. This allows the agent to learn to recognise colours and learn which constraints are a part of the goal. We address two shortcomings of their paper by: 1) utilising the evidence that the teacher has *not* corrected the agent’s latest action, and 2) extending the model to capture extended dialogue, allowing us to deal with anaphoric expressions, which are ubiquitous in spontaneous natural language interactions.

2 Related Work

Teaching agents through interaction is a central theme in areas such as Learning through Demonstration ([Argall et al., 2009](#)), advice giving ([Maclin and Shavlik, 1996](#); [Kuhlmann et al., 2004](#); [Benavent and Zanuttini, 2018](#)), and learning reward functions in Reinforcement Learning ([Christiano et al., 2017](#); [Hadfield-Menell et al., 2016](#)). However, the area that shares our goals most is Interactive Task Learning (ITL) ([Laird et al., 2017](#)).

ITL focuses on teaching agents the parameters or rules which govern a task, rather than optimising a known task (such as in Reinforcement Learning), through interaction with a teacher (e.g. ([Scheutz et al., 2017](#); [Lindes et al., 2017](#); [She et al., 2014](#); [Chai, 2018](#))). The main contribution of our work and of [Appelgren and Lascarides \(2019\)](#) is to extend the types of interaction which teachers perform beyond instructions and definitions, with a focus in this paper on correction and elaboration. Correction has only been studied with use of very simple language; e.g. “no” ([Nicolescu and Mataric, 2003](#)).

The goal in our task is to learn to identify valid sequential plans autonomously, as opposed to learning how to perform new actions by combining primitive actions ([Chai, 2018](#); [Scheutz et al., 2017](#)) or learning low level motor control directly ([Knox and Stone, 2009](#)). The agent must also refine its domain model with unforeseen concepts that are discovered through interaction, as opposed to having a full domain conceptualisation and needing only to learn to map language onto these known concepts (contra [Wang et al. \(2016\)](#); [Kuhlmann et al. \(2004\)](#)). To do this language grounding we follow an approach where individ-



Figure 1: The shades used for blocks within each colour category.

ual classifiers are trained for each concept ([Matuszek, 2018](#)).

3 Task

Agents must learn a tower building task in the blocks world. Each scenario consists of 10 coloured blocks that must be placed into a tower. The resulting tower must conform to a set of constraints, or rules, which are part of the task’s goal description, G . In this paper we consider rules of two forms:

$$r_1^{c_1, c_2} = \forall x. c_1(x) \rightarrow \exists y. c_2(y) \wedge on(x, y) \quad (1)$$

$$r_2^{c_1, c_2} = \forall y. c_2(y) \rightarrow \exists x. c_1(x) \wedge on(x, y) \quad (2)$$

where c_1 and c_2 are colours (e.g., red, blue, maroon).

The task is implemented in a virtual environment, where each scenario is defined in the Planning Domain Definition Language (PDDL). Agents interact with the world through the action $put(x, y)$, which simply places object x on object y . In each scenario, the agent must build a tower consistent with G . However, it begins ignorant of the specific constraints that define G . Further, the agent can see what blocks exist and their spatial relation to each other, but it is unaware of what colour terms are used to describe them. Instead, it only observes the RGB values of each block (henceforth referred to as $F(x)$). Additionally, the agent begins with no knowledge of what colour terms exist or what parts of the RGB spectrum divide into different colour terms (ie, it is unaware of the terms in Figure 1 and what disparate RGB values map to a particular concept). As such, the agent faces the problem of jointly learning: (a) the vocabulary of colour terms; (b) how to ground those terms in the embodied environment (i.e. finding a mapping from colour term

to the range of RGB values it denotes); and (c) the constraints on the goal G , in terms of those colours.

A teacher observes the agent attempting to build the towers. Every time the agent takes an action which breaks one of the rules in G (or which leads to a situation where a rule will inevitably be broken) the teacher provides verbal feedback. The feedback serves to correct the agent's mistake by providing an explanation as to why the action was incorrect. However, the verbal component may be ambiguous between several rules (see Section 4 for details). Thus, the agent must disambiguate the teacher's intended message while simultaneously learning to ground new terms in the embodied environment by learning the partition of RGB values into the concepts in Figure 1.

4 Coherence

The agent must learn the task by exploiting evidence supplied by the teacher's dialogue actions. It does this by reasoning about how the teacher's utterance coherently connects to the context in which it was uttered. To simplify matters we assume that all the teacher's dialogue moves are coherent, sincere (i.e. she believes what she says) and competent (i.e. what she believes is true).

The basic dialogue move the teacher makes is a correction of the form $u = \text{"no, put red blocks on blue blocks"}$ (or any other pair of colours). This utterance is ambiguous between rules $r_1^{red,blue}$ and $r_2^{red,blue}$ (henceforth shortened to $r_1^{r,b}$ and $r_2^{r,b}$). The semantics of correction stipulate that the content of the correction must negate some part of the corrected action (Asher and Lascarides, 2003). In our planning domain, this means that the teacher will utter u if the agent's latest action $a = put(x, y)$ violates the rule that she intended u to express, as stipulated in (3), where $V(r, a)$ represents that rule r was violated by action a :

$$Corr(a, u) \leftrightarrow (r_1^{r,b} \in G \wedge V(r_1^{r,b}, a)) \vee (r_2^{r,b} \in G \wedge V(r_2^{r,b}, a)) \quad (3)$$

The action a can violate a rule in the goal in two ways. For the first case, consider S_1 in Figure 2. If $r_1^{r,b} \in G$, then an action resulting in S_1 would directly violate the rule since $r_1^{r,b}$ requires each red block to be on a blue block, but here a red block was put on a non-blue block. Where

$a = put(o_1, o_2)$, this *Direct* violation is expressed as (4), and similarly S_2 directly violates $r_2^{r,b}$ because of (5):

$$V_D(r_1^{r,b}, a) \leftrightarrow red(o_1) \wedge \neg blue(o_2) \wedge on(o_1, o_2) \quad (4)$$

$$V_D(r_2^{r,b}, a) \leftrightarrow \neg red(o_1) \wedge blue(o_2) \wedge on(o_1, o_2) \quad (5)$$

$r_1^{r,b}$ is not directly violated in S_2 and $r_2^{r,b}$ is not directly violated in S_1 . However, these rules are respectively *Indirectly* violated: it is impossible to complete a rule-compliant tower without first removing the top block from it. That is, an indirect violation of $r_1^{r,b}$ means that there are more red blocks on the table than blue ones, and furthermore (given that it violates the latest action $put(o_1, o_2)$), this was not the case before this action, and therefore o_2 must be blue and o_2 not red. Formally, indirect violations of the rule $r_1^{r,b}$ (which is satisfied by S_2) and $r_2^{r,b}$ (which is satisfied by S_1) are respectively defined by (6) and (7):

$$V_I(r_1^{r,b}, a) \leftrightarrow \neg red(o_1) \wedge blue(o_2) \wedge on(o_1, o_2) \wedge |\{o_3 : red(o_3) \wedge on(o_3, table)\}| > |\{o_4 : blue(o_4) \wedge on(o_4, table)\}| \quad (6)$$

$$V_I(r_2^{r,b}, a) \leftrightarrow red(o_1) \wedge \neg blue(o_2) \wedge on(o_1, o_2) \wedge |\{o_3 : blue(o_3) \wedge on(o_3, table)\}| > |\{o_4 : blue(o_4) \wedge on(o_4, table)\}| \quad (7)$$

When uttering u , our teacher helps the agent to determine which type of violation has happened by pointing at the tower if it's a *Direct* violation V_D or pointing at the block which can no longer be placed in the tower if it's an *Indirect* violation V_I .

If the agent can ground either the colour term "red" and/or "blue" to blocks of those colours, then it can use the coherence equations (4)–(7) to infer whether the teacher's utterance u was intended to convey $r_1^{r,b}$, or $r_2^{r,b}$. Conversely, if an agent knows the intended meaning of u , then it can use these equations to make inferences about the colours of the blocks. However, our agent may know neither how to ground the colour terms (i.e., it can observe the RGB values but doesn't know what colour terms denote them) nor know how to

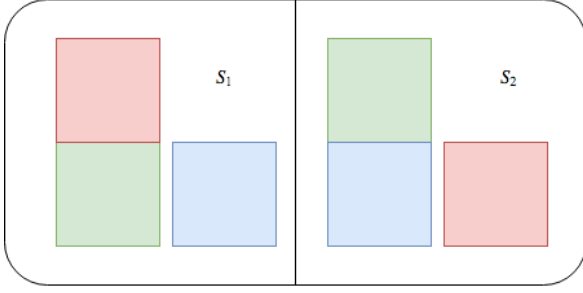


Figure 2: These two states would both be corrected if either $r_1^{(r,b)}$ or $r_2^{(r,b)}$ were in the goal.

disambiguate u . Therefore, in a context where the agent is sufficiently unsure as to the correct interpretation of the correction, because it is also unsure about how to ground the colour terms in u in the current visual scene, we allow the agent to utter a query, all of whose possible answers resolve the ambiguity. Due to the constraints expressed in (4)–(7), finding out the colour of just one of the blocks is sufficient to infer both the intended meaning of u and whether the situation is like S_1 or S_2 above. So the agent will simply ask such a yes/no question: for example, “is the top block red?”. The teacher’s answer provides the agent with an anchoring point, from which it can make further inferences via (4)–(7).

Additionally, when the teacher *doesn’t* correct the agent’s latest action a , then via the above equations, together with the agent’s current beliefs about which blocks are which colours, the agent can infer beliefs about which rules are *not* a part of the goal (on the grounds that if that rule had been in the goal, a correction of a would have been uttered).

Interpreting u only requires knowledge of the action a it corrects. However, certain utterances are only interpretable through their coherent connection to previous dialogue. In this paper, our teacher uses two such utterances: u_2 = “no, that is wrong for the same reason” and u_3 = “no, that is not red either”. u_2 presupposes a prior (identical) reason (in our task, a rule violation) is a part of the multimodal context; u_3 presupposes that something else (in the context) is not red.

In line with existing coherence-based theories of discourse (eg., Hobbs (1985); Kehler (2002); Asher and Lascarides (2003)) we assume that any utterance containing an anaphor or presupposition must be coherently connected to the unit that contains its antecedent. Thus u_2 (or u_3) must coher-

ently attach to more than just the agent’s latest action a ; it must also attach to a prior utterance—this is why starting a dialogue with u_2 or u_3 sounds anomalous. Constraints on which parts of an *embodied* dialogue context the current utterance can coherently connect to are not yet fully understood (though see (Hunter et al., 2018) for initial work). We therefore take a very permissive approach: in principle, u_2 (or u_3) can coherently attach to any prior dialogue move. However, in line with existing theories of discourse interpretation, we adopt a preference for attaching to the most recent utterance u that supports a coherent interpretation, and in particular resolves the anaphor. In other words, an utterance of the form u_2 or u_3 attaches with *correction* to the latest agent’s action a , but also to the most recent prior utterance u where a coherence relation $R(u, u_2)$ (or $R(u, u_3)$) can be established and an antecedent identified.

The utterance u_2 can be interpreted as an *elaboration* of any prior correction u : even if u were simply the expression “no”, thanks to (3) a violation can be accommodated as part of the content of u precisely because it corrects an agent’s (prior) action. Thus in embodied dialogue (1), u_2 attaches to a_2 with *correction* and also to u_1 with *elaboration* (because u_1 is more recent than u_0):

- (1) a. a_0 : $put(o_1, o_2)$
- b. u_0 : “No, put green blocks on orange blocks”
- c. a_1 : $put(o_3, o_4)$
- d. u_1 : “No, put red blocks on blue blocks”
- e. a_2 : $put(o_5, o_6)$
- f. u_2 : “No, that is wrong for the same reason”

The relation $elaboration(u_1, u_2)$ entails that however u_1 is disambiguated—ie, $r_1^{r,b}$, or $r_2^{r,b}$ —“the reason” in u_2 refers to the same rule. So a_1 and a_2 both violate the same rule, and so impose joint constraints on the colours of the four blocks o_3, o_4, o_5 and o_6 . This differs from the interpretation of a similar dialogue where the agent says u'_2 below:

- (2) a. a_1 : $put(o_3, o_4)$
- b. u_1 : “No, put red blocks on blue blocks”
- c. a_2 : $put(o_5, o_6)$
- d. u'_2 : “No, put red blocks on blue blocks”

u_2' doesn't feature any anaphoric expression, and so coherence does *not* demand that it be related to u_1 . Thus the ambiguities in u_1 and u_2 may resolve in different ways. This illustrates how anaphora can impose additional constraints on interpretation of both the linguistic and non-linguistic moves. Our model (Section 5) and experiments (Section 6) show that exploiting anaphora in the interaction helps the agent to learn faster.

The utterance u_3 = “that is not red either” requires an antecedent individual that's not red. With this in mind, consider dialogue (3):

- (3) a. $a_0: put(o_1, o_2)$
- b. u_0 : “No, put orange blocks on red blocks”
- c. $a_1: put(o_3, o_4)$
- d. u_1 : “No, put red blocks on blue blocks”
- e. $a_2: put(o_5, o_6)$
- f. u_2 : “No, put purple blocks on pink blocks”
- g. $a_3: put(o_7, o_8)$
- h. u_3 : “No, that is not red either”

The utterance u_3 corrects a_3 , and coherence demands that it also attach to a prior utterance that entails that something isn't red. It cannot attach to u_2 with elaboration or with any other relation: in particular, it cannot elaborate either of the rules that u_2 might express while at the same time violating a rule that's expressed in terms of red, which it must do given that u_3 corrects an action (i.e., a_3). On the other hand, if the agent's beliefs about the colours of o_3 and o_4 are consistent with resolving the ambiguity in u_1 to $r_2^{r,b}$, then by (5) this interpretation provides an antecedent that's not red—namely o_3 —and moreover it supports an elaboration relation between u_1 and u_3 . Thus discourse coherence results in u_3 attaching to u_1 with *elaboration*, u_1 gets resolved to mean $r_2^{r,b}$, and hence (via equation (5)) o_3 and o_7 are not red and o_4 and o_8 are blue.

5 Method

We build an agent which utilises coherence to learn from the teacher's feedback. The agent architecture is the same as in Appelgren and Lascarides (2019) except the model for learning from correction is replaced. Figure 3 shows an overview of the system. The main components are the action

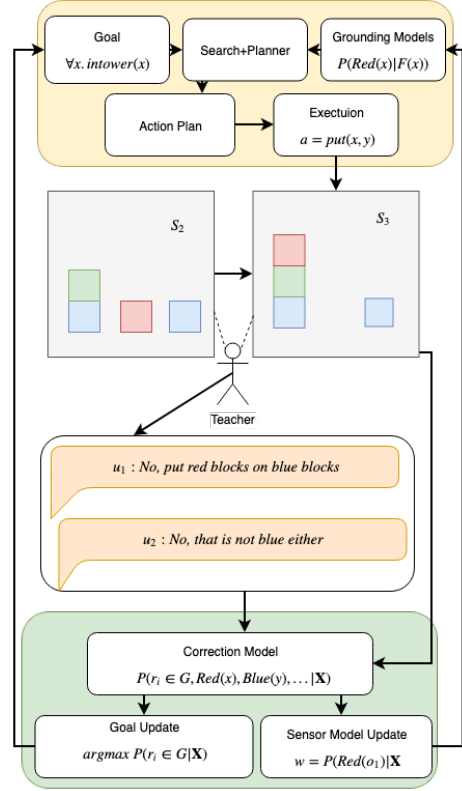


Figure 3: The agent consists of an action selection system (yellow) and a learning system (green). Action selection uses a symbolic planner to find a plan given the most likely goal and grounding of colour terms. The learning system uses coherence to build a probability model, used to learn what rules are in the goal and how to ground colour terms.

selection system, which makes use of a symbolic planner (Hoffmann and Nebel, 2001), a search strategy to find the most likely plan consistent with what has been learned so far, and the correction handling system, which learns from the dialogue.

The agent learns to find the most likely rules that are part of the goal G and learns classifiers for colour terms. The classifiers (which we call the grounding models) are binary classifiers for every relevant colour, such as $P(Red(x)|F(x))$ and $P(Blue(x)|F(x))$. These are estimated using Bayes Rule, utilising Weighted Kernel Density Estimation (KDE) (Gisbert, 2003) for estimating $P(F(x)|Red(x))$.

5.1 Learning from Dialogue moves

The agent must learn what the rules are and how to ground colour terms. To learn the rules the agent must resolve the ambiguity in the teacher's messages. To learn the colours the agent must associate the RGB values to colour words, thus creat-

ing training data. Both outcomes are linked, since disambiguation of messages leads to learning the rules and associates colour words to blocks, but resolving the ambiguity requires knowledge of colours.

To resolve the ambiguity and use the available evidence to learn, we build a probability model which captures the semantics of the dialogue, and how it links to the environment, by capturing equations (3)–(7) in a probabilistic graphical model. The model is updated dynamically each time a new dialogue move is performed, adding new factors and new evidence, rather than creating a new model for every correction as is done by Appelgren and Lascarides (2019).

Specifically, when a correction is uttered a set of nodes are added to the current model of the dialogue. As an example we shall use $u = \text{“no, put red blocks on blue blocks”}$ being directly violated. The nodes added from this correction can be seen in Figure 4. Here we know that equation (3) must hold, with rules $r_1^{r,b}$ or $r_2^{r,b}$. This is captured by adding a node $Corr(a, u)$, which is binary and observed to be *True*. Connected to this node are nodes $V_D(r_1^{r,b}, a)$, $V_D(r_2^{r,b}, a)$, $r_1^{r,b} \in G$, and $r_2^{r,b} \in G$. These are also binary variables, but they are latent. In the probability model this creates a factor

$$P(Corr_i(a, u) | V_D(r_1^{r,b}, a), V_D(r_2^{r,b}, a), r_1^{r,b} \in G, r_2^{r,b} \in G) \quad (8)$$

Which gives probability 1 to any outcome which satisfies equation (3).

For each $V_D(r_i^{r,b}, a)$, additional nodes are created to capture equations (4) and (5). The nodes capture the colour of the relevant objects: $Red(o_1)$ and $Blue(o_2)$. The probability distribution (9) is 1 whenever the values of the variables satisfy those in equations (4) and (5).

$$P(V_D(r_1^{r,b}, a) | Red(o_1), Blue(o_2)) \quad (9)$$

Since $Red(o_1)$ and $Blue(o_2)$ aren’t observable, nodes are also added for the observable RGB values of the objects: $F(o_1)$ and $F(o_2)$. $P(Red(o_1) | F(o_1))$ and $P(Blue(o_2) | F(o_2))$, which are the aforementioned grounding models, are learned using a weighted KDE. We also add priors for $P(r_i^{r,b} \in G)$ which are set to 0.01 and for $P(Red(o_1))$ and $P(Blue(o_2))$ which is set to 0.5.

The difference from Appelgren and Lascarides (2019) comes from the fact that when further corrections are given the model is updated by adding new nodes for the new correction and the possible violations it denotes. These nodes will be linked together if, for example, the same rule or the same blocks appear in the several corrections. This allows the agent to make inferences which change a belief from a previous correction given the new evidence. However, the biggest strength comes from modelling the interpretation of the anaphoric utterances, as discussed in Section 4.

5.1.1 Updating when no correction occurs

When a correction is given the agent adds nodes for the rules which are entailed by the content of the correction and observes $Corr(u, a) = True$. When no correction is given the agent instead adds nodes for all known rules and observes $Corr(u, a) = False$. That is, the agent adds a correction node which captures the fact that no rule which is in the goal was violated (through the negative case of equation (3)) as well as the nodes for direct violation of rules, capturing equations (4) and (5). Thus, the only non-zero probability interpretations of a non-corrected action are those which ensure these equations hold.

5.1.2 Handling Anaphoric Utterances

As discussed in Section 4, when an elaboration, such as $Elaboration(u_1, u_2)$ from dialogue (1), is given, the agent knows that the content of u_1 applies to the current action (a_2) and that the same rule must be violated by both a_1 and a_2 . Thus, the nodes which were added for u_1 and a_1 are also added for the action a_2 , as seen in Section 5.1. Further, an additional factor is added to capture that the same rule must be violated. This factor depends on $V(r_i, a_i)$ for the relevant rules and actions:

$$\phi(V(r_1^{r,b}, a_1), V(r_2^{r,b}, a_1), V(r_1^{r,b}, a_2), V(r_2^{r,b}, a_2)) \quad (10)$$

The factor gives a score of 0 to any situation where one of the rules, $r_i^{r,b}$, is violated for one of a_1 or a_2 but not the other, thus enforcing the constraint that the same rule must be violated in both situations.

When it comes to “no, that is not red either” (see dialogue (3)) the same applies. Further, we know that o_3 and o_7 are $\neg red$. The effect of this in our model is to add $\neg red(o_3)$ and $\neg red(o_7)$ as observed variables, whereas they would be latent

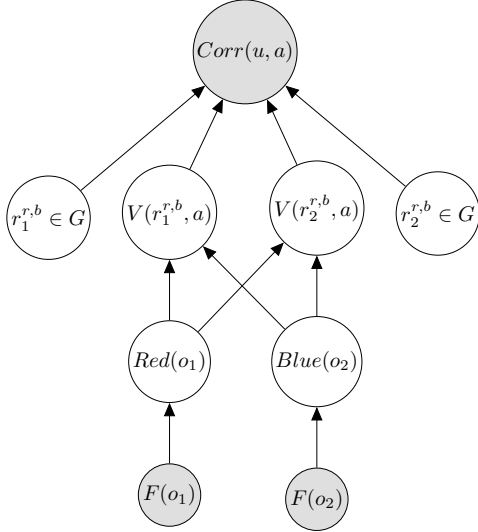


Figure 4: The nodes added to the probabilistic graphical model after a correction $u = \text{“no, put red blocks on blue blocks”}$. Grey nodes are observed and white ones are latent.

otherwise. Similarly, after a question is answered, the agent adds the ‘colour’ node’s value to the observed variables.

As we see, the structure of our model makes it straight forward to add new logical constraints, as imposed by symbolic semantic models of coherent discourse, by adding factors that force probabilities to zero when an interpretation is deemed by the symbolic semantic model to be incoherent.

5.2 Updating the Goal

The graphical model is used by the agent to estimate which rules are most likely to be in the goal. This is done by finding the rules which have the highest probability of being in the goal:

$$\operatorname{argmax} P(r_i \in G | \mathbf{X}) \quad (11)$$

where \mathbf{X} represents the available observations (including the RGB values, correction variables, and observed colour variables). Since r_i being in G is a binary decision, this means all rules which have a probability higher than 0.5 of being in the goal are added and the rest are not.

5.3 Updating the Grounding Models

To update the grounding models we seek labels for individual colours. Since our graphical model creates constraints on what colours blocks may have, we use the probability estimate as a soft label to update our model. For example let

$$w = P(\text{Red}(o_1) = \text{True} | \mathbf{X}) \quad (12)$$

Cumulative Regret for the Three Rules planning problem

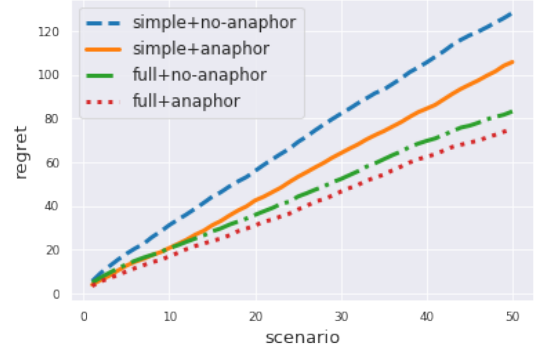


Figure 5: Cumulative regret averaged over 50 different trials on a planning problem containing three rules.

then w is used as a weighted label for o_1 which is used to update the probability density using weighted KDE.

6 Experiments

Our experiments test two hypotheses: exploiting anaphora when repeating a prior correction will lead to the agent learning to make fewer planning errors; and exploiting the evidence given by uncorrected moves will also lead to faster learning.

To test this we run four different set ups. First, we vary the teacher’s dialogue strategy between **anaphor**, in which she uses utterances like u_2 and u_3 in dialogues (1) and (4) when the agent repeats an error, and **no-anaphor**, in which even for the agent’s repeated errors, she repeats an utterance of the form $u = \text{“no, put red blocks on blue blocks”}$. Second, we vary the agent’s reasoning about the task, between **simple**, in which it updates its probabilistic model only when the teacher says something, vs. **full**, in which it updates its model every time it performs an action $put(x, y)$, taking the teacher’s silence to be a meaningful act as described in Section 5.1.1. The two types of teachers and agents gives four different combinations.

Each of these four combinations is run on 50 different planning problems—that is, we experiment with 50 different goals G , where each G is defined by two or three rules drawn from pairs of different colour categories and shades, such as red, purple, or maroon. Each planning problem (or G) is learned via a sequence of 50 scenarios: in each scenario the agent is presented with a (distinct) initial state of 10 blocks on the table, and the agent has to complete a tower that satisfies G , aided by the teacher’s feedback. The colour of the 10



Figure 6: Cumulative regret averaged over 50 different trials on a planning problem containing two rules.

blocks in each scenario is randomly generated by either selecting a completely random colour with probability 0.2 or (with probability 0.8) selecting a pair of colours present in one of the rules (e.g. red and blue for $r_1^{r,b}$), selecting randomly from the hues of those colours, which biases states to include many of the constrained colours. We filter out any scenarios for which no correct tower exists. To measure the effectiveness of the agent we measure regret, accumulated over the 50 scenarios. Regret is simply the number of mistakes the agent makes, i.e. the number of corrected actions it performs.

6.1 Results

We present results for experiments where each goal consists of two rules (Figure 6) and three rules (Figure 5).

Our hypothesis was that anaphors would help the agent make fewer mistakes; similarly for using the full evidence. Both of these results can be observed in the Figures 6 and 5. To test the significance of these results we performed a paired t-test on the total regret of each agent. The tests are made pairwise between agents using simple vs full, but keeping anaphor fixed, and between anaphor and no-anaphor, keeping simple vs full fixed. These significance tests are in Table 1.

These tests confirm that learning from the teacher’s silence, as well as from corrective moves, speeds up learning significantly. These benefits stem mainly from the ability to observe more training examples of colours, colour learning being the major bottleneck in this problem. The effects of anaphora on learning is more nuanced, however. The fact that exploiting anaphora sig-

	Two Rules	Three Rules
s/s+a	t=1.5, p=0.14	t=2.6, p=0.012
s/f	t=4.39, p=6.1e-5	t=4.4, p=6.1e-5
s+a/f+a	t=2.1, p=0.046	2.1, p=0.043
f/f+a	t=2.3, p=0.024	t=3.9, p=3.1e-4

Table 1: Results of t-test between combinations of simple (s) with and without anaphora (a) and full (f) with and without anaphora (the superior system in bold).

nificantly improves performance for the three-rule case, but in the two-rule case it is not quite significant, suggests that the more complex the (latent) goal, the more useful anaphora will be. A further issue concerning the utility of anaphora could also be linked to the way we constructed the 50 initial states for each planning problem (see earlier discussion), which does *not* guarantee that if $r_1^{r,b}$, say, is a rule in G , then the initial state contains at least two red blocks of a different hue and/or two blue blocks of a different hue.

7 Conclusion

We presented a novel graphical model which exploits the semantics of coherent discourse to jointly learn three tasks via natural language interaction with a teacher: how to refine the domain model to include new concepts; how to ground novel natural language terms to those concepts; and how to infer the correct goal description, so as to construct valid sequential plans. The graphical model extends on previous work by allowing it to learn from uncorrected moves in the dialogue as well as from utterances containing anaphoric expressions. Our experiments show that these extensions can help reduce the number of mistakes made by the agent while learning the task. In the future we intend to tackle more complex planning problems, featuring goal constraints with more complex structure, that are expressed in terms of unforeseen concepts other than colour. Additionally we intend to drop assumptions about the infallibility of the teacher.

Acknowledgements: We thank EPSRC for funding Mattias Appelgren, Ram Ramamoorthy and Yordan Hristov for helpful advice and discussions, and two anonymous reviewers for helpful feedback. All remaining errors are our own.

References

- Mattias Appelgren and Alex Lascarides. 2019. Learning plans by acquiring grounded linguistic meanings from corrections. In *In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13-17, 2019, IFAAMAS*, page 9 pages.
- Brenna Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. 57:469–483.
- Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.
- Florian Benavent and Bruno Zanuttini. 2018. An experimental study of advice in sequential decision-making under uncertainty. In *AAAI*.
- Joyce Yue Chai. 2018. Language to action: Towards interactive task learning with physical agents. In *AAMAS*.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4302–4310.
- Francisco J Goerlich Gisbert. 2003. Weighted samples, kernel density estimators and convergence. *Empirical Economics*, 28(2):335–351.
- Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart J. Russell. 2016. Cooperative inverse reinforcement learning. In *NIPS*.
- J. R. Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University.
- Jörg Hoffmann and Bernhard Nebel. 2001. The FF planning system: Fast plan generation through heuristic search. 14:253–302.
- Julia Hunter, Nicholas Asher, and Alex Lascarides. 2018. [A formal semantics for situated conversation](#). *Semantics and Pragmatics*.
- A. Kehler. 2002. *Coherence, Reference and the Theory of Grammar*. CSLI Publications, Cambridge University Press.
- W. Bradley Knox and Peter Stone. 2009. [Interactively shaping agents via human reinforcement: the TAMER framework](#). In *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009), September 1-4, 2009, Redondo Beach, California, USA*, pages 9–16.
- Gregory Kuhlmann, Peter Stone, Raymond J. Mooney, and Jude W. Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer.
- John E. Laird, Kevin A. Gluck, John R. Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario D. Salvucci, Matthias Scheutz, Andrea Lockerd Thomaz, J. Gregory Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive task learning. *IEEE Intelligent Systems*, 32:6–21.
- Peter Lindes, Aaron Mininger, James R. Kirk, and John E. Laird. 2017. Grounding language for interactive task learning. In *RoboNLP@ACL*.
- Richard Maclin and Jude W. Shavlik. 1996. Creating advice-taking reinforcement learners. *Machine Learning*, 22:251–281.
- Cynthia Matuszek. 2018. Grounded language learning: Where robotics and nlp meet. In *IJCAI*.
- Monica N. Nicolescu and Maja J. Mataric. 2003. [Natural methods for robot task learning: instructive demonstrations, generalization and practice](#). In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2003, July 14-18, 2003, Melbourne, Victoria, Australia, Proceedings*, pages 241–248.
- Matthias Scheutz, Evan A. Krause, Bradley Oosterveld, Tyler M. Frasca, and Robert Platt. 2017. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *AA-MAS*.
- Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Yue Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *SIGDIAL Conference*.
- Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. [Learning language games through interaction](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

The Devil is in the Details: A Magnifying Glass for the GuessWhich Visual Dialogue Game

Alberto Testoni

University of Trento
alberto.testoni@unitn.it

Raquel Fernández

University of Amsterdam
raquel.fernandez@uva.nl

Ravi Shekhar

Queen Mary University of London
r.shekhar@qmul.ac.uk

Raffaella Bernardi

University of Trento
raffaella.bernardi@unitn.it

Abstract

Grounded conversational agents are a fascinating research line on which important progress has been made lately thanks to the development of neural network models and to the release of visual dialogue datasets. The latter have been used to set visual dialogue games which are an interesting test bed to evaluate conversational agents. Researchers’ attention is on building models of increasing complexity, trained with computationally costly machine learning paradigms that lead to higher task success scores. In this paper, we take a step back: We use a rather simple neural network architecture and we scrutinize the GuessWhich task, the dataset, and the quality of the generated dialogues. We show that our simple Questioner agent reaches state-of-the-art performance, that the evaluation metric commonly used is too coarse to compare different models, and that high task success does not correspond to high quality of the dialogues. Our work shows the importance of running detailed analyses of the results to spot possible models’ weaknesses rather than aiming to outperform state-of-the-art scores.

1 Introduction

The development of conversational agents that ground language into visual information is a challenging problem that requires the integration of dialogue management skills with multimodal understanding. Recently, visual dialogue settings have entered the scene of the Machine Learning and Computer Vision communities thanks to the construction of visually-grounded human-human dialogue datasets (Mostafazadeh et al., 2017; Das et al., 2017a; de Vries et al., 2017) against which neural network models have been challenged. Artificial agents have been developed to learn either to ask or answer questions. Most of the work has focused on developing better Answerer agents, with a few exceptions (e.g., Manuvinakurike et al., 2017;

Zhang et al., 2018; Jiaping et al., 2018; Sang-Woo et al., 2019; Shekhar et al., 2019). Interesting and efficient machine learning methods (such as hierarchical co-attentions and adversarial learning) have been put at work to improve the Answerer agent (Lu et al., 2017b,a; S. and D., 2018; Kottur et al., 2018; Wu et al., 2018; Yang et al., 2019; Gan et al., 2019). Also when work has been proposed to highlight weaknesses of the available datasets, this has been done from the perspective of the Answerer (Masiceti et al., 2019). Much less is known about the Questioner agent, on which our work focuses.

The Questioner is evaluated through visually-grounded dialogue games like GuessWhat?! and GuessWhich introduced by de Vries et al. (2017) and Das et al. (2017b), respectively.¹ The two games share the idea of having two agents, a Questioner and an Answerer, playing together so that the Questioner, by asking questions to the Answerer, at the end of the game can make its guess on what is the object or which is the image they have been speaking about; however, the two games differ in many respects. Crucially in GuessWhich the Questioner sees a description (i.e., a caption) of the target image it will have to guess at the end of the game, but does not see any of the candidate images among which it has to select the target one (see Figure 1 for an example). Most, if not all, the work proposed for these two games heavily relies on Reinforcement Learning (RL).

The purpose of this work is to dive into the GuessWhich task and dataset through a simple Questioner model trained in a supervised setting, with a standard encoder-decoder architecture. The model learns to process the image caption and the dialogue history (the sequence of question-answer

¹The name GuessWhich has been used only lately by Chatopadhyay et al. (2017) to evaluate the Answerer agent playing the game with a Human. We take the liberty to use it for the game when played by two agents.

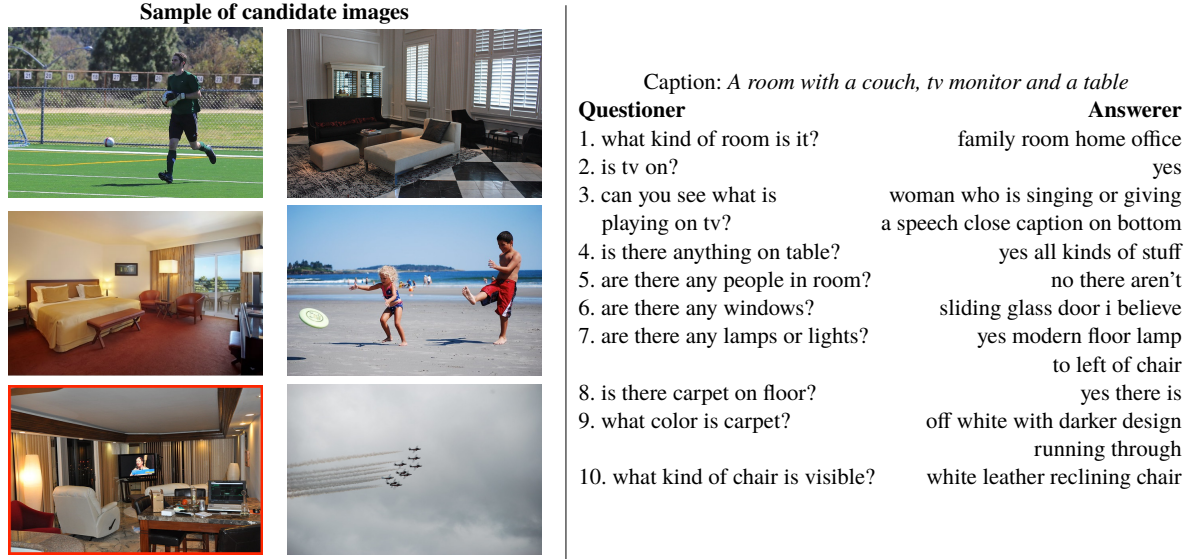


Figure 1: GuessWhich: two Bots are given a caption describing an image that one of the two bots (the answerer) sees while the other (the questioner) does not see. The Questioner has to ask 10 questions about the image and then select among about 10K candidates the image they have been speaking about. The dialogues given as example were generated by AMT workers, who were asked to chit-chat about the image, without having to select the target image at the end. The target image is the one on the left corner on the bottom, marked by the red box.

pairs), to generate questions, and to retrieve the target image at the end of the game by ranking the candidate images. We show that a simple model like ours outperforms state-of-the-art (SoA) models based on RL. Most importantly, by scrutinizing the model, we show that the SoA result obtained hides important weaknesses of the model and of the dataset:

- The question generator plays a rather minor role on task-success performance.
- The dialogues do not help much to guess the image, in the test phase. During training, they play the role of a *language incubator*, i.e., they help enrich the linguistic skills of the model, but the most informative linguistic input to guess the image is its caption.
- The distribution of game difficulty in the dataset is rather skewed: our simple model performs very well on half of the games, while half of the games appear to have issues that make them intrinsically difficult.

2 Related Work

Reinforcement Learning (RL) has become the default paradigm in visually-grounded dialogue. Strub et al. (2018) and Das et al. (2017b) show that RL improves the Questioner’s task success with respect to supervised learning (SL) in both GuessWhat?! and GuessWhich. Two crucial com-

ponents of the Questioner in visual dialogue guessing games are the question generator and the guesser. Shekhar et al. (2019) show that by training these two components jointly good performance can be achieved, and that a level of task success comparable to that attained by RL-based models can be reached by training the two modules cooperatively (i.e., with generated dialogues). Furthermore, Shekhar et al. (2019) show the linguistic poverty of the dialogues generated with RL methods, highlighting the importance of going beyond task success in evaluating visually-grounded Questioner agents. Inspired by this work, we study how far a simple model can go within the GuessWhich game and how the dialogue history is exploited in such a game.

Jiaping et al. (2018) propose a Questioner model based on hierarchical RL which, besides using RL to play the GuessWhich game, learns to decide when to stop asking questions and guess the image. In their approach, questions are retrieved (rather than generated) and the model is trained and evaluated on 20 pre-selected candidate images (instead of the full list of around 10K candidates as in the original game). A decision-making module has been introduced also by Shekhar et al. (2018), who train a discriminative model to play the GuessWhat?! game end-to-end without RL. In GuessWhat?!, the Questioner model has to identify a target object among 20 candidate objects

within an image. Thanks to the decider module, SoA results are achieved with shorter dialogues.

In the original GuessWhich game, the image has to be guessed among a very high number of candidates ($\sim 10k$); moreover, neither the target nor the other candidate images are seen during the dialogue. Hence the role of the decider module is vanished in such a setting, since the agent will never be sure to have gathered enough information to distinguish the target from the other images. As we focus on the original GuessWhich game, we do not include a decision-making module in our Questioner model. The number of questions is set to 10, as with the human players (see the next section).

Finally, a novel model is proposed by Sang-Woo et al. (2019), where the Questioner exploits a probabilistic calculus to select the question that brings about the most information gain. Their code has just been released. Hence, we leave for the future a thorough comparison with this approach.

3 Task and Dataset

We evaluate our model on the GuessWhich game proposed by Das et al. (2017b), which is based on the Visual Dialogue (VisDial) dataset by Das et al. (2017a).²

VisDial is the dataset used to play the GuessWhich game. It consists of 68K images from MS-COCO (Lin et al., 2014) of which 50,729 and 7663 are used for the training and validation set, respectively, and 9628 are used for the test set. There is no image overlap across the three sets. Each image is paired with one dialogue. The dialogues have been collected through Amazon Mechanical Turk (AMT) by asking subjects to chat in real-time about an image. The two AMT workers were assigned distinct roles: the questioner, who does not see the image but sees an MS-COCO caption of it, has to imagine the scene and ask questions about it; the answerer, who sees the image and the caption, has to answer the other player’s questions. The workers are allowed to end the chat after 10 rounds of question-answer pairs. An example of a dialogue by AMT workers is shown in Figure 1.

GuessWhich is a two-player game proposed by Das et al. (2017b). Two agents, Q-bot and A-Bot, have to play the role of the Questioner and the Answerer AMT workers in VisDial, but at the

end of the dialogue the Qbot has to guess the target image among a set of candidates (this task-oriented aspect was not present in the human data collection). The authors have released two versions of the test set: one with the original MS-COCO ground-truth captions and one with captions automatically generated with Neuraltalk (Karpathy and Fei-Fei, 2015) using the implementation by Vinyals and Le (2015). Usually, models are trained with the ground-truth captions and evaluated using the generated ones to check their robustness.

4 Models

We focus on developing a model of the Questioner agent. As the Answerer, we use the A-bot model by Das et al. (2017b) described below.

4.1 The Answerer Model

The A-Bot by Das et al. (2017b) is based on a Hierarchical Recurrent Encoder-Decoder neural network. It consists of three 2-layered LSTM encoders with 512-d hidden states and one LSTM decoder: A *question encoder* encodes the question asked by the Q-Bot; a *history encoder* takes, at each turn t , (i) the encoded question Q_t , (ii) the VGG image features (recall that the Answerer does see the image, unlike the Questioner), and (iii) the previous question-answer pair encodings to produce a state-embedding of the question being asked that is grounded on the image and contextualized over the dialogue history; an *answer decoder* takes the state encoding of the history encoder and generates an answer by sampling words; a *fact encoder* encodes the question-answer pairs.

The VGG features are obtained from a CNN pre-trained on ImageNet (Russakovsky et al., 2015). The vocabulary contains all tokens that occur at least 5 times in the training set; its size is 7,826 tokens. The model is trained with a cross-entropy loss. We use the code released in the authors’ Github page.

4.2 State of the Art Questioner Models

The Q-Bot by Das et al. (2017b) has a similar structure to the A-bot described above and shares its vocabulary, but it does not receive the image features as input. The goal of Q-Bot is to generate a question based on the caption and the dialogue history (the sequence of previous question-answer pairs). To this end, an encoder receives first the caption and then the question-answer pairs sequential-

²We use the version v0.5 available from the authors’ github at <https://github.com/batra-mlp-lab/visdial-rl>.

ly; it outputs a state-embedding at t that is jointly used by the decoder (an LSTM which learns to generate the next question) and by a Feature Regression Network (FRN, a fully connected layer which learns to approximate the visual vector of the target image). The decoder and the FRN are updated at every turn.

In the supervised learning (SL) phase, the two agents (A-Bot and Q-Bot) are separately trained under a Maximum Likelihood Estimation objective on the train set of VisDial human-human dialogues for 60 epochs. The FRN of the Q-Bot is trained to regress to the true image representation at each turn using Mean Square Error, i.e. l_2 loss. We will refer to this setting as Q-Bot-SL.

In the Reinforcement Learning (RL) phase, the Q-Bot and A-Bot are initialized by the models trained with SL for 15 epochs and then are fine-tuned with RL gradually by continuing SL for the first k rounds, and with RL for the $10 - k$ rounds, and annealing down k by 1 at every epoch. The authors have released the versions in which the model is trained with RL for 10 and 20 epochs. The reward is given to the two bots at each turn jointly. It is based on the change in distance (l_2) between the image representation produced by the FRN of Q-Bot and the true image vector before and after a round of dialogue. The total reward is a function only of the initial and final states. We will refer to this setting as Q-Bot-RL.

Recently, Sang-Woo et al. (2019) have proposed an interesting new model, AQM+, within the Answerer in Questioner’s Mind (AQM) framework introduced by Lee et al. (2017). Their Questioner asks questions based on an approximated probabilistic model of the Answerer, generating the question that gives the maximum information gain. The authors evaluate two versions of their model corresponding to the Q-Bot-SL and Q-Bot-RL settings described above: the two agents are trained (a) independently using human data (hence, AQM+/indA) or (b) together using the generated data (AQM+/depA).

4.3 Our Questioner Model

The architecture of our model is similar to the Q-Bot model of Das et al. (2017b) with two important differences: (i) the Encoder receives the caption at each turn, as it happens with humans who can reread the caption each time they ask a new question, and (ii) in the training phase, the

image regression module “sees” the visual vector of the target image only once, at the end of the game (i.e., as is the case for the human participants, there is no direct visual feedback during the dialogue).

As illustrated in Figure 2, in our model the Encoder receives two linguistic features: one for the caption and one for the dialogue. These features are obtained through two independent LSTM networks (Cap-LSTM and QA-LSTM) whose hidden states of 1024 dimensions are scaled through two linear layers to get linguistic features of 512-d. These two representations are passed to the Encoder: they are concatenated and scaled through a linear layer with a \tanh activation function. The final layer (viz. the dialogue state) is given as input to both the question decoder (QGen) and the Guesser module. QGen employs an LSTM network to generate the token sequence for each question. The Guesser module acts as a feature regression network (FRN): it takes as input the dialogue hidden state produced by the Encoder, and passes it through two linear layers with a ReLU activation function on the first layer. The final representation is a 4096-d vector which corresponds to the fc7 VGG representation of the target image. In contrast to the FRN by Das et al. (2017b), as mentioned above, our Guesser “sees” the ground-truth image only at the end of the game.

We use the same vocabulary as the A-Bot model. We apply the supervised training paradigm of Das et al. (2017b) and refer to our simple Questioner model as ReCap.

5 Experiment and Results

In this section, we present our experimental setup and report the results obtained, comparing them to the state of the art. We also analyse the role of the caption and the dialogue, as well the joint training regime on the performance of the model.

5.1 Evaluation Metrics and Implementation

Following Das et al. (2017b), we report the Mean Percentile Rank (MPR) of the target image, which is computed from the mean rank position of the target image among all the candidates. An MPR of e.g., 95% means that, on average, the target image is closer to the one chosen by the model than the 95% of the candidate images. Hence, in the VisDial test set with 9628 candidates, 95% MPR corresponds to a mean rank of 481.4, and a difference of $\pm 1\%$ MPR corresponds to ± 96.28 mean rank, which is a substantial difference. The chance level is 50.00

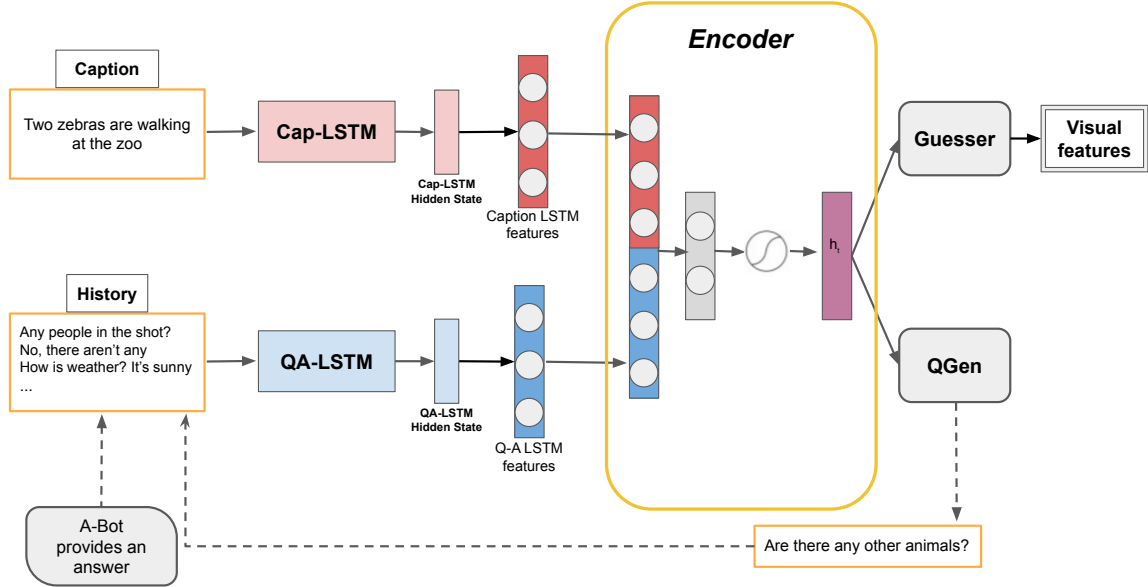


Figure 2: The ReCap questioner model: A simple encoder-decoder architecture that builds a hidden state of the dialogue by combining the representation of the caption and the dialogue; it rereads the caption at each turn while processing the dialogue history incrementally. The hidden state is used by the decoder (Question Generator) to generate the follow-up question at each turn, and by the Guesser to select the target image at the end of the game.

MPR, viz., 4814 mean rank position.

Our ReCap model has been trained for 41 epochs. Like Das et al. (2017b), our QGen and Guesser are trained jointly. However, following Shekhar et al. (2019), we use a modulo- n training regime, where n indicates after how many epochs of QGen training the Guesser is updated – we use $n = 5$. For the Q-Bot by Das et al. (2017b), we report the results we have obtained using the code released by the authors since they are higher than those reported in their paper.³

5.2 Comparison with SoA Models

Following Das et al. (2017b); Sang-Woo et al. (2019), we evaluate the models on the version of the test set containing captions generated with NeuralTalk2. As already shown by these authors, at test time, SoA models achieve rather good performance at round 0, i.e., just being exposed to the caption, without the dialogue history. For instance, the Q-Bot-SL trained on both captions and dialogues, when tested only on the caption achieves 89.11 MPR; in other words, it obtains just 2.08% less than what the same model achieves with the full 10-round dialogue. As we can see in Table 1, the

same holds for all the models we consider.

	MPR@0	MPR@10
Chance	50.00	50.00
Q-Bot-SL	89.11	91.19
Q-Bot-RL	95.72	94.19
AQM+/indA	88.50	94.64
AQM+/depA	88.50	97.45
ReCap	89.38	95.54

Table 1: Models tested with captions generated with NeuralTalk2. We evaluate the Mean Percentile Rank (MPR) of the models when receiving only the caption (at round 0) or the full dialogue (round 10). The results of the AQM model are from Sang-Woo et al. (2019).

Two things stand out regarding the performance of our model ReCap: First, although it is simpler, it obtains results 4.35% higher than Q-Bot-SL and comparable to Q-Bot-RL (ReCap +1.35%) as well as to the “supervised” version of the AQM model (ReCap +0.90%). Its performance is only lower than the more complex version of AQM (−1.91%). Second, our model appears to be able to exploit the dialogue beyond the caption to a larger degree than Q-Bot-SL and Q-Bot-RL, as evidenced by the larger difference between the results at round 0 and round 10.

5.3 Role of the Caption and the Dialogue

Given the results by Das et al. (2017b); Sang-Woo et al. (2019) with respect to the high performance obtained by the model at round 0 with just the

³In the authors’ github, there are various versions of the code: the QBot-RL model trained with 10 vs. 20 epochs, starting from the pre-trained Q-Bot-SL, and with and without optimizing the delta parameter. We use the code without the delta parameter, since it is the one explained in the paper, and with 20 epochs since it gives better results than the other one.

		GEN	GT
ReCap	MPR@0	89.38	87.95
	MPR@10	95.54	95.65
Q-Bot-SL	MPR@0	89.11	87.53
	MPR@10	91.19	89.00
Q-Bot-RL	MPR@0	95.72	94.84
	MPR@10	95.43	94.19

	MPR@10
Guesser caption	49.99
Guesser dialogue	49.99
Guesser caption + dialogue	94.92
Guesser+QGen	94.84
ReCap	95.65
Guesser-USE caption	96.90

Table 2: **Left:** Comparison of models performance when tested on generated (GEN) vs. ground truth (GT) captions **Right:** Ablation study of ReCap reporting MPR: We evaluate the Guesser when trained by receiving only the GT caption (Guesser caption); only the GT dialogues (Guesser dialogue) or both the GT caption and the GT dialogues (Guesser caption + dialogue). Furthermore, we report the results of QGen and Guesser trained separately (Guesser + QGen). Finally, Guesser-USE caption shows the MPR obtained by the Gusser when using pre-trained linguistic features.

caption, we aim to better understand the role of the captions and the dialogues in GuessWhich.

First of all, we check how the models behave on the ground truth captions (GT) of MS-COCO. As we can see from Table 2 (left), having the generated captions instead of the GT ones, facilitates the task (all models experience a gain in performance of around 1 to 2% MPR with GEN). However, at round 10, our model is somewhat more stable: it is less affected by the use of GT vs. generated captions than the other two versions of Q-Bot.

Secondly, we check whether the lack of improvement through the dialogue rounds is due to the quality of the dialogues. Therefore, we run the evaluation on the GT dialogues. Figure 3 reports the performance of our ReCap model when tested with the GT dialogues at each question-answer round, and compares it with the performances obtained by ReCap and the two Q-Bot models with the generated dialogues. As we can see, using the generated or GT dialogues does not affect ReCap’s performance very much: Also with human dialogues, after round 3 the performance does not increase significantly. Of course, these results do not say anything about the quality of the dialogues generated by the models, but they show that the per-round pattern common to all the models is not due to the linguistic quality of the dialogues.

Finally, we evaluate our Guesser trained and tested when receiving as input only the caption (Guesser caption), only the GT dialogues (Guesser dialogue) or both (Guesser caption and dialogue). Interestingly, as we can see from Table 2 (right), training the model only on the caption or only on the dialogue does not provide the Guesser with enough information to perform the task: it stays at chance level. Instead, training it with both types of linguistic in-

put doubles its performance (from 49.99 to 94.92). Based on these findings, we check also how the Guesser, trained and tested only on the caption, performs when the caption embedding is obtained using pre-trained and frozen linguistic features from the Universal Sentence Encoder (USE) by [Cer et al. \(2018\)](#) (Guesser-USE caption). This model reaches 96.90 MPR. This shows that in ReCap the caption and the dialogue play a complementary role: the caption provides a good description of the image but it is not enough by itself to train the model lexical knowledge; whereas the dialogues improve the linguistic knowledge of the encoder but do not provide a self-contained description of the image since they were produced as a follow-up to the image caption.

5.4 Role of the Joint Learning

By comparing the results in Table 1 and Table 2, we can see that QGen plays a rather minor role: already the Guesser alone reaches 94.92 MPR. Below we verify whether the multi-task setting in which the Guesser and QGen modules are trained improves the task success. [Shekhar et al. \(2019\)](#) show that the accuracy of a model that jointly learns to ask a question and guess the target object in the GuessWhat?! game obtains a 9% increase over its counterpart in which the two modules are trained separately. We check whether this result holds for the GuessWhich game too and compare ReCap with its counterpart with the two modules trained independently (Guesser+QGen). As we can see from Table 2, the joint training brings an increase of +0.81% MPR, viz. a lower increase than the one found in the GuessWhat?! game. We conjecture that this difference is due to the fact that the Guesser in GuessWhich does not have access to the distractor images during the dialogue, viz., the

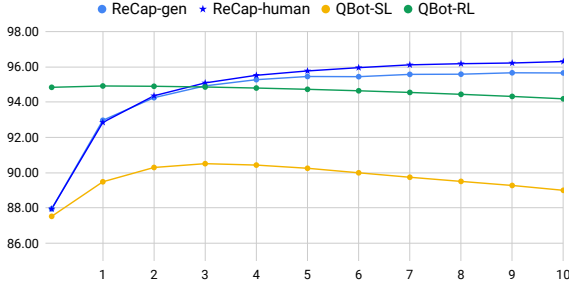


Figure 3: MPR distribution per dialogue round: comparison of ReCap model tested on human dialogues vs. ReCap and QBot models tested on generated dialogues.

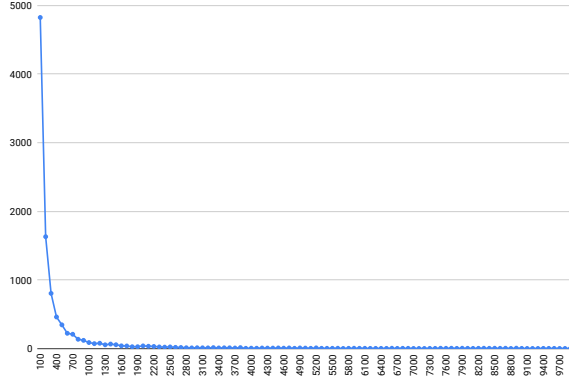


Figure 4: Distribution of rank assigned to the target image by ReCap tested on human dialogues. Each column aggregates 300 ranks.

candidate images that it has to learn to distinguish from the target image.

6 Analysis

To better understand why the dialogues do not help to rank the target image higher, below we further analyse the dataset.

Analysis of the Ranking For each of the 9628 images in the test set, we look into the ranks chosen by the ReCap model tested on the human dialogues. As shown Figure 4, the distribution is very skewed. On the one hand, in 126 games the target image has been ranked below chance level (below rank 4814); these images effect the MPR quite negatively. On the other hand, half of the games played by our simple model have a rank lower than 100, which means approximately 99 percentile or higher. Of these, 1032 are ranked above the 10th position.

Qualitative analysis of the 126 instances ranked below chance level has revealed that they are mostly cases of dialogues about images whose objects are hard to recognize, where the caption contains wrong information or unknown words (see examples in Figure 5.) Interestingly, Gusser-USE

caption has failed to rank high only 60 of these 126 outliers. For instance, the example in Figure 5 (up right) with the unknown word “roosters” is ranked at position 1082 by Guesser-USE caption and at 7006 by ReCap. As for the 1032 games with highly ranked target images, they concern images where the main objects are easily identifiable and are mentioned in the caption.

Analysis of the Visual Space To further understand the MPR results obtained by the models, we have carried out an analysis of the visual space of the candidate images. To check how the images in the high vs. low position in the rank differ, we have looked into their neighbourhood in the semantic space. We see that the highly ranked images have a denser neighbourhood than the ones ranked low, where density is defined as mean cosine distance between the image visual vector and its 20 closest neighbours. There is a 0.61 Spearman correlation, with p -value < 0.05 , between the rank of the retrieved image and the density of the neighbourhood.

Analysis of the Dialogues Following Shekhar et al. (2019), we look into the quality of the dialogues generated by the models by computing lexical diversity, measured as type/token ratio over all games; question diversity, measured as the percentage of unique questions over all games, and the percentage of games with at least one question repeated verbatim within the dialogue. Table 3 reports the statistics for our ReCap model and Q-bot. As we can see, ReCap produces a much richer and less repetitive output than both versions of Q-Bot. In particular, it has a more diverse vocabulary, generates more unique questions, and repeats questions within the same dialogue at a much lower rate. An illustration of these difference in the dialogue quality is provided by the example in Table 4.

	Lexical diversity	Question diversity	% Games with repeated Q's
Q-bot-SL	0.11	1.66	100
Q-bot-RL	0.05	0.35	100
ReCap	0.45	31.25	41.66
Humans	1.85	46.70	1.06

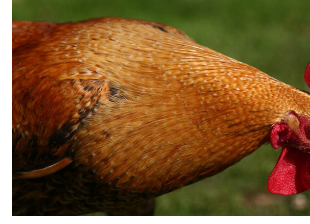
Table 3: Linguistic analysis of the generated dialogues. For lexical and question diversity higher scores are better, while lower scores are better for percentage of games with repeated questions.



A person cutting up a soccer ball cake.



A black and white image of a vehicle that is decorated like a dog.



This is a close up picture of a roosters neck



Two children are playing on the beach with a frisbee.



A room with a couch, tv monitor and a table.



A man that is on a soccer field with a ball.

Figure 5: **Top:** Images ranked lower than 8000 by ReCap: the objects are difficult to identified (left), the caption contains wrong information (middle) or unknown words (right); the wrong and the unknown words are underlined. **Bottom:** Images ranked higher than 10 by ReCap: the captions mention the main objects which are easily identifiable.

ReCap		QBot-SL		QBot-RL	
1. is tv flat screen?	no, it is small screen of some sort	1. is tv on?	no	1. any people?	no
2. does table have wood table?	yes, it does	2. is there couch?	yes	2. is it sunny?	yes
3. any people?	no people	3. what color is couch?	brown	3. is it sunny?	yes
4. is tv on?	no	4. is there a table?	yes	4. is it sunny?	yes
5. what color is couch?	brown	5. is there a table?	yes	5. is it sunny?	yes
6. are there any boots?	no, there aren't any	6. is there a table?	yes	6. is it sunny?	yes
7. can you see wintows?	no	7. is there a table?	yes	7. is it sunny?	yes
8. does table have wood on table?	no it doesn't	8. what color is table?	brown	8. is it sunny?	yes
9. what color are walls?	white	9. is there table?	yes	9. is it sunny?	yes
10. does table look clean?	not really	10. what color is table?	brown	10. is it sunny?	yes

Table 4: Examples of dialogues produced by the ReCap and Q-Bot models for the game about the image in Figure 5 (bottom, middle) which has been highly ranked by ReCap.

7 Conclusion

We have presented a simple model of the GuessWhich Questioner player. We have shown that it achieves SoA task-success scores. We have used this model as a magnifying glass to scrutinize the GuessWhich task and dataset aiming to further understand the model's results and, by so doing, to shed light on the task, the dataset, and the evaluation metric.

Our in-depth analysis shows that the dialogues play the role of a language incubator for the agent, i.e., they simply enrich its linguistic skills, and do not really help in guessing the target image. Furthermore, the difficulty distribution of the GuessWhich datapoints seems to be rather skewed: on the one hand, our model performs very well on half of the games; on the other hand, there

are outliers which have intrinsic difficulty and have a high impact on the final score. All this shows that the metric used in previous work to evaluate the models is way too coarse and obscures important aspects. Finally, we have shown that the linguistic quality of the dialogues produced by our simple model is substantially higher than that of the dialogues generated by SoA models.

References

- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*.
- Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrashekhara, Abhishek Das, Stefan Lee,

- Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *The fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (ICCV)*.
- Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. <https://arxiv.org/abs/1902.00579>.
- Zhang Jiaping, Zhao Tiancheng, and Yu Zhou. 2018. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceeding of the SigDial Conference*, pages 140–150. Association for Computational Linguistics.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of ECCV 2018*.
- Sang-Woo Lee, Yujung Heo, and Byoung-Tak Zhang. 2017. Answerer in questioner’s mind for goal-oriented visual dialogue. In *NIPS Workshop on Visually-Grounded Interaction and Language (ViGIL)*.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Dollar, P., and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV (European Conference on Computer Vision)*.
- Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017a. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS 2017*.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2017b. Hierarchical question-image co-attention for visual question answering. In *Conference on Neural Information Processing Systems (NIPS)*.
- Ramesh Manuvinakurike, David DeVault, and Kalliroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the SIGDIAL 2017 Conference*.
- Daniela Massiceti, Puneet K. Dokania, N. Siddharth, and Philip H.S. Torr. 2019. Visual dialogue without vision or dialogue. In *NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. [Image-grounded conversations: Multimodal context for natural question and response generation](#). In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.
- Zarrieß S. and Schlangen D. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of INLG 2018*.
- Lee Sang-Woo, Gao Tong, Yang Sohee, Yao Jaejun, and Ha Jung-Woo. 2019. Large-scale answerer in questioner’s mind for visual dialog question generation. In *ICLR*.
- Ravi Shekhar, Tim Baumgärtner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernández. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1218–1233.
- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and guess-what. In *NAACL*.
- Florian Strub, Mathieu Seurin, Ethan Perez, Harm de Vries, Jérémie Mary, Philippe Preux, Aaron Courville, and Olivier Pietquin. 2018. Visual reasoning with multi-hop feature modulation. In *Proceedings of ECCV*.
- Oriol Vinyals and Quoc V. Le. 2015. [A neural conversational model](#). In *Proceedings of the 31st International Conference on Machine Learning*.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! Visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of CVPR 2018*.
- Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. <https://arxiv.org/abs/1902.09326>.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018. Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)*.

Meet Up! A Corpus of Joint Activity Dialogues in a Visual Environment

Nikolai Ilinykh
Dialogue Systems Group
Bielefeld University

nikolai.ilinykh@uni-bielefeld.de

Sina Zarriß *
Digital Humanities
University of Jena

sina.zarriess@uni-jena.de

David Schlangen *
Computational Linguistics
University of Potsdam

david.schlangen@uni-potsdam.de

Abstract

Building computer systems that can converse about their visual environment is one of the oldest concerns of research in Artificial Intelligence and Computational Linguistics (see, for example, Winograd’s 1972 SHRDLU system). Only recently, however, have methods from computer vision and natural language processing become powerful enough to make this vision seem more attainable. Pushed especially by developments in computer vision, many data sets and collection environments have recently been published that bring together verbal interaction and visual processing. Here, we argue that these datasets tend to oversimplify the dialogue part, and we propose a task—MeetUp!—that requires both visual and conversational grounding, and that makes stronger demands on representations of the discourse. MeetUp! is a two-player coordination game where players move in a visual environment, with the objective of finding each other. To do so, they must talk about what they see, and achieve mutual understanding. We describe a data collection and show that the resulting dialogues indeed exhibit the dialogue phenomena of interest, while also challenging the language & vision aspect.

1 Introduction

In recent years, there has been an explosion of interest in language & vision in the NLP community, leading to systems and models able to ground the meaning of words and sentences in visual representations of their corresponding referents, e.g. work in object recognition (Szegedy et al., 2015), image captioning (Fang et al., 2015; Devlin et al., 2015; Chen and Lawrence Zitnick, 2015; Vinyals et al., 2015; Bernardi et al., 2016), referring expression resolution and generation (Kazemzadeh et al., 2014; Mao et al., 2015; Yu et al., 2016;

Schlangen et al., 2016), multi-modal distributional semantics (Kiela and Bottou, 2014; Silberer and Lapata, 2014; Lazaridou et al., 2015), and many others.

While these approaches focus entirely on visual grounding in a static setup, a range of recent initiatives have extended existing data sets and models to more interactive settings. Here, speakers do not only describe a single image or object in an isolated utterance, but engage in some type of multi-turn interaction to solve a given task (Das et al., 2017b; De Vries et al., 2017). In theory, these data sets should allow for more dynamic approaches to *grounding* in natural language interaction, where words or phrases do not simply have a static multi-modal meaning (as in existing models for distributional semantics, for instance), but, instead, where the meaning of an utterance is *negotiated* and *established* during interaction. Thus, ideally, these data sets should lead to models that combine visual grounding in the sense of Harnard (1990) and conversational grounding in the sense of Clark et al. (1991).

In practice, however, it turns out to be surprisingly difficult to come up with data collection setups that lead to interesting studies of both these aspects of grounding. Existing tasks still adopt a very rigid interaction protocol, where e.g. an asymmetric interaction between a question asker and a question answerer produces uniform sequences of question-answer pairs (as in the “Visual Dialogue” setting of Das et al. (2017b) for instance). Here, it is impossible to model e.g. turn-taking, clarification, collaborative utterance construction, which are typical phenomena of conversational grounding in interaction (Clark, 1996b). Others tasks follow the traditional idea of the *reference game* (Rosenberg and Cohen, 1964; Clark and Wilkes-Gibbs, 1986) in some way, but try to set up the game such that the referent can only be

*Work done while at Bielefeld University.

established in a sequence of turns (e.g. De Vries et al., 2017). While this approach leads to goal-oriented dialogue, the goal is still directly related to reference and visual grounding. However, realistic, every-day communication between human speakers rarely centers entirely around establishing reference. It has been argued in the literature that reference production radically changes if it is the primary goal of an interactive game, rather than embedded in a dialogue that tries to achieve a more high-level communicative goal (Stent, 2011).

Another strand of recent work extends the environments about which the language can talk to (simulated) 3D environments (Savva et al. (2019, 2017); see Byron et al. (2007) for an early precursor). On the language side, however, the tasks that have been proposed in these environments allow only limited interactivity (navigation, e.g. Anderson et al. (2018); Ma et al. (2019); question answering, Das et al. (2017a)).

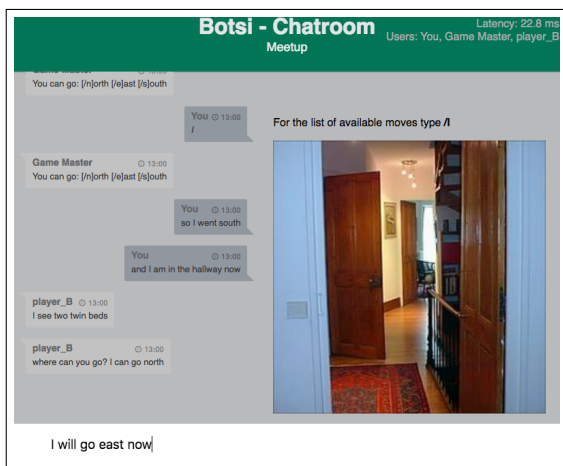


Figure 1: The game interface

What is lacking in these tasks is a real sense of the interaction being a *joint task* for which both participants are equally responsible, and, phrased more technically, any need for the participants to jointly attempt to track the dialogue state. In this paper, we propose a new task, MeetUp!, for visually grounded interaction, which is aimed at collecting conversations about and within a visual world, in a collaborative setting. (Figure 1 gives a view of the game interface and an excerpt of an ongoing interaction.)

Our setup extends recent efforts along three main dimensions: 1) the task’s main goal can be defined independently of reference, in high-level

communicative terms (namely “try to meet up in an unknown environment”), 2) the task is symmetric and does not need a rigid interaction protocol (there is no instruction giver/follower), 3) the requirement to *agree* on the game state (see below) ensures that the task is a true *joint activity* (Clark, 1996a), which in turn brings out opportunity for *meta-semantic* interaction and negotiation about perceptual classifications (“there is a mirror” – “hm, could it be a picture?”). This is an important phenomenon absent from all major current language & vision datasets.

This brings our dataset closer to those of unrestricted natural situated dialogue, e.g. (Anderson et al., 1991; Fernández and Schlangen, 2007; Tokunaga et al., 2012; Zarri   et al., 2016), while still affording us some control over the expected range of phenomena, following our design goal of creating a challenging, but not too challenging modelling resource. The crowd-sourced nature of the collection also allows us to create a resource that is an order of magnitude larger than those just mentioned.¹

We present our data collection of over 400 dialogues in this domain, providing an overview of the characteristics and an analysis of some occurring phenomena. Results indicate that the task leads to rich, natural and varied dialogue where speakers use a range of strategies to achieve communicative grounding. The data is available from <https://github.com/clp-research/meetup>.

2 The Meet Up Game

MeetUp! is a two-player coordination game. In the discrete version described here, it is played on a gameboard that can be formalised as a connected subgraph of a two-dimensional grid graph.² See Figure 2 for an example.

Players are located at vertices in the graph,

¹Haber et al. (2019) present a concurrently collected dataset that followed very similar aims (and is even larger); their setting however does not include any navigational aspects and concentrates on reaching agreement of whether images are shared between the participants or not.

²The game could also be realised in an environment that allows for continuous movement and possibly interaction with objects, for example as provided by the simulators discussed above. This would complicate the navigation and visual grounding aspects (bringing those more in line with the “vision-and-language navigation task”; (e.g. Anderson et al., 2018; Ma et al., 2019)), but not the coordination aspect. As our focus for now is on the latter, we begin with the discrete variant.

which we call “rooms”. Players never see a representation of the whole gameboard, they only see their current room (as an image). They also do not see each other’s location. The images representing rooms are of different types; here, different types of real-world scenes, such as “bathroom”, “garage”, etc., taken from the ADE20k corpus collected by Zhou et al. (2017). Players can move from room to room, if there is a connecting edge on the gameboard. On entering a room, the player is (privately) informed about the available exit directions as cardinal directions, e.g. “north”, “south”, etc., and (privately) shown the image that represents the room. Players move by issuing commands to the game; these are not shown to the other player.

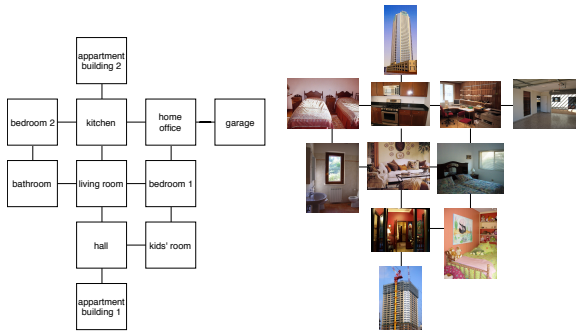


Figure 2: An abstract layout with room types (left), and a full gameboard with assigned images (right).

The goal of the players is to be in the same location, which means they also have to be aware of that fact. In the variant explored here, the goal is constrained in advance in that the meetup room has to be of a certain type previously announced to the players; e.g., a kitchen. The players can communicate via text messages. As they do not see each other’s location, they have to describe the images they see to ascertain whether or not they are currently in the same room, and move to a different room if they decide that they aren’t. If they have reached the conclusion that they are, they can decide to end the game, which they do via a special command. If they are then indeed in the the same room, and it is of the target type, the game is counted as a success, of which they are informed. The gameboard can be arranged such that there is type-level ambiguity; for example, there may be more than one room of type “bedroom” (as in Figure 2).

The game as implemented does not impose strict turn taking on the players; however, mes-

sages are only shown to the other player once they are sent via pressing the return key, as is usual in chat tools. There is thus no possibility for perceptibly overlapping actions, but it may happen that both players have been typing at the same time and the message that is received second is not a response to the first.

To make this more concrete, and to explain our expectations with respect to phenomena and required capabilities, we show a realistic, but compressed and constructed example of an interaction in this domain in the following. We will discuss attested examples from our data collection further below.

- (1)
 - a. Game Master: You have to meet in a room of type *utility room*.
 - b. A: Hi. I’m in a bedroom with pink walls.
 - c. B: I seem to be in a kitchen.
 - d. A: I’ll go look for a utility room.
 - e. A (privately): *north*
 - f. A (privately): *west*
 - g. B (privately): *east*
 - h. A: Found a room with a washing machine. Is that a utility room?
 - i. B: Was wondering as well. Probably that’s what it is.
 - j. B: I’m in the pink bedroom now. I’ll come to you.
 - k. B (privately): *north*
 - l. B (privately): *west*
 - m. B: Poster above washing machine?
 - n. A: Mine has a mirror on the wall.
 - o. B: yeah, could be mirror. Plastic chair?
 - p. A: And laundry basket.
 - q. A: *done*
 - r. B: Same
 - s. B: *done*

In (1-a), the Game Master (realised as a software bot in the chat software) gives the type constraint for the meetup room, which sets up a **classification task** for the players, namely to identify rooms of this type. (1-b) and (1-c) illustrate a common strategy (as we will see below), which is to start the interaction by providing state information that potentially synchronises the mutual representations. This is done through the production of **high-level descriptions of the current room**; for which the agents must be capable of providing *scene categorisations*. (1-d) and (1-j) show, among other things, the **coordination of strategy**, by announcing plans for action. In (1-e) – (1-g), private navigation actions are performed, which here are both **epistemic actions** (changing the environment to change perceptual state) as well as **pragmatic actions** (task level actions that potentially advance towards the goal), in the

sense of Kirsh and Maglio (1994). (1-h) and (1-i), where the classification decision itself and its basis is discussed (“what is a utility room?”); and (1-m)–(1-o), where a classification decision is revised (*poster* to *mirror*), illustrate the potential for **meta-semantic interaction**. This is an important type of dialogue move (Schlangen, 2016), which is entirely absent from most other language and vision datasets and hence outside of the scope of models trained on them. (1-j), also illustrates the need for **discourse memory**, through the co-reference to the earlier mentioned room where A was at the start. Finally, (1-p) as reply to (1-o) shows how in conversational language, **dialogue acts** can be **performed indirectly**.

As we have illustrated with this constructed example, the expectation is that this domain challenges a wide range of capabilities; capabilities which so far have been captured separately (e.g., visual question answering, scene categorisation, navigation based on natural language commands, discourse co-reference), or not at all (discussion and revision of categorisation decisions). We will see in the next section whether this is borne out by the data.

3 Data Collection

To test our assumptions, and to later derive models for these phenomena, we collected a larger number of dialogues in this domain (430, to be precise). We realised the MeetUp game within the *slurk* chat-tool (Schlangen et al., 2018), deployed via the Amazon Mechanical Turk platform.

We constructed maps for the game in three steps. First, we create a *graph* through a random walk over a grid graph, constrained to creating 10 nodes. The nodes are then assigned room types, to form what we call a *layout*. We identified 48 categories from the ADE20k corpus that we deemed plausible to appear in a residential house setting, from which we designated 20 categories as possible (easy to name) target types and the remaining 28 as distractor types. Additionally, we identified 24 plausible outdoor scene types, from which we sampled for the leaf nodes. The full set is given in the Appendix. We designate one type per layout to be the target type; this type will be assigned to 4 nodes in the graph, to achieve type ambiguity and potentially trigger clarification phases. We then sample actual images from the appropriate ADE20k categories, to create the *gameboards*.

In a final step, we randomly draw separate starting positions for the players, such that both of the players start in rooms not of the target type. For each run of the game, we randomly create a new gameboard following this recipe.

We deployed the game as a web application, enlisting workers via the Mechanical Turk platform. After reading a short description of the game (similar to that at the beginning of Section 2, but explaining the interface in more detail), workers who accepted the task were transferred to a waiting area in our chat tool. If no other worker appeared within a set amount of time, they were dismissed (and paid for their waiting time). Otherwise, the pair of users was moved to another room in the chat tool and the game begun. Player were paid an amount of \$0.15 per minute (for a maximum of 5 minutes per game), with a bonus of \$0.10 for successfully finishing the game (as was explained from the start in the instruction, to provide an additional incentive).³

4 Results

4.1 Descriptive Statistics

Over a period of 4 weeks, we collected 547 plays of the game. Of these, 117 (21%) had to be discarded because one player left prematurely or technical problems occurred, which left us with 430 completed dialogues. Of these, 87% ended successfully (players indeed ending up in the same room, of the correct type), 10% ended with the players being in different rooms of the correct type; the remaining 3% ended with at least one player not even being in a room of the target type. Overall, we spent around \$700 on the data collection.

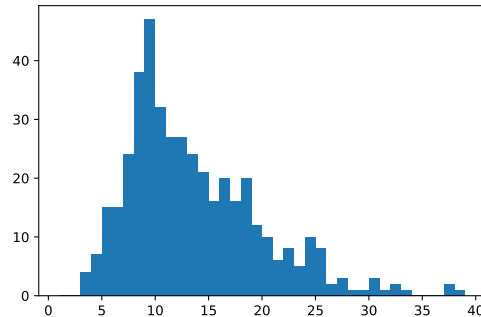


Figure 3: Histogram of number of turns per dialogue

³By the time of the conference, we will publish the code required to run this environment, as well as the data that we collected.

The average length of a dialogue was 13.2 turns (66.9 tokens), taking 165 seconds to produce. (The distribution of lengths is shown in Figure 3.) Altogether, we collected 5,695 turns, of an average length of 5.1 tokens. Over all dialogues, 2,983 word form types were introduced, leading to a type/token ratio of 0.10. The overlap of the vocabularies of the two players (intersection over union) ranged from none to 0.5, with a mean of 0.11.

On average, in each dialogue 28.3 navigation actions were performed. (Resulting in a MOVE/SAY ratio of a little over 2 to 1). The median time spent in a room was 12.2 secs. On average, each player visited 5.9 rooms without saying anything; when a player said something while in a room, they produced on average 3.5 turns. It hence seems that, as expected, players moved through some rooms without commenting on them, while spending more time in others.

We calculated the contribution ratio between the more talkative player and the less talkative one in each dialogue, which came out as 2.4 in terms of tokens, and 1.7 in terms of turns. This indicates that there was a tendency for one of the players to take a more active role. To provide a comparison, we calculated the same for the (role-asymmetric) MapTask dialogues (Anderson et al., 1991),⁴ finding a 2.8 token ratio and a 1.3 turn ratio.

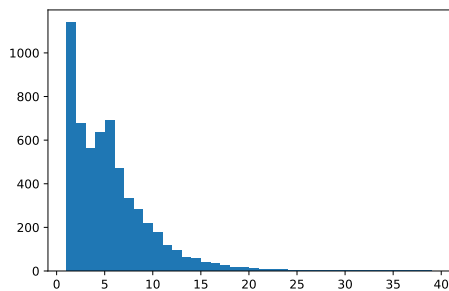


Figure 4: Histogram of number of tokens per turn

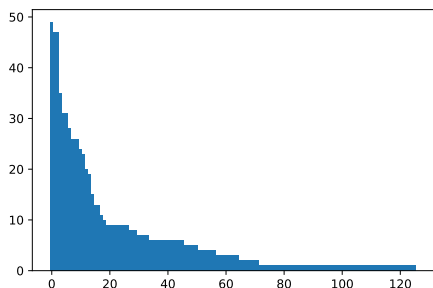


Figure 5: Number of Games Played, by Worker

Crosstalk occurs: On average, there are 1.4 in-

stances of one turn coming within two seconds or less than the previous one (which we arbitrarily set as the threshold for when a turn is likely not to be a reaction to the previous one, but rather has been concurrently prepared). The mean pause duration between turns of different speakers is 11.2 secs – with a high standard deviation of 9.46, however. This is due to the structure of the dialogues with phases of intense communicative activity, when a matching decision is made, and phases of individual silent navigation. If we only take transition times within the first 3 quartiles, the average transition time is 5.04 secs.

As Figure 4 indicates, most turns are rather short, but there is a substantial amount of turns that contain 4 or more tokens.

Figure 5 shows a frequency distribution of number of games played, by crowdworker. Overall, we had 126 distinct participants (as indicated by AMT ID). Our most prolific worker participated in 49 games, and the majority of workers played in more than one game. In only 22 games, two novices played with each other. In 81 games, there was one novice, and in 305 games, both players had played before. (For a few games we could not reconstruct the workerIDs for technical reasons, so this does not sum up to 430.)

4.2 Examples

Figure 6 shows a full interaction from the corpus. The public actions are represented in bold font, private actions are marked with “privately”, and responses by the Game Master are shown in italics. This example has many of the features illustrated with the constructed example (1) shown earlier. In lines 20 and 22, the players begin the game by providing high-level categorisations of their current positions, in effect synchronising their mutual game representations. Lines 22 and 23 then show coordination of game playing strategy. After separately moving around, individually solving the categorisation task—by moving through rooms that are not of the right type—the players resume interaction in lines 50ff. (with 50/51 showing crosstalk). Line 54 provides a justification for the negative conclusion from line 53, by providing information that contradicts l. 52. After more coordination of strategy in l.s 55 & 56, player B explicitly queries for more information. In line 75, player A justifies their positive verdict by confirming details previously given by B, extending it with

⁴Using the transcripts provided at <http://groups.inf.ed.ac.uk/maptask/maptasknxt.html>.





time	main	
10 00:03	GM (to A): url: /b/bathroom/ADE-train-00002739.jpg	
12 00:03	GM (to A): You can go: [/s]outh [/n]orth [/w]est	
14 00:03	GM (to A): You have to meet in: b/basement	
15 00:03	GM (to B): url: /a/apartment-building/outdoor/ADE-train-00001414.jpg	
17 00:04	GM (to B): You can go: [/s]outh	
19 00:04	GM (to B): You have to meet in: b/basement	
20 00:18	A: i am in the bathroom	
21 00:29	A: where are you at?	
22 00:33	B: I'm outside. Let's move around a bit.	
23 00:38	A: k	
24 00:40	B (privately): n	
25 00:41	GM (to B): Nothing happened.You can go: [/s]outh	
26 00:43	A (privately): s	
27 00:44	GM (to A): url: /b/bedroom/ADE-train-00003553.jpg	
29 00:44	GM (to A): You can go: [/s]outh [/n]orth [/w]est	
30 00:45	B (privately): s	
31 00:45	GM (to B): url: /b/basement/ADE-train-00002497.jpg	
33 00:45	GM (to B): You can go: [/s]outh [/n]orth [/e]ast	
34 00:50	A (privately): s	
35 00:50	GM (to A): url: /c/corridor/ADE-train-00006285.jpg	
37 00:51	GM (to A): You can go: [/n]orth [/e]ast	
38 00:52	B (privately): s	
39 00:52	GM (to B): url: //living-room/ADE-train-00000774.jpg	
41 00:53	GM (to B): You can go: [/s]outh [/n]orth [/e]ast	
42 00:57	B (privately): s	
43 00:57	GM (to B): url: /b/basement/ADE-train-00002482.jpg	
45 00:57	GM (to B): You can go: [/n]orth [/e]ast	
46 01:01	A (privately): e	
47 01:01	GM (to A): url: /b/basement/ADE-train-00002494.jpg	
49 01:01	GM (to A): You can go: [/e]ast [/w]est	
50 01:11	A: I am in the basement	
51 01:11	B: I'm in a basement.	
52 01:23	B: Mine has a white staircase	
53 01:28	A: no	
54 01:37	A: mine has wooden stair case	
55 01:55	B: Okay. Should I try to move towards you?	
56 02:09	A: Sure	
57 02:11	B: Wooden? What else?	
58 02:16	B (privately): e	
59 02:17	GM (to B): url: /b/bedroom/ADE-train-00003553.jpg	
61 02:17	GM (to B): You can go: [/s]outh [/n]orth [/w]est	
62 02:25	B (privately): s	
63 02:25	GM (to B): url: /c/corridor/ADE-train-00006285.jpg	
65 02:25	GM (to B): You can go: [/n]orth [/e]ast	
66 02:27	A: water heater and washer and dryer	
67 02:35	B (privately): s	
68 02:35	GM (to B): Nothing happened.You can go: [/n]orth [/e]ast	
69 02:40	B (privately): e	
70 02:40	GM (to B): url: /b/basement/ADE-train-00002494.jpg	
72 02:41	GM (to B): You can go: [/e]ast [/w]est	
73 02:42	A: a plastic chair and a screen door	
74 03:01	GM: Attention: you are in the game for 3 minutes	
75 03:18	B: I'm there! I see the water heater, washer and dryer, sink, chair laying on top of the screen door	
76 03:26	B (privately): done	
77 03:27	GM: The '/done' command has been issued by you or your partner. To end the game, both players need to enter '/done'	
78 03:27	A: yep	
79 03:32	A (privately): done	
80 03:34	GM: Well done Both of you are indeed in the same room of type: b/basement	

Figure 6: One Example Dialogue (mux36), with Images Overlayed

even more details. B confirms explicitly in 78, before also choosing SOLVE.

The excerpt from another dialogue in (2) shows an example of classification uncertainty being negotiated and dealt with.

(2) (Excerpt from mux39)

A: i think i am in a basement

B: i think i might be too.

A: maybe not though

A: wood panel?

A: two doors?

B: there's a tan couch, and a tan loveseat/chair. brown coffee table.

bar. tv
 B: nope, different room
 A: ok i am not there
 B: want me to meet you, or do you want to meet me?
 A: i think mine is more basement like
 B: okay. i'll try to find it.

4.3 Phases and Phenomena

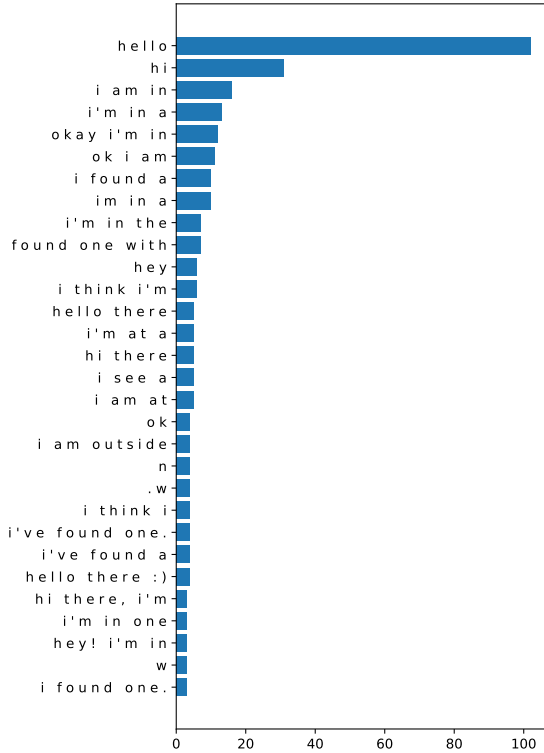


Figure 7: Prefixes of first turns

Figure 7 shows the most frequent beginnings of the very first turn in each dialogue. As this indicates, when not opening with a greeting, players naturally start by locating themselves (as in the example we showed in full). Figure 8 gives a similar view of the final turn, before the first *done* was issued. This shows that the game typically ends with an explicit mutual confirmation that the goal condition was reached, before this was indicated to the game.

What happens inbetween? Figure 9 shows the most frequent overall turn beginnings. As this illustrates, besides the frequent positive replies (“yes”, “ok”; indicating a substantial involvement of VQA-like interactions), the most frequent constructions seem to locate the speaker (“I’m in a”) or talk about objects (“I found a”, “there is a”, “is there a”). Using the presence of a question mark at

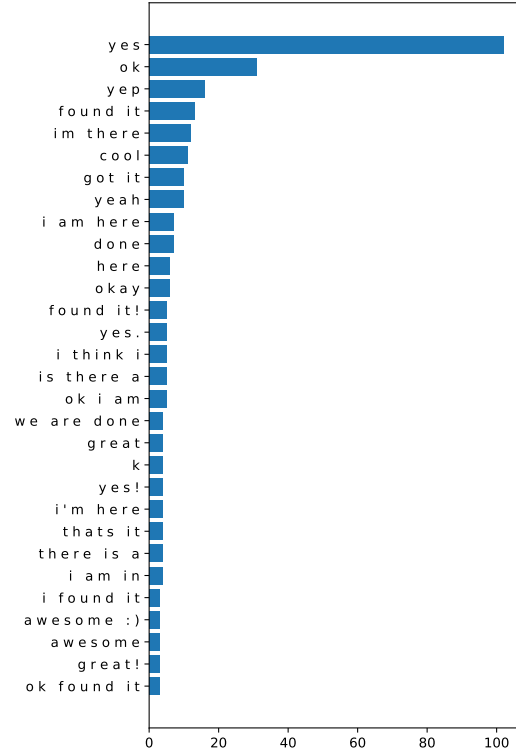


Figure 8: Prefixes of final turns (before *done*)

the end of the turn as a very rough proxy, we find 615 questions over all dialogues, which works out as 1.43 on average per dialogue. Taking only the successful dialogues into account, the number is slightly higher, at 1.48. Figure 10 shows the beginnings of these turns.

5 Modelling the Game

The main task of an agent playing this game can be modelled in the usual way of modelling agents in dynamic environments (Sutton and Barto, 1998), that is, as computing the best possible next action, given what has been experienced so far. The questions then are what the range of possible actions is, what the agent needs to remember about its experience, and what the criteria might be for selecting the best action.

In the action space, the clearest division is between actions that are directly observable by the other player—actions of type SAY—and actions that are targeted at changing the observable game state for the agent itself: actions of type MOVE and the END action. Since we did not restrict what the players could say, there is an infinite number of SAY actions (see Côté et al. (2018) for a formalisation of such an action space).

The total game state consists of the positions of

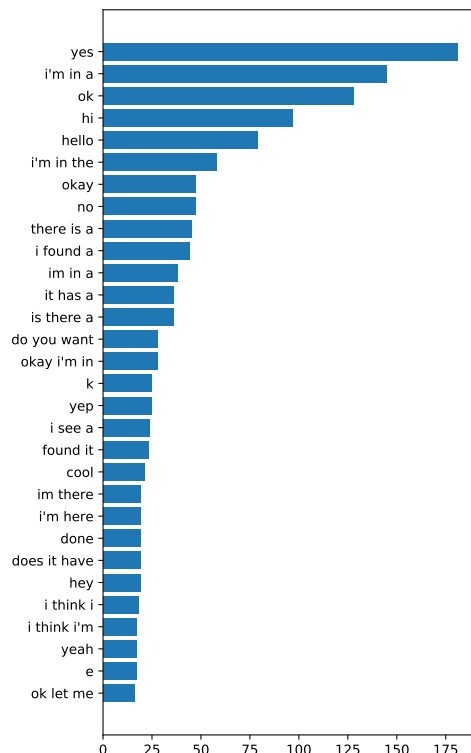


Figure 9: Most frequent turn beginnings

the players on the gameboard. Of this, however, only a part is directly accessible for either agent, which is their own current position. The topology of the network must be remembered from experience, if deemed to be relevant. (From observing the actions of the players in the recorded dialogues, it seems unlikely that they attempted to learn the map; they are however able to purposefully return to earlier visited rooms.) More importantly, the current position of the other player is only indirectly observable, through what they report about it. Finally, as we have seen in the examples above, the players often negotiate and agree on a current strategy (e.g., “I find you”, “you find me”, “we walk around”). As this guides mutual expectations of the players, this is also something that needs to be tracked. On the representation side, we can then assume that an agent will need to track a) their own history of walking through the map (raising interesting questions of how detailed such a representation needs to be or should be made; an artificial agent could help itself by storing the full image for later reference, which would presumably be not entirely plausible cognitively); b) what has been publicly said and hence could be antecedent to later co-references; c) what they infer about the other player’s position; and d)

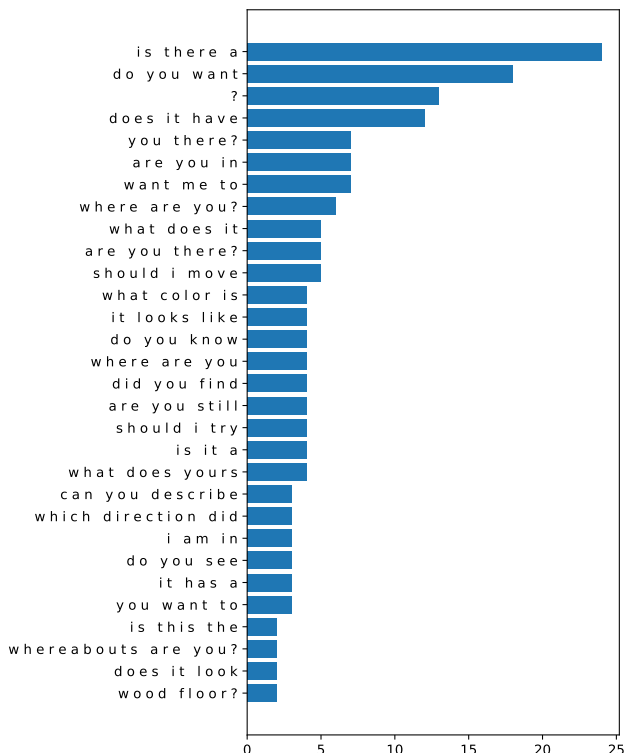


Figure 10: Prefixes of questions (utt.s ending in “?”)

what they assume the current agreed upon strategy is. This clearly is a challenging task; we will in future work first explore hybrid approaches that combine techniques from task-oriented dialogue modelling (Williams and Young, 2007; Buß and Schlangen, 2010) with more recent end-to-end approaches (Côté et al., 2018; Urbanek et al., 2019).

6 Conclusions

We have presented a novel situated dialogue task that brings together visual grounding (talking about objects in a scene), conversational grounding (reaching common ground), and discourse representation (talking about objects that were introduced into the discourse, but aren’t currently visible). An agent mastering this task will thus have to combine dialogue processing skills as well as language and vision skills. We hence hope that this task will lead to the further development of techniques that combine both. Our next step is to scale up the collection, to a size where modern machine learning methods can be brought to the task. Besides use in modelling, however, we also think that the corpus can be a valuable resource for linguistic investigations into the phenomenon of negotiating situational grounding.

A Room Types

1. Target room types: bathroom, bedroom, kitchen, basement, nursery, attic, child's room, playroom, dining room, home office, staircase, utility room, living room, jacuzzi/indoor, doorway/indoor, locker room, wine cellar/bottle storage, reading room, waiting room, balcony/interior
2. Distractor room types: home theater, storage room, hotel room, music studio, computer room, street, yard, tearoom, art studio, kindergarten classroom, sewing room, shower, veranda, breakroom, patio, garage/indoor, restroom/indoor, workroom, corridor, game room, pool room/home, cloakroom/room, closet, parlor, hallway, reception, carport/indoor, hunting lodge/indoor
3. Outdoor room types (nodes with a single entry point): garage/outdoor, apartment building/outdoor, jacuzzi/outdoor, doorway/outdoor, restroom/outdoor, swimming pool/outdoor, casino/outdoor, kiosk/outdoor, apse/outdoor, carport/outdoor, flea market/outdoor, chicken farm/outdoor, washhouse/outdoor, cloister/outdoor, diner/outdoor, kennel/outdoor, hunting lodge/outdoor, cathedral/outdoor, newsstand/outdoor, parking garage/outdoor, convenience store/outdoor, bistro/outdoor, inn/outdoor, library/outdoor

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 hrc map task corpus. *Language and speech*, 34(4):351–366.
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. [Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments](#). In *CVPR 2018*.
- Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. [Automatic description generation from images: A survey of models, datasets, and evaluation measures](#). *J. Artif. Int. Res.*, 55(1):409–442.
- Okko Buß and David Schlangen. 2010. Modelling sub-utterance phenomena in spoken dialogue systems. In *Proceedings of the 14th International Workshop on the Semantics and Pragmatics of Dialogue (Pozdial 2010)*, pages 33–41, Poznan, Poland.
- Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating instructions in virtual environments (give): A challenge and an evaluation testbed for nlg. *Position Papers*, page 3.
- Xinlei Chen and C Lawrence Zitnick. 2015. Mind’s eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431.
- Herbert H. Clark. 1996a. *Using Language*. Cambridge University Press, Cambridge.
- Herbert H Clark. 1996b. Using language. 1996. *Cambridge University Press: Cambridge*, pages 274–296.
- Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.
- Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.
- Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. [TextWorld: A Learning Environment for Text-based Games](#). *ArXiv*.
- Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017a. [Embodied question answering](#). *CoRR*, abs/1711.11543.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017b. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.
- Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.
- Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*, Boston, MA, USA. IEEE.
- Raquel Fernández and David Schlangen. 2007. Referring under restricted interactivity conditions. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 136–139, Antwerp, Belgium.
- Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In *Proceedings of the 2019 meeting of the Association for Computational Linguistics*, Florence, Italy.
- Stevan Harnard. 1990. The symbol grounding problem. *Physica D*, 42:335–346.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.
- Douwe Kiela and Léon Bottou. 2014. [Learning image embeddings using convolutional neural networks for improved multi-modal semantics](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar. Association for Computational Linguistics.
- David Kirsh and Paul Maglio. 1994. [On Distinguishing Epistemic from Pragmatic Action](#). *Cognitive Science*, 18(4):513–549.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. [Combining language and vision with a multimodal skip-gram model](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.
- Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019. [Self-Monitoring Navigation Agent via Auxiliary Progress Estimation](#). *ArXiv*, pages 1–18.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. [Genera-](#)

- tion and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.
- Seymour Rosenberg and Bertram D. Cohen. 1964. Speakers’ and Listeners’ Processes in a Word-Communication Task. *Science*, 145(3637):1201–1204.
- Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. 2017. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*.
- Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. *Habitat: A Platform for Embodied AI Research*. *ArXiv*.
- David Schlangen. 2016. Grounding, Justification, Adaptation: Towards Machines That Mean What They Say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.
- David Schlangen, Tim Diekmann, Nikolai Illykh, and Sina Zarriß. 2018. slurk – A Lightweight Interaction Server For Dialogue Experiments and Data Collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / semdial 2018)*.
- David Schlangen, Sina Zarriß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.
- Carina Silberer and Mirella Lapata. 2014. *Learning grounded meaning representations with autoencoders*. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics.
- Amanda J Stent. 2011. Computational approaches to the production of referring expressions: Dialog changes (almost) everything. In *PRE-CogSci Workshop*.
- Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning*. MIT Press, Cambridge, USA.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA.
- Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. 2012. The rex corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. *Learning to Speak and Act in a Fantasy Text Adventure Game*. *ArXiv*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.
- Jason Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):231–422.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing, Cham.
- Sina Zarriß, Julian Hough, Casey Kennington, Ramesh Manuvinaurike, David DeVault, Raquel Fernandez, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *10th edition of the Language Resources and Evaluation Conference*.
- Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A Sigh of Positivity: An Annotation Scheme for Sighs in Dialogue

Christopher Cash
Trinity College, Dublin
ccash@tcd.ie

Jonathan Ginzburg
Université de Paris, CNRS
Laboratoire de Linguistique Formelle
LabEx-EFL

yonatan.ginzburg@univ-paris-diderot.fr

Abstract

In this paper, an annotation scheme is developed to investigate the emotional quality of sighs in relation to three criteria; their placement in dialogue, their reason of expression and the emotion expressed by the sigh. Plutchik’s Wheel of Emotions is used to categorize the emotions and identify the extent of their arousal. We recognize two recurring kinds of sighs: those of low arousal and negative valence, and those of high arousal and positive valence. In particular, our results suggest the existence of cues indicating that a sigh is positive, as 75% of sighs *between pauses* were classified as positive and, moreover, when a sigh is classified as High arousal, there exists a 82% probability that the sigh will be positive. We conclude with a brief sketch of a formal analysis of sighs within the framework of KoS integrated with Scherer’s component process model of appraisal.

1 Introduction

Sighs are non-verbal vocalisations that can carry important information about a speaker’s emotional and psychological state (Truong et al., 2014). Their emotional quality is generally regarded as expressing a negative emotion, and are studied less than stereotypical positive utterances such as laughter and smiles. Truong et al. (2014) developed an annotation scheme for sighs which acknowledges their phonetic variations, in the hope of shedding light on the possibility of sighs expressing positive emotions. This scheme introduced two different sighs differentiated by whether or not they retained an audible inhalation and exhalation or just an audible exhalation. A basic finding was that not all sighs are produced in the same emotional context.

The physiological element of a sigh has been investigated thoroughly since Charles Darwin, and Straus (Straus, 1952) cites Darwin’s assertion that “the movements of expression in the face and

body, whatever their origin may have been, are in themselves of much importance for our welfare.” Much of Straus’ (1952) paper details the physiological element of the sigh, resulting in the conclusion that “the sigh, obviously, has no physiological causation.” Thus, we can understand sighing as the expression of emotion as it establishes a relation between a solitary individual and the world, as it is “a variation of the experiencing of breathing.” Bearing this in mind, we can discuss the expression of sighs in relation to theories of appraisal.

In cognitive theories of emotion, the general consensus is that emotions are caused by appraising events. Appraisal theories predict that emotions are elicited entirely on the basis of an individual’s subjective evaluations of the event (Oatley and Johnson-Laird, 2014). Scherer (Scherer, 2009a) formulates an appraisal theory which insists that an organism’s analysis of a situation correlates with their emotional reactions. Thus, it seems beneficial to analyse sighing in relation to a range of contextual criteria, such as investigating which person in a dialogue expresses the sigh and what is the topic of the dialogue— though this conflicts with Goffman’s influential theory that sighs are produced by spontaneous eruptions of negative emotions (Goffman, 1978). To evaluate whether a sigh could retain a positive connotation, Teigen (Teigen, 2008) conducted three studies. He claims that a prototypical sigh is a mismatch between ideals and reality writing that “the sigh accordingly carries two messages: One of discrepancy (something is wrong) and one of acceptance (there is nothing to be done)”. Hoey (Hoey, 2014), from a conversation analysis perspective, analyzes 54 sighs from the Santa Barbara Corpus of Spoken American English and from the Language Use and Social Interaction archive at the University of California, Santa Barbara. He distinguishes effects sighs have by position: speakers were found to use pre-beginning sighs for presaging the onset of talk

and indicating its possible valence; speakers used post-completion sighs for marking turns as being complete and displaying a (typically resigned) stance toward the talk. However, Hoey does not attempt any emotional analysis of the sighs.

Evidently, the categorisation of a sigh’s emotional valency is complex. This paper will explore Straus’ (1952) assertion that “If expressions are immediate variations of fundamental functions, every form of behaviour needs to be expressive”, by analysing the intensity of the emotion expressed by a sigh. By elaborating on two recent studies conducted to investigate the effects of emotional valence and arousal on respiratory variability during picture viewing and imagery, this paper will analyse the emotional quality of a sigh in relation to its context. (Vlemincx et al., 2015). Vlemincx et al.(2015), employed a method which separated emotions into dimensions and found that these dimensions yielded significantly different results, as fear imagery increased the expiration time of a sigh. These studies highlight the importance of analysing emotions with respect to a scale and found that high arousal emotions increase sigh rate, which contrasts with Straus’ theory. For this purpose, the employment of Plutchik’s Wheel of Emotions (Plutchik, 1988) establishes a consistent categorisation of emotions. Plutchik’s diagram (see Figure 8) highlights the existence of eight basic emotions, comprised of other emotions of various intensities. The use of this classification model can guide economists when developing a decision-making model (Gomes, 2017), highlighting the application it has for appraisal theories of emotions. It is the aim of this paper to develop a model in which the emotional quality of a sigh can be predicted.

An annotation scheme was developed iteratively and applied to a corpus consisting of conversation extracts—a sampling of the British National Corpus, thereby establishing contextual criteria for sigh classification. Using this methodology, it was found that sighs indicate a positive emotion more than other studies have accounted for.

The paper is structured as follows: A description of the corpus is given in Section 2, and the annotation scheme is clearly outlined in Section 3. An Analysis is conducted in Section 4, a Discussion is provided in Section 5, and a formal analysis is sketched in Section 6. Section 7 contains some

concluding remarks.

2 Material

One hundred samples of spoken dialogue were randomly selected for sigh annotation and analysis from spoken portion of the British National Corpus(British National Corpus). This was achieved using Matt Purver’s search engine SCoRE (Purver, 2001). Sighs are denoted as non-linguistic vocalisations in the corpus, and were added to the transcription scheme in 2014, as noted in the (British National Corpus 2014: User Manual and Reference Guide).

3 Development of the annotation scheme

Annotation guidelines were decided upon before analysis of samples. These guidelines focused on three dimensions;

- (1) Who produced the sigh and the vocal environment of the sigh
- (2) An interpretation of the reason for the sigh
- (3) An interpretation of the emotion expressed by the sigh by evaluating the first two dimensions.

The annotation was conducted by the first author. An inter-annotator study was conducted using annotation by a third individual, on the entire sample, for all three dimensions; three κ values were calculated.

3.1 Annotation Guidelines

First Dimension Guidelines

1. *Determine the Speaker and Addressee in the dialogue. The Speaker is defined as the person initiating the topic of conversation in the section of dialogue that is being analysed, and the Addressee is defined as the person who is perceived to be responding to the topic of conversation in the dialogue.*
2. *Determine the vocal environment of the sigh, focusing solely on the line of dialogue in which the sigh exits, by first distributing the data between two sets: Sighs expressed in relation to Speech and sighs expressed in relation to Pause. A pause in the corpus is indicated by “<pause>” and should only be taken into account if it immediately precedes or follows the sigh. Distribute this data into three subsets; Before, Between and After.*
 - Before: describing a sigh existing directly before speech or pause.*
 - Between: describing a sigh existing between two forms of speech or two indicated pauses.*
 - After: describing a sigh existing directly after speech or pause.*¹

¹If the sigh exists in a line of dialogue independently, distribute regularly into the set and subset however indicate that it refers to the other person’s speech or pause, looking at the lines of dialogue directly preceding and following.

Distinguishing between a Speaker and an Addressee yields interesting results when informed by their emotional valence. This is further elaborated on in Section 4.

As seen in Example 1, it is clear that Andrew is identified as the Speaker as he initiates the conversation.

Andrew: (SPK)
1. < sigh> Well we're keen to get here aren't we?
2. < pause> We're in the right place
I suppose? < pause> < unclear>
Anon 1: (ADR)
3. Mm. < pause dur=20>
Andrew:
4. Aha < pause dur=12> Well they'll be asking
the rest of us to take
a cut in salary soon < unclear>. < pause>
Anon 1:
5. < unclear> Well if I can < unclear> < laugh>

Example 1

Example 2 illustrates how a sigh is categorised as Between Pause.

Anon 3: (SPK)
133. Erm < pause> < sigh> < pause> (BTP) well I tried
for years to live with my second husband and it
just was impossible!
Anon 9: (ADR)
134. Mm.
Anon 3:
135. Not for just my own children but for my own health.
136. I'm now in a stable relationship with my fiancé and
it's fantastic!
137. What a difference!

Example 2

Second Dimension Guidelines Determine the reason for the sigh by analysing the entire excerpt of dialogue and proposing a category for the conversation. Building on the first dimension, it is clear from Example 3 that John is the Addressee, and the sigh is expressed *Before Speech* and the reason is denoted as *Answering*.

John: (SPK)
1594. Yeah I'll I'll check what I've got booked where
and then I'll I'll get in touch you for next week.
1595. Er
Andrew: (ADR)
1596. As long as it doesn't cause too much disruption
for you.
John:
1597. < sigh> (BS) It doesn't (Answering)

Example 3

Third Dimension Guidelines Determine the emotion expressed by the sigh using Plutchik's Wheel of Emotions:

1. Assign the sigh to one of the twenty four emotions of varying intensity in the model

2. Note which of the eight basic emotions it corresponds to,
3. Determine whether the basic emotion is positive or negative, based on the following partition: Anticipation, Joy, Surprise and Trust are positive, and Anger, Disgust, Fear and Sadness are negative.

Example 4 highlights how a sigh can be interpreted as positive, as the emotion expressed is categorised as *Joy*, and as *Neutral Intensity*.

Clare:
350. < laughing>: [That's the one, yes.] < laugh>.
Wendy:
351. < laugh>.
Clare: (SPK)
352. < sigh> (BS) After having hear his discourse on < pause>
the wonders of interchangeable brain chips and the lunar
landscape just above the ceiling border in thirty
[address] Road, I think he would probably be < pause>
quite a good candidate. (Discussing) (Joy)

Example 4

3.2 Results

3.2.1 First Dimension

The data indicates that 62% of the sighs were expressed by *Speakers* as opposed to 38% of the sighs which were expressed by the *Addressee*. This data highlights that a *Speaker* in a dialogue is more likely to express a sigh than the *Addressee*.

The placements of the sighs in the sample dialogues were then analysed with regard to the sub-sets *Before*, *Between* and *After*. This data is presented in Figure 1.

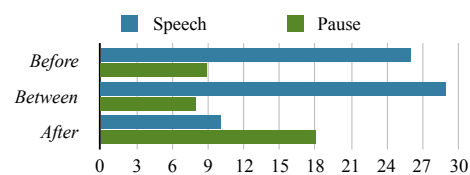


Figure 1: Placement of sigh in dialogue

Interestingly, the results indicate that 65% of the sighs were produced in relation to speech. Out of the 29 sighs produced *Between speech*, 41% of sighs were produced by a person during the other person's speech, as opposed to 59% of these sighs being produced during the person's own speech.

The results also indicated that 83% of the sighs produced between the other person's speech, were produced by an *Addressee* while the *Speaker* was speaking. Also, 11% of the sighs produced *After pauses* were produced after the other person's pause.

3.2.2 Second Dimension

The second dimension analysed the reason for the sigh. There were 29 possible reasons for sighs recorded, however only the reasons which received at least two entries were included in the data presented in Figure 2

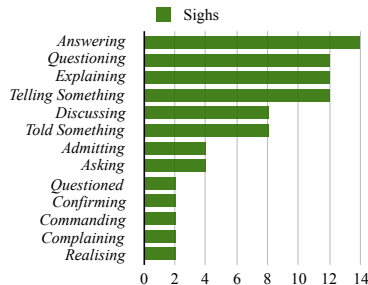


Figure 2: Reasons for expressing a sigh

These reasons were decided upon after analysis of dialogue, taking into account the words and expressions used. The results show that the most common reason for expressing a sigh is when *Answering*.

3.2.3 Third Dimension

The final dimension considered was the emotion expressed by the sigh. This data emerged through analysis of the data found in the previous two dimensions. The emotions were categorised into the twenty-four emotions, of varying intensity, on Plutchik's Wheel of Emotions, and then further categorised into the eight basic emotions, outlined in Figure 3. These eight emotions are grouped

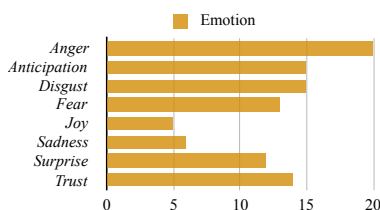


Figure 3: Classification of emotions expressed by sighs

together as polar counterparts in the pairings of *Joy/Sadness*, *Trust/Disgust*, *Anger/Fear* and *Anticipation/ Surprise*. This annotation scheme further distinguishes whether these emotions are either generally positive or negative. The positive emotions are *Anticipation*, *Joy*, *Surprise* and *Trust*, and the negative emotions are *Anger*, *Disgust*, *Fear* and *Sadness*. Interestingly, the results indicate that 46% of emotions recorded were positive, as opposed to 54% of emotions recorded as negative. The distribution of the sighs into the twenty-four

emotions of varying intensity is given in Figure 4. Interestingly, the majority of all emotions were

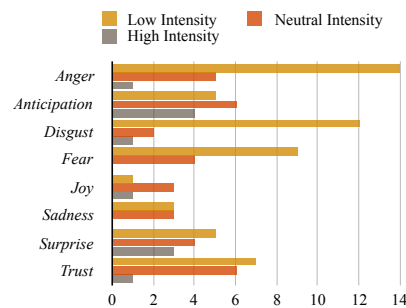


Figure 4: Distribution of emotional intensity

of *Low Intensity*, except *Anticipation* and *Joy*, in which the majority of these emotions were of *Neutral Intensity*. This indicates that it is more likely for a sigh that is expressing a positive emotion to be of a higher intensity.

3.3 Inter-Annotator Reliability

Following these guidelines, a third individual annotated 100% of the samples on all three dimensions. Cohen's Kappa Value was then computed from this inter-annotation scheme.

For the first dimension, a κ value of 0.52 was obtained for whether the sigh was expressed by a speaker or addressee. This highlights the difficulty in labelling participants when the full dialogue is not analysed. However, a κ value of 1 was obtained for the annotation of vocal environments, indicating that due to the guidelines provided by the Reference Guide (British National Corpus 2014: User Manual and Reference Guide), this dimension is deterministic as there are no discrepancies between annotators when analysing the vocal environment of a sigh.

For the second dimension, a κ value of 0.6 was obtained for analysing the reason for sighing, indicating a moderate level of agreement.

For the third dimension, a κ value of 0.62 was obtained for the emotion expressed by the sigh, suggesting that the inter-rater agreement for this study is moderately high. Interestingly, the majority of discrepancies occurred with respect to the classification of the basic emotions *Anger* and *Anticipation*. Out of the samples that were categorised as either expressing *Anger* and *Anticipation* by both annotators, it was found that in 33% of samples, the annotators disagreed about whether the emotion was *Anger* or *Anticipation*. This discrepancy could be accounted for

by the existence of eight extra emotions outlined in Plutchik’s model, which are combinations of two basic emotions. The basic emotions *Anger* and *Anticipation* exist beside each other on the wheel, and their combination emotion is *Aggressiveness*. Thus, the addition of these eight combination emotions could account for these discrepancies.

4 Analysis

As indicated in Figure 5, the results show that 50% of the sighs expressed by the speaker were positive, indicating that there is no efficient way of predicting the valence of a sigh when it is expressed by the speaker of a dialogue.

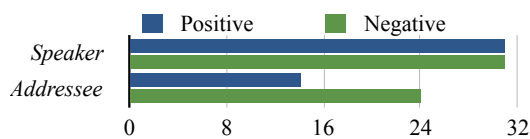


Figure 5: Distribution of sighs in all sound environments

This contrasts with the data recorded for the addressee, as the data indicates that 37% of the sighs expressed were positive. Subsequently we can deduce that it is more probable for an addressee to express a negative sigh than a speaker. Figure 6 distributes the sighs expressed between speech or pause into categories of positive and negative. It is clear that between speech, 41% of sighs are positive which contrasts with 75% of sighs expressed between pauses being positive. Thus, it is evident that it is more likely that a sigh expressed between a pause is positive than when it is expressed between speech. In Example 5, a sigh is ex-

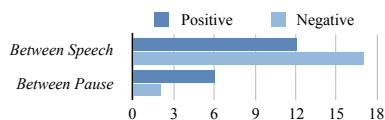


Figure 6: Distribution of sighs between speech or pause

pressed between a pause, which is positive as the emotion expressed is *Anticipation*. It’s clear that the dialogue directly preceding the pause and sigh is of a positive nature, indicating that the pause is used to establish equilibrium before asking a question. This example highlights the significance of pauses when analysing sighs. From these results it is clear that the emotion of a sigh is directly related to the vocal environment that it is

```
Terry: (SPK)
596. <singing>:[ I wanna to take you to outer space
<pause dur=7> outer space ].
597. <pause dur=17> Reggae Hits, thirteen.
598. <pause dur=12> <sigh> <pause> How long will
it be mum? (BTP) (Asking) (Anticipation)
Mother: (ADR)
599. What?
Terry:
600. How lo , how long are you gonna be?
Mother:
601. Up the park, er for dinner?
Terry
602. Yeah.
```

Example 5

expressed in. By focusing on the second dimension, and the motivations for expressing a sigh, the data is categorised into positive and negative emotions. It is clear that the highest recorded reasons for expressing a sigh were 1. *Answering*, 2. *Questioning*, *Explaining* & *Telling Something*, 3. *Discussing* & *Told Something* and 4. *Admitting* & *Asking*. These reasons were categorised by valency and this data is presented in Figure 7. In-

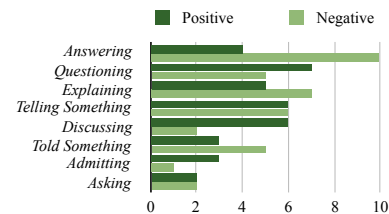


Figure 7: Distribution of the reasons for sighs by positive or negative emotions

terestingly, 29% of sighs expressed when *Answering* were positive, in comparison to the 58% of sighs when *Questioning*. Thus, for the most common reason it is clear that the majority of sighs expressed are negative. However, for the second most common reason, the data indicates positive values of 58%, 42% and 50%, respectively. This highlights the difficulty in predicting whether or not a sigh will be positive or negative based on the reason for the sigh. Strikingly, 75% of sighs expressed while discussing yielded a positive result, which provides an excellent probability score for future sigh interpretation. Example 6 provides an example in which the sigh is expressed while discussing money, categorised as expressing the emotion of *Acceptance*, which is *Trust* at a *Low Intensity*. Finally, by analysing the Third Dimension, the data presented in Figure 4 indicates that the most common emotion expressed by a sigh is *Anger*, found in twenty samples, followed by *An-*

Katriane: (SPK)
 175. What are you gonna do, go and tell them?
 <counting money>
 Sandy: (ADR)
 176. Give him a <pause> <unclear> this afternoon.
 177. Er <pause> tell him then.
 178. Twenty.
 179. I'll charge six from silver.
 Katriane:
 180. Mhm.
 Sandy:
 <sigh> (BTspkS) (Discussing) (Trust)
 Katriane:
 181. Six, six, seven

Example 6

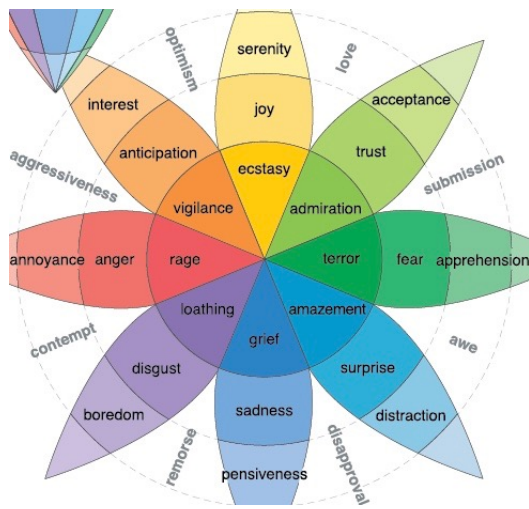


Figure 8: Plutchik's Wheel of Emotions

anticipation and *Disgust*, both found in fifteen samples respectively. Interestingly, this study found that 46% of sighs recorded were of a positive nature which contrasts with the other studies, such as Teigen(2008). Plutchik's model, displayed in Figure 8, illustrates how each of the eight basic emotions is constituted from two other emotions which exist at the extremes of the emotion spectrum. It is clear that *Annoyance* is a mild form of *Anger* whereas *Rage* is an intense form of *Anger*. Figure 9 distributes the emotions recorded into the categories of *Low*, *Neutral* and *High intensity*, and distributes them according to whether or not these emotions are positive or negative. The data indicates that the majority of emotions observed were of *Low Intensity*, as it accounts for 56% of the data. Surprisingly, only 11% of the data indicates a sigh of *High Intensity*, which suggests that a person rarely expresses intense emotions when sighing. The data also highlights that 82% of *High Intensity* emotions expressed were of a positive nature, contrasting to 32% of *Low Intensity* emotions

that were positive. Example 7 indicates a posi-

Bev: (SPK)
 5139. So, I don't know.
 5140. They said work from eleven point three.
 5141. I mean this is the last which is there.
 Wendy: (ADR)
 5142. Yeah.
 Bev:
 5143. From what I understand.
 5144. I dunno!
 5145. <sigh> Ah!
 5146.<pause dur=9> Only I've <pause> get a
 <unclear> if you want one.
 5147. <unclear>. I thought there's no point
 in leaving it in here.
 (BS) (Discussing) (Amazement)

Example 7

itive sigh that was expressed of a *High Intensity*. Thus, this sigh can be categorised as a sigh of high arousal and positive valence, as opposed to the majority of low arousal sighs of negative valence. Figure 9 indicates that the majority of sighs

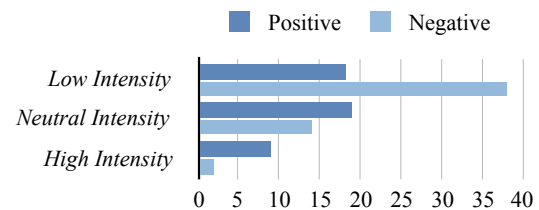


Figure 9: Distribution of intensity of emotions by positive or negative

expressed were of *Low Intensity* and negative, as these sighs account for 38% of all data. 58% of sighs expressed at *Neutral Intensity* were positive also, indicating that if a sigh is expressed above *Low Intensity*, it is more probable for it to be positive.

5 Discussion

The development of this annotation scheme in relation to three distinct dimensions informs our categorisation of sighs by their emotional valency. The results of the first dimension indicate that the vocal environment has a direct relationship with the valency of a sigh, and that the *Speaker* in a dialogue is more likely to express a sigh than an *Addressee*. Interestingly, when a *Speaker* expresses a sigh the probability for positive valence is 50%, contrasting to the probability for an *Addressee* to express a sigh of positive valence which lies at 37%. 75% of sighs expressed *Between Pause* were recorded as positive, providing an excellent prob-

ability score for future sigh interpretation. Thus, it is evident that by investigating the vocal environment of the sigh, we can predict whether or not a sigh will be positive or negative.

The results from the second dimension indicate that the most common reason for expressing a sigh is when *Answering* and there exists a 71% probability that this sigh is negative. However, for the second most popular reason, it is difficult to predict whether or not a sigh is positive or negative. The only reason which provides a great probability score is *Discussing*, as 75% of these sighs were recorded as positive.

The results of the third dimension indicate that 46% of sighs can be interpreted as expressing a positive emotion. This result is crucial in understanding the complex nature of the sigh and how subjective it's emotional interpretation is. However, it is clear that the most common emotion for expressing a sigh is *Anger* (which is negative) and the least common emotion is *Joy* (which is positive). The majority of the emotions recorded were of *Low Intensity*, accounting for 56% of sighs, and of these emotions at *Low Intensity*, there exists a 68% probability that the sigh is negative. However, of the recorded sighs at *Neutral* or *High Intensity*, the majority of sighs were positive. For emotions of high arousal, 82% of the sighs were of positive valency, making it easy to predict that an emotion of high arousal retains positive valency. By engaging with Straus' paper, it is clear that by looking at emotions of low arousal, a sigh will more likely be of negative valency. However, this paper highlights the importance of an emotional scale when interpreting the emotional quality of a sigh.

6 Formal Analysis

In this section, we sketch how lexical entries for sighs can be provided within a dialogue semantics. We follow the approach to non-verbal social signals sketched in (Ginzburg et al., 2018). Their approach involves two basic steps: (i) integrating Scherer's component process model (CPM) of appraisal (Scherer, 2009b) with the dialogical framework KoS (Ginzburg, 2012). (ii) reifying non-verbal social signal content by positing an external real world event as trigger.

Within the component process model an agent evaluates events she perceives and their consequences by means of a number of criteria

or stimulus evaluation checks (SECs) (e.g., *Is the event intrinsically pleasant or unpleasant, independently of my current motivational state? Who was responsible and what was the reason? Do I have sufficient power to exert control if possible?*). Each appraisal is, therefore modelled in Type Theory with Records in terms of a type given in (8). Pleasantness is specified via a scalar predicate *Pleasant* which can be positively aroused or negatively aroused or both; Power is specified in terms of a scalar predicate *Powerful* whose lower bound arises when the arousal value is zero.

$$(8) \text{Appraisal} = \left[\begin{array}{l} \text{pleasant} : \left[\begin{array}{l} \text{Pred} = \text{Pleasant} : \text{EmotivePred} \\ \text{arousal} : \left[\begin{array}{l} \text{pve} : \mathbb{N} \\ \text{nve} : \mathbb{N} \end{array} \right] \end{array} \right] \\ \text{responsible} : \text{RecType} \\ \text{power} : \left[\begin{array}{l} \text{Pred} = \text{Powerful} : \text{EmotivePred} \\ \text{arousal} : \mathbb{N} \end{array} \right] \end{array} \right]$$

Appraisal is incorporated in the dialogue game-board, the public part of information states in KoS, in terms of an additional repository MOOD—a weighted sum of appraisals. In this way MOOD represents the publicly accessible emotional aspect of an agent that arises by publicly visible actions (such as non-verbal social signals), which can but need not diverge from the private emotional state. The resulting type of DGBs is given in (9).

$$(9) \text{DGBType} \mapsto \left[\begin{array}{l} \text{spkr} : \text{Ind} \\ \text{addr} : \text{Ind} \\ \text{utt-time} : \text{Time} \\ \text{c-utt} : \text{addressing}(\text{spkr}, \text{addr}, \text{utt-time}) \\ \text{Facts} : \text{Set}(\text{Prop}) \\ \text{Pending} : \text{list}(\text{LocProp}) \\ \text{Moves} : \text{list}(\text{LocProp}) \\ \text{QUD} : \text{poset}(\text{Question}) \\ \text{Mood} : \text{Appraisal} \end{array} \right]$$

An update rule that increments by δ the positive pleasantness recorded in Mood given the weight ϵ (between new appraisal and existing Mood) is given in (10); the converse operation of incrementing the negative pleasantness is entirely analogous with the obvious permutation on the pve/nve values *mutatis mutandis*.

$$(10) \text{PositivePleasantnessIncr}(\delta, \epsilon) =_{def}$$

$$\left[\begin{array}{l} \text{preconditions} : \left[\text{LatestMove.cont} : \text{IllocProp} \right] \\ \text{effect} : \left[\begin{array}{l} \text{Mood.pleasant.arousal.pve} = \\ \epsilon(\text{preconds.Mood.pleasant.arousal.pve}) \\ + (1 - \epsilon)\delta : \text{Real} \\ \text{Mood.pleasant.arousal.nve} = \\ \epsilon(\text{preconds.Mood.pleasant.arousal.nve}) : \\ \text{Real} \end{array} \right] \end{array} \right]$$

Given our earlier discussion, we can posit two distinct lexical entries for sighs. We distinguish non-high arousal sighs from high arousal ones, associating the former with negative pleasantness and a sense of powerlessness, the latter with positive pleasantness. Respective lexical entries are (11a,c), where p is the *sighable*, the event triggering the sigh, identified with an Austinian proposition; (11b) is an update rule associated with (11a), incrementing the negative pleasantness and setting the power arousal level to zero. The force of a positive sigh (11c) is postulated to be simply Pleasant, which makes it trigger the positive pleasantness update, like laughter and smiling do.

$$\begin{aligned}
(11a) \quad & \left[\begin{array}{l} \text{phon : sighphontype} \\ \\ \text{dgb-params : } \left[\begin{array}{l} \text{spkr : Ind} \\ \text{addr : Ind} \\ \text{t : TIME} \\ \text{c1 : addressing(spkr,addr,t)} \\ \delta : \text{Int} \\ \text{c2 : Arousal}(\delta, \text{phon}) \\ \text{c3 : } \delta < \text{HighArousal} \\ \text{s : Rec} \\ \text{p = } \left[\begin{array}{l} \text{sit = 1} \\ \text{sit-type = L} \end{array} \right] : \text{prop} \end{array} \right] \\ \\ \text{content = } \left[\begin{array}{l} \text{sit = s} \\ \text{sit-type = } \\ \text{c4 : Unpleasant-accept}(p, \delta, \text{spkr}) \end{array} \right] \\ \text{Prop} \end{array} \right] : \\
(11b) \quad & \left[\begin{array}{l} \text{preconditions : } \left[\begin{array}{l} \text{LatestMove.cont =} \\ \text{Assert(spkr,} \\ \text{Unpleasant-accept}(p, \delta, \text{spkr})) : \\ \text{IllocProp} \end{array} \right] \\ \\ \text{effect : } \left[\begin{array}{l} \text{NegativePleasantnessIncr}(\delta, \epsilon) \\ \text{Mood.Power.arousal = 0} \end{array} \right] \end{array} \right] \\
(11c) \quad & \left[\begin{array}{l} \text{phon : sighphontype} \\ \\ \text{dgb-params : } \left[\begin{array}{l} \text{spkr : Ind} \\ \text{addr : Ind} \\ \text{t : TIME} \\ \text{c1 : addressing(spkr,addr,t)} \\ \delta : \text{Int} \\ \text{c2 : Arousal}(\delta, \text{phon}) \\ \text{c3 : } \delta \geq \text{HighArousal} \\ \text{s : Rec} \\ \text{p = } \left[\begin{array}{l} \text{sit = 1} \\ \text{sit-type = L} \end{array} \right] : \text{prop} \end{array} \right] \\ \\ \text{content = } \left[\begin{array}{l} \text{sit = s} \\ \text{sit-type = } \left[\text{c4 : Pleasant}(p, \delta, \text{spkr}) \right] \end{array} \right] : \text{Prop} \end{array} \right]
\end{aligned}$$

7 Conclusion

In this paper an annotation scheme was developed to investigate the quality of sighs in relation

to three dimensions; their placement in dialogue, their reasoning and emotion expressed by the sigh. There is clearly potential subjectivity when interpreting the data and recording the sighs, and the possibility that by using a different emotion scale, the results may differ. The inter-annotator study indicates a moderately high agreement but highlights also the discrepancies regarding emotion interpretation, indicating that broadening the categories of emotions would account for some difference in interpretation. We recognize two recurring kinds of sighs: those of low arousal and negative valence, and those of high arousal and positive valence. From our study it emerges that the probability for a sigh expressing a positive or negative emotion is almost equal, which contrasts with past research, which used fewer examples and no systematic emotion analysis. With this annotation scheme proposed, this paper hopes to have laid a firm basis for the future study and annotation of sighs. The complexity of sigh denotation could be reconciled through focus on contextual criteria of sighs and the establishment of a multitude of emotions with varying arousal. We concluded with a sketch of a formal analysis of sighs within the framework of KoS integrated with Scherer's component process model of appraisal.

Acknowledgments

We acknowledge the support of the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) and a senior fellowship from the Institut Universitaire de France to the second author. We thank Tristan Stérin for providing another critical perspective on this subject which pushed us to delve further into our analysis and for continually providing kind help, support and feedback. Thank also to three anonymous reviewers for SemDial for their detailed and perceptive comments.

References

- British National Corpus. British National Corpus. <http://www.natcorp.ox.ac.uk>. Accessed: 2019-05-15.
- British National Corpus 2014: User Manual and Reference Guide. British National Corpus 2014: User Manual and Reference Guide. <http://corpora.lancs.ac.uk/bnc2014/>.
- Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.
- Jonathan Ginzburg, Chiara Mazzocchi, and Ye Tian. 2018. Interaction, appraisal, and non-verbal social signal meaning. In *Proceedings of Cognitive Structures: Linguistic, Philosophical and Psychological Perspectives (CoSt18)*, Heinrich Heine University, Düsseldorf.
- Erving Goffman. 1978. *Response cries*. *Language*, 54(4):787–815.
- Orlando Gomes. 2017. *Plutchik and economics: Disgust, fear, and, oh yes, love*. *Economic Issues Journal Articles*, 22(1):37–63.
- Elliott M Hoey. 2014. Sighing in interaction: Somatic, semiotic, and social. *Research on Language and Social Interaction*, 47(2):175–200.
- Keith Oatley and P. N. Johnson-Laird. 2014. *Cognitive approaches to emotions*. *Trends in Cognitive Sciences*, 18(3):134–140.
- Robert Plutchik. 1988. *The Nature of Emotions: Clinical Implications*, pages 1–20. Springer US, Boston, MA.
- Matthew Purver. 2001. Score: A tool for searching the bnc. Technical Report TR-01-07, King’s College, London.
- Klaus Scherer. 2009a. *The dynamic architecture of emotion: Evidence for the component process model*. *Cognition and Emotion*, 23:1307–.
- Klaus R Scherer. 2009b. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351.
- Erwin W. Straus. 1952. The sigh: An introduction to a theory of expression. *Tijdschrift Voor Filosofie*, 14(4):674–695.
- KARL HALVOR Teigen. 2008. *Is a sigh “just a sigh”? sighs as emotional signals and responses to a difficult task*. *Scandinavian Journal of Psychology*, 49(1):49–57.
- Khiet P. Truong, Gerben J. Westerhof, Franciska de Jong, and Dirk Heylen. 2014. *An annotation scheme for sighs in spontaneous dialogue*. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, Singapore, September 14-18, 2014, pages 228–232. ISCA.
- Elke Vlemincx, Ilse Van Diest, and Omer Van den Bergh. 2015. *Emotion, sighing, and respiratory variability*. *Psychophysiology*, 52(5):657–666.

Posture Shifts in Conversation: An Exploratory Study with Textile Sensors

Sophie Skach

Human Interaction Lab
Cognitive Science Research Group
Queen Mary University of London
s.skach@qmul.ac.uk

Patrick G.T. Healey

Human Interaction Lab
Cognitive Science Research Group
Queen Mary University of London
p.healey@qmul.ac.uk

Abstract

Posture shifts involving movement of half or more of the body are one of the most conspicuous non-verbal events in conversation. Despite this we know less about what these movements signal about the interaction than we do about smaller scale movements such as nods and gestures. This paper reports an exploratory study of posture shifts in seated conversation. Using data from video analysis and bespoke pressure sensors in clothing, we are able to distinguish different types of posture shifts and detect them in speakers and listeners. The results show that large scale posture shifts are performed by both speaker and listener, appear to be distinct from smaller movements associated with preparing to speak and that the posture shifts associated with speech offset are less clearly defined. We discuss the potential of using pressure sensors to investigate these salient conversational states.

1 Introduction

One of the most salient body movements people make in natural conversation is a general posture shift in which most or all of the body goes through a momentary adjustment. While these movements could, of course, be explained by fatigue or physical discomfort there is also an intuition that they have communicative significance. Unlike, say, iconic gestures or nods that accompany each utterance these are relatively global, infrequent movements that seem to mark larger conversational units or signal something about participant's stance towards an issue. Schefflen (1964) was one of the first to document these moments in detailed case studies of psychotherapy sessions. He defined posture shifts as movements involving at least half the body and proposed that they are organised around changes in *position* or point of view.

Others have since elaborated on Schefflen's findings, describing posture shifts as self synchronised movements to speaker turns (Condon and Ogston, 1966), as signals for different levels of engagement in a conversation (Schegloff, 1998) or to correlate with tonic stress (Bull and Connelly, 1985). In most cases, postural changes are linked to speaker behaviours. They can accentuate it in fine grained ways (Ekman and Friesen, 1969), and also accompany the change of speech categories (Bull and Brown, 1977). Posture shifts can also appear outside of speech and may be interesting signals in interaction in their own right. For example Bull has considered frequent posture changes as a marker of boredom (Bull, 2016).

Although there is an intuition that posture shifts are important non-verbal signals, not least because of their relative scale, the literature on them is limited. More attention has been given to posture as a static feature of participation in conversation, especially in relation to posture matching as indication of affiliation or attraction (Beattie and Beattie, 1981; Bianchi-Berthouze et al., 2006; Mehrabian, 1969), and in their spatial formation (Kendon, 1976).

The work on posture reviewed above relies on video combined with human coded judgements of posture type for analysis. More recently there has been an increase of interest in the use of motion capture and markerless computer vision techniques as ways of automatically measuring posture (Wei et al., 2016). Here we extend this to considering the use of pressure sensors as a way of sensing changes in seated postures. This has the advantage that it is not susceptible to problems with occlusion that can affect camera-based techniques (e.g. see Tan et al. (2001), Meyer et al. (2010) and Skach et al. (2017)). It can also detect subtle changes in pressure that do not necessarily translate to overt visual cues. Furthermore, we in-

troduce pressure sensors made of conductive textiles integrated into fabric surfaces as a method to capture shifts in movement and behavioural cues. We use bespoke 'smart' trousers with an integrated sensor matrix to record pressure data around the thighs and the buttocks. This is used in an exploratory study of changes of posture both for listeners and speakers in multiparty seated conversations. We explore the potential of pressure sensors in trousers to detect changes of state in the conversation and discuss the qualitative characteristics of some of the events they detect.

2 Background

There are several suggestions as to what postural shifts mean and what role they play in punctuating communication between interactants, between speakers and addressees, and also when in conversation they are most likely to appear. Generally, posture shifts have been associated with changes in topic ("locution cluster", coined by Kendon (1970)) or situations (Gumperz, 1982). Condon (1976) and Lafrance (1976) also reported on postural synchrony, leading to higher rapport, or if incongruent, are indicators for negative relations between people (Lafrance and Broadbent, 1976). Furthermore, the exposure and intensity of such movement may present cues to interpersonal relationships. For example, Wiemann et al. (1975) suggested that the more familiar interactants are with each other, the more subtle the postural shifts and bodily movement, moving parts of limbs (fingers) rather than entire body parts. This can be linked to Kendon's observation (1972) that generally, those body parts are in more motion than the torso and the legs.

2.1 Speakers and Listeners

Postural changes have been reported most commonly in connection to speaker behaviours, or listeners' perception of speakers. Hadar (1984) reports that they appear primarily at the start of a speaking turn, when the interactant changes their state from listener to speaker, or after a long speaker pause.

Speakers are said to punctuate the end of their turn and maintain a more upright posture overall, leaning rather forward than backwards (Wiemann and Knapp, 1975), or emphasise words and phrases. Even micro-movements like facial expressions, e.g. raising an eyebrow, can be in line

with changes in tonality, e.g. lowering voice (Condon and Ogston, 1966). Bull and Brown (1985) identified 3 postures related to the upper body and 5 related to the lower body, evaluating them in relation to 6 different categories of speech.

Listeners' postures are examined less often. It is suggested that the status of an addressee can be interpreted by the openness of their legs and arms (Mehrabian, 1968), and that listeners synchronise with speakers (Condon and Ogston, 1966) and shared postures between them are linked to a high rapport (Lafrance and Broadbent, 1976). Also pauses between speech as listener turns are associated with postural adjustments by Hadar et al. (1984).

2.2 Sensing Social Cues

Sensing bodily movement as behavioural or affective signals has been subject to numerous human-centred research, both for interaction with each other or with a device (HCI). While conventionally, video and audio recordings were used, other modalities have been explored in more recent years. One of the goals for new technologies is to maintain an undisturbed environment, deploying sensors unintrusively. Ambient interior design and the utilisation of everyday objects has been successful a contribution to such ubiquitous computing (see e.g. Vinciarelli et al. (2009) or Venkatarayan and Shahzad (2018)).

A material that is closest to our skin, follows our movements organically and is a natural interface we have used for thousands of years is fabric. Therefore, using our own clothing as a sensing surface seems appropriate to capture bodily actions and behaviours (such as in Mattmann et al. (2007)).

Here, we exploit trousers and test their performance to capture postural shifts as dynamic movements as opposed to static postures, which we have proven to reliably identify with sensing trousers before (Skach et al., 2018).

3 Methodology

In this section, we report on the design of the 'smart' trousers and the development of custom-made textile sensors, as well as on the process of collecting video data in a user study.

The data is drawn from a corpus of seated, three-way unscripted conversations (Figure 1). The conversations were video recorded to allow

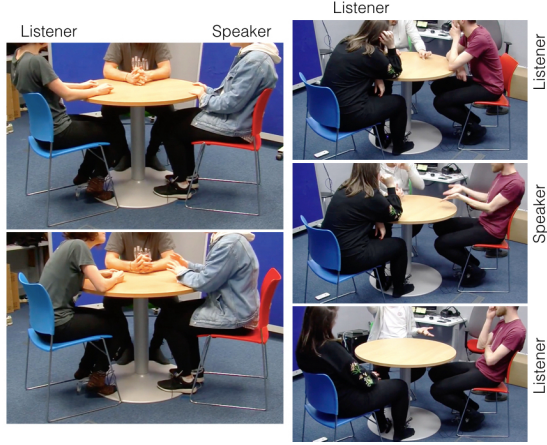


Figure 1: Examples of postural shifts, to be read from top to bottom: Left: Listener leans towards speaker, responds to their posture change (leaning forward). Right: Listener posture shifts on the left; postural transitions from listener to speaker and back, on the right.

for hand coding of non-verbal behaviours using two cameras from different angles to reduce problems with occlusion. In addition, participants wore specially constructed trousers with an array of fabric pressure sensors built in (see below) to detect lower body movements. These sensing trousers continuously recorded changes of pressure across thighs and buttocks.

This combination makes it possible to identify conversational phenomena such as speaking or listening and identify whether they are systematically associated with lower body movements.

3.1 Textile Sensors in Trousers

A fabric sensor matrix of 10x10 data points was designed (adapted from [Donneaud and Strohmeier \(2017\)](#)) and embedded in custom made stretch trousers, as seen in Figure 1. Each leg’s matrix therefore consists of 100 sensors and is deployed around the upper leg, covering the area from the knee upwards to the buttocks in the back and the crotch in the front, as illustrated in Figure 2. Placement, shape, amount and type of the sensors, as well as the type of trousers that were chosen for the sensor integration derived from ethnographic observations of multi-party social interactions. The use of soft, textile conductive materials, of which the pressure sensors consist, enables unintrusive sensing without augmenting conventional trousers’ properties. A detailed documentation of the design and manufacturing process of this wearable sensing system is reported in ([Skach](#)

Tier	Description
Talk	on- and offset of overt speech
Pre-Speech	2 sec immediately before talk
Post-Speech	2 sec immediately after talk
Posture Shift	gross movement of torso & legs

Table 1: Overview of the hand coded annotations in Elan

et al., 2018).

3.2 Data Collection

3.2.1 Participants

A total of 42 participants were grouped into 14 three-way conversations, each of them was given a pair of sensing trousers¹. A subset of 5 participants were annotated and analysed here: 4 female, 1 male. These participants were selected randomly from a predetermined subset of participants that performed above average in preliminary posture classification tasks.

3.2.2 Procedure

Participants were seated on a round table and given the task to resolve a moral dilemma between them. Conversations lasted 15 to 20 minutes and were captured from two angles with video cameras in addition to recording the pressure data from the sensing trousers that each participant was wearing during the entire time of the recording.

3.2.3 Sensor Data Processing

The pressure readings from the fabric sensors were recorded with a time stamp and were stored on a microSD card integrated and hidden in the hem of the trousers. The data was captured at 4Hz by a microcontroller placed in the hem, too. For further processing, the data of each of the 200 sensors was normalised.

3.3 Annotations

The recorded videos were hand coded using the software package Elan ([Brugman and Russel, 2004](#)), with two annotators for determining speech, and one annotator to code posture shifts.

First pass coding focused on overt speech with starts and ends of annotations defined by onset and offset of audible speech. Second pass annotation then coded the moments immediately before and after speaking arbitrarily defined as 2 sec-

¹We manufactured multiple trousers in different sizes (Small, Medium, Large) to accommodate all participants.

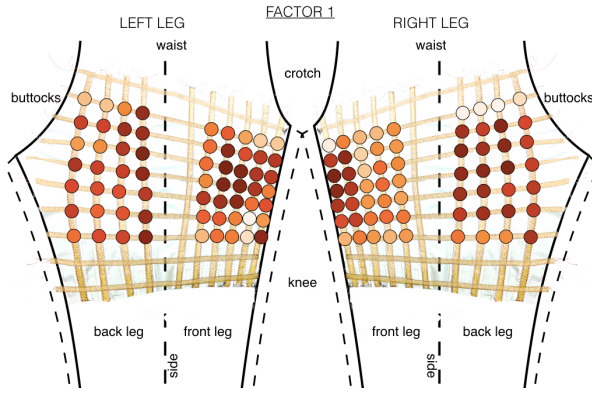


Figure 2: Visualisation of component 1 for each sensor, mainly discriminating posture shifts, talk, and pre-speech movement. Dark colours show positive associations, bright tones negatives (similar to a heat map).

onds just before, and 2 seconds just after speech. These were coded regardless of other non-verbal behaviours or marked bodily movement. Third pass coding was used to identify posture shifts defined as gross body movement involving either or both, the upper and the lower body. This includes leaning forwards, backwards, and sideways, but also performing leg crossing and adjusting sitting position with thighs and hips (shifting the weight within a seated counterpose). Both, speaker and listener posture shifts were included. Again, some movement coincided with other behavioural cues, verbal and non-verbal. An overview of the coding scheme can be seen in Table 1.

Later, the annotations were synchronised with the sensor data of both legs by merging and approximating the time lines of both recordings with each other. Broken sensors were removed from further processing and analysis.

4 Results

The results are reported in two steps: a) analysis of the pressure sensor data and b) observations of the interactional context of the posture shifts.

Across all participants in our video, posture shifts occurred on a regular basis. In a time window of 15 minutes, an average of 35 posture shifts were annotated, which equates to 2-3 posture shifts each minute. By posture shift, we define the positional movement of the torso and / or the lower body including the legs. In the scope of this work, we exclude gaze and gestures from postural shifts, but acknowledge that gestures in particular are often described as part of a postural shift that affects the dynamics of the entire torso

Comp.	Total	% of Variance	Cumulative in %
1	38.182	30.303	30.303
2	31.003	24.606	54.909
3	25.184	19.988	74.896
4	9.523	7.558	82.454
5	6.491	5.152	87.606
6	3.624	2.876	90.482
7	1.994	1.583	92.065
8	1.575	1.250	93.315
9	1.218	0.966	94.218

Table 2: Variance Explained (Extraction Sums of Squares Loadings)

(Cassell et al., 2001).

4.1 Posture Shifts and Pressure Changes

4.1.1 Factor Analysis

The 200 pressure sensors on each participant (100 right leg, 100 left leg) produce a relatively complex array of pressure measurements with a significant amount of redundancy between sensors. Hardware failures reduced this to 165. If a sensor failed on one participant the data were deleted for all participants to ensure equivalent sensor arrays were used for each person. The sensors yielded a total of 6278 pressure measurements across the whole sample (in total for both legs, per participant). In order to reduce the complexity of the sensor data a factor analysis was calculated using SPSS (v.25). This yielded 9 components that account for 94% of the variance.

The influence of the four coded behaviours listed in Table 1 on pressure changes was analysed using Automatic Linear Modelling with forward stepwise model selection. Talk (1/0), Beforetalk (1/0), Aftertalk (1/0), and Participant (1-5) were used as predictors and the regression factor score for each component from the factor analysis for each pressure observation as the target.

For Component 1 the model fit is 88%, Information Criterion -10,438. The analysis shows that Participant ($p < 0.000$), Postureshift (Coefficient = -0.133 $p = 0.003$), Talk (Coefficient = -0.047, $p < 0.000$) and Beforetalk (Coefficient = -0.041 $p < 0.004$) predict changes in first factor (component) of the pressure data. The effect of the individual sensors for component 1 are visualised in Figure 2, showing which sensors have positive and negative associations. From this, we see that the

front mid thigh on the left leg, and the mid buttocks of the right leg affect the predictions most positively, while the sensors in crotch proximity, on the upper buttocks, as well as on lower mid thighs have negative associations. Interestingly, these patterns are not symmetrical.

The estimated means of these effects for Factor 1 are illustrated in Figure 3. Components 2-8 are primarily predicted by Participant with different Components picking out different subgroups of participants. There are two exceptions: Component 3 is also marginally predicted by Aftertalk (Coefficient -0.031, $p < 0.000$) and Component 6 is also predicted by Postureshift. Component 9 which has a relatively poor model fit (4.5% accuracy, and Information Criterion -216.0) is predicted by Postureshift (Coefficient = -0.204, $p < 0.000$), Aftertalk (Coefficient = 0.125, $p = 0.001$) and Beforetalk (Coefficient = 0.101 $p < 0.005$).

The pressure data changes corresponding to the predictors found for Component 1 are illustrated in Figures 4, 5 and 6. Note that, in effect 'Beforetalk' is the inverse of Talk but sampled over a smaller data set. Together they show that talking is associated with an overall increase in lower body pressure (when seated) and that the shift takes place in a two second window prior to speaking. Conversely, large scale posture shifts are associated with an overall decrease in lower body pressure.

Overall, these preliminary results suggest that the array of pressure sensors can be used to discriminate between global posture shifts and also the movements people make immediately before and after speaking. This replicates an earlier analysis of the pressure data comparing talking vs. listening using machine learning techniques. The results also highlight the substantial individual variation in the pattern of the pressure data. Individual identities form the largest and most consistent predictor of pressure patterns across all the analyses.

4.2 Observational Findings

The posture shifts coded from the videos were explored to develop hypotheses about the possible functions of the large scale posture shifts in this corpus. We divide types of posture shifts according to the time of their appearance in relation to overt speech: before, during, after and between speakers' turns.

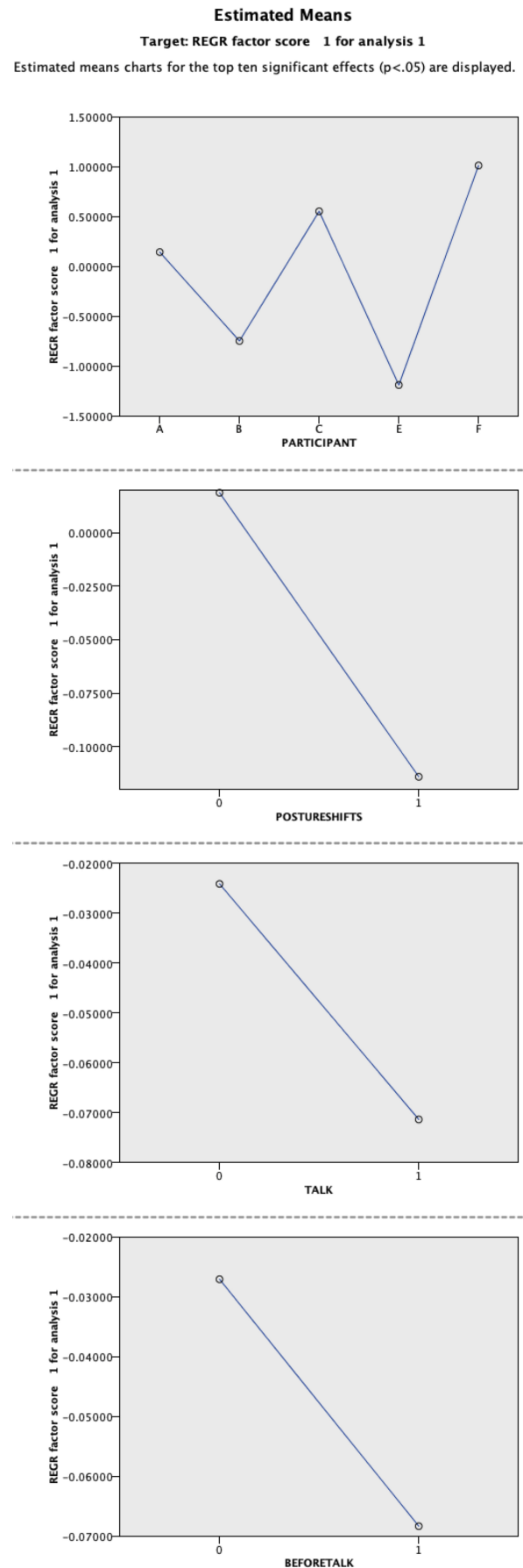


Figure 3: Estimated Means of the first factor for the top ten significant effects ($p < 0.05$) are displayed

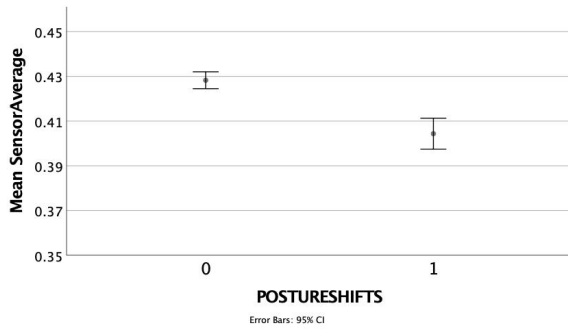


Figure 4: Pressure Change with Posture Shifts: Average Normalised Sensor Data

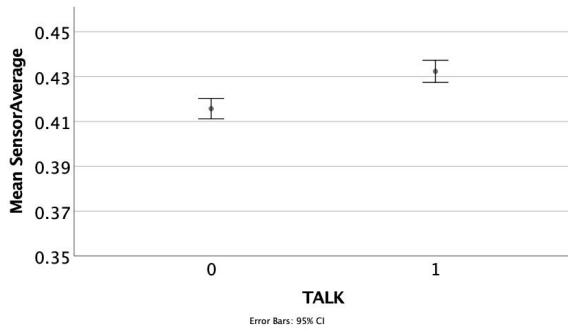


Figure 5: Pressure Change when Talking: Average Normalised Sensor Data

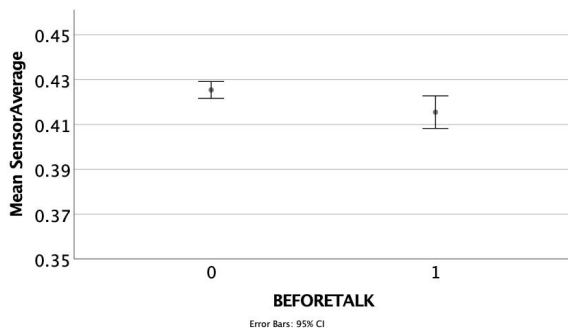


Figure 6: Pressure Change Before talking: Average Normalised Sensor Data

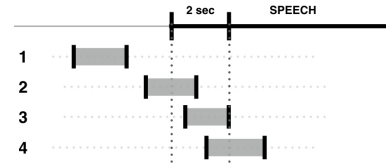


Figure 7: Preparatory Movement Types: 1) performed several seconds before utterance; 2) completion within 2 sec before talk; 3) start & end within 2 sec window; completion often precisely at start of talk; 4) start within 2 sec window, overlap with talk.

4.2.1 Preparatory Movement

Listed below are the four categories of posture shifts before a speaker's turn, also illustrated in Figure 7:

1. Start and end of movement several seconds before utterance (end of movement ≥ 2 sec before talk), however still close enough to be seen as preparatory.
2. Start of movement before speech, outside of 2 sec window, but completion within this time window, up to the very start (onset) of speech.
3. Occurrence of posture shift precisely within 2 seconds before speech, ending at the very start of utterance.
4. Posture shift starts within 2 sec just before, and is executed and completed during speech

The evaluation of the sample set of 5 participants shows that, considering the frequency of these categories, 80% of preparatory postural movements can be captured in part or as a whole through the time window of the 2 seconds annotations. The rest of preparatory posture shifts happens largely between 4 and 3 seconds before speech. One approach therefore, with the aim of capturing these movements, is to extend the specified time window to 4 seconds before talk. This, however, would often mean that postural preparation is longer than talk itself, whose duration is 3.21 seconds on average across all participants. These findings confirm our initial hypothesis on posture as preparation for speech, and also align with previous suggestions that posture change indicates turn taking and interactants signal their next speaking turn through these movements.

4.2.2 Delayed Post-Speech Shifts

We observed that postural shifts that are not classified as preparatory movement, but rather as

post-speech movement, follow a different pattern. Overall, they occur less frequently and are only rarely performed in the immediate aftermath of talking utterances (inside the 2 seconds time frame). This is not to say they don't exist, but more commonly, they seem to be performed with a short delay. We can categorise this delay in similar ways as the preparatory movement (mirroring Figure 7):

1. overlap with speech: posture adjustment performed towards the end of speech and beyond: start of movement within speech, completion after speech has ended.
2. no delay: start of postural movement immediately after offset of speech
3. short delay: after utterance ends, postural shift is performed with a delay of ≤ 2 sec (within the specified time window)
4. long delay: considered as a movement being performed more than 2 sec after speech has ended (outside specified time window)

In numbers, we have found that only 2 out of 47 post-speech movements are performed immediately at the offset of speech. Most postural shifts that are associated with the end of an utterance are performed with a delay between 1 and 4 seconds after talking (with rare outliers up to 5 seconds after, everything later than this was not linked as a post-speech postural adjustment) - with 49% of them falling into the specified time window of 2 seconds post talk. In fact, most movements of this category started within this time window, but at the same time, 28% of posture shifts started only clearly after 2 seconds post speech. In other words, this means categories 3) and 4) are the most common amongst post-talk postural movement - with a short or long delay.

4.2.3 Active Listener Postures

Although postural adjustments have been closely linked to speaking, they are interesting phenomena in their own rights. In our data set, speakers' shifts only account for 40.44% of posture shifts. Listeners' posture shifts, however, often co-occur with other conversational behaviours, appearing to signal something about participants' relation to what is happening in an interaction. We observed that in most cases where not linked to speaker behaviour, they are often related to specific 'active'

listener signals, such as nodding, backchanneling or laughing, which go somewhat beyond these specific forms of concurrent feedback. Two examples are depicted in Figure 1. In some cases, shifts in postures seem to predict these behaviours, too, similar to the patterns of preparatory movement for talk. In general, the movement patterns for backchannels were most similar to the ones for talk. During nodding, the movement of both torso and legs appeared visibly more subtle and was observed to only become more embodied when close (within 5 seconds) to a speaking turn. This could be discussed as another extended preparation for speech, too. However, posture shifts related to nodding only make up 6.56%, the smallest category. When looking at laughter, postural movement was expectedly the most marked and obvious, and also forms the second largest set of all posture shifts, 28.96%. In comparison, 8.20% of all postural movements relate to backchannels. Additionally, our observations suggest that not only during these active listener behaviours, but also for the embodied transition from inattentive to attentive listeners, postural shifts play an important role, accounting for 17.76% of all movements, and expanding on the reports of Kendon (1972), Schefflen (1964) and Blom and Gumperz (1982).

5 Discussion

The results of this exploratory study suggest that posture shifts are a significant and rich interactional phenomenon that deserve more attention. Nonetheless, it is important to acknowledge that the data set presented here is small and the observations made here can only be considered preliminary.

5.1 Topic Changes in Speech

Kendon (1972) has discussed posture shifts in relation to changes in topics, and Bull and Brown (1985) have also noted different postural patterns in specific categories of speech (e.g. drawing back legs or raising a foot during a statement). In this work, we have not considered differences in what is being said, but have treated talk as a broad, overt event. Posture shifts performed during speech were coded and included in the analysis, but were not further divided into more fine grained categories of nuanced speech. Therefore, we did not examine whether postural movement during a speaker turn correlates with topic changes. From

observation, however, it is suggested that in some occasions, there is evidence to confirm the works of Kendon, Cassell (2001), Schulman (2011) and others. For example, the participants of our sample set that have embodied such topic changes in a marked way, have moved both their torso and lower body significantly. Following this, it would be interesting to explore whether different markedness of posture shifts correlate with different conversational events not only in individual cases, but in a general conversational structure.

5.2 Individual Variation

The most obvious point about the data presented here is the large amounts of individual variation. Individual participants showed patterns of movement that seemed specific to them, and may be a starting point towards an approach to identify individuals through postural movement. Nonetheless, the analysis suggests that there are still commonalities in the patterns of posture change that may generalise across individuals.

In consideration of individual variation, there were some nuances in postural movements we observed that were distinct for different participants. Rhythmic, continuous events were leg bouncing and back- and forwards swinging with the torso. These events occurred alongside other, previously mentioned behaviours that present more specified social signals and are to find for each participant: nodding and laughter. In some cases, they also appeared to correlate with affective states. One participant, for example, bounced their leg in supposedly uncomfortable moments. Another participant, when listening and not giving any other cues to speakers, continuously moved his torso back and forth, lightly swinging. Others have performed smaller movements like fidgeting more frequent than gross postural shifts.

5.3 Familiarity and Synchrony

The idea that interactants move in different ways depending on how familiar they are with each other comes from Wiemann and Knapp (1975), and suggests more subtle movement when participants know each other. This aligns with the works of Kendon (1976), discussing spatial organisation as a signifier for interpersonal relationships. We have noted this phenomenon in individual cases and have not gathered enough evidence to support Wiemann and Knapp's suggestion in full, but have observed that the number of gross body move-

ments decreased after the first 5 minutes into the conversation. After that, movements became more subtle. In this context, it is also to note that the participants we have grouped together, were in different personal relationships: some knew each other briefly, while others were not familiar with each other at all.

Furthermore, it is also not clear and has not been investigated in this study, whether posture shifts are always noticed by conversation partners. This especially refers to smaller scale movements, whose interactional relevance could followingly be discussed, too.

5.4 Handedness and 'Footedness'?

One additional suggestion emerging from this study is that the pressure sensors of the left leg appear to be more discriminative of posture shifts than the right leg. This might have two reasons: the variation of the sensor performance, considering self made sensors as difficult to calibrate; or a potential correlation with handedness. There are some indications that people gesture differently with their dominant hand we speculate that this might also influence the pressure distribution of legs, too. To elaborate on these ideas, more information about the participants is required, that was not asked for in our studies.

6 Conclusion

This exploratory study contributes to the discourse on the meaning of posture shifts and their role in conversation. We have showed that it is possible to identify different types of postural movements through a novel multimodal approach: video recordings and a wearable sensing system made of fabric pressure sensors in trousers. These were used for a study in which we recorded the data of three-way conversations. The results show that there is a lot to draw from posture shifts in general, in relation to speech, as well as to active listener behaviours, verbal and non-verbal, and that smart clothing can be used to detect them.

7 Acknowledgements

This research is funded by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). We thank all participants and colleagues for their support - in particular Tom Gurion for his help with data formatting.

References

- Geoffrey W Beattie and Carol A Beattie. 1981. Postural congruence in a naturalistic setting. *Semiotica*, 35(1-2):41–56.
- N Bianchi-Berthouze, Paul Cairns, Anna Cox, Charlene Jennett, and Ww Kim. 2006. [On Posture as a Modality for Expressing and Recognizing Emotions](#). *Emotion in HCI workshop at BCS HCI*, pages 74–80.
- Hennie Brugman and Albert Russel. 2004. [Annotating multi-media/multi-modal resources with ELAN](#). *International Conference on Language Resources and Evaluation*, pages 2065–2068.
- P. E. Bull and R. Brown. 1977. [The role of postural change in dyadic conversations](#). *British Journal of Social and Clinical Psychology*, 16(1):29–33.
- Peter Bull and Gerry Connelly. 1985. [Body movement and emphasis in speech](#). *Journal of Nonverbal Behavior*, 9(3):169–187.
- Peter E Bull. 2016. *Posture & gesture*, volume 16. Elsevier.
- Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. 2001. Non-Verbal Cues for Discourse Structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 114–123. Association for Computational Linguistics.
- William S. Condon. 1976. [An Analysis of Behavioral Organization](#). *Sign Language Studies*, 13:285–318.
- William S Condon and William D Ogston. 1966. Sound film analysis of normal and pathological behavior patterns. *Journal of nervous and mental disease*.
- Maurin Donneaud and Paul Strohmeier. 2017. Designing a Multi-Touch eTextile for Music Performances. In *Proceedings of the 17th International Conference on New Interfaces for Musical Expression (NIME17)*, pages 15–19, Aalborg, Denmark.
- Paul Ekman and Wallace V Friesen. 1969. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding, Semiotical.
- John J. Gumperz. 1982. [Discourse Strategies](#). Cambridge University Press.
- U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose. 1984. [The timing of shifts of head postures during conversation](#). *Human Movement Science*, 3(3):237–245.
- Adam Kendon. 1970. Movement coordination in social interaction: Some examples described. *Acta psychologica*, 32:101–125.
- Adam Kendon. 1972. Some relationships between body motion and speech. *Studies in dyadic communication*, 7(177):90.
- Adam Kendon. 1976. [Spatial organization in social encounters: The F-formation system](#). *Man Environment Systems*, 6:291–296.
- Marianne Lafrance and Maida Broadbent. 1976. Group Rapport : Posture Sharing as a Nonverbal Indicator. *Group & Organizations Studies*, 1(3):328–333.
- Corinne Mattmann, Oliver Amft, Holger Harms, Gerhard Tröster, and Frank Clemens. 2007. [Recognizing upper body postures using textile strain sensors](#). *Proceedings - International Symposium on Wearable Computers, ISWC*, (2007):29–36.
- Albert Mehrabian. 1968. [Relationship of attitude to seated posture, orientation, and distance](#). *Journal of Personality and Social Psychology*, 10(1):26–30.
- Albert Mehrabian. 1969. [Significance of posture and position in the communication of attitude and status relationships](#). *Psychological Bulletin*, 71(5):359–372.
- Jan Meyer, Bert Arnrich, Johannes Schumm, and Gerhard Troster. 2010. [Design and modeling of a textile pressure sensor for sitting posture classification](#). *IEEE Sensors Journal*, 10(8):1391–1398.
- Albert E Schefflen. 1964. [The Significance of Posture in Communication Systems](#). *Psychiatry*, 27(4):316–331.
- Emanuel Schegloff. 1998. [Body torque](#). *Social Research*, 65(3):535–596.
- Daniel Schulman and Timothy Bickmore. 2011. [Posture, relationship, and discourse structure: Models of nonverbal behavior for long-term interaction](#). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6895 LNAI:106–112.
- Sophie Skach, Patrick G T Healey, and Rebecca Stewart. 2017. Talking Through Your Arse: Sensing Conversation with Seat Covers. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London.
- Sophie Skach, Rebecca Stewart, and Patrick G T Healey. 2018. [Smart Arse: Posture Classification with Textile Sensors in Trousers](#). *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 116–124.
- Hong Z. Tan, Lynne A. Slivovsky, and Alex Pentland. 2001. [A sensing chair using pressure distribution sensors](#). *IEEE/ASME Transactions on Mechatronics*, 6(3):261–268.
- Raghav H. Venkatnarayan and Muhammad Shahzad. 2018. [Gesture Recognition Using Ambient Light](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–28.

Alessandro Vinciarelli, Maja Pantic, and Herv Bourlard. 2009. [Social signal processing: Survey of an emerging domain](#). *Image and Vision Computing*, 27(12):1743–1759.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

John M. Wiemann and Mark L. Knapp. 1975. [Turntaking in Conversations](#). *Journal of Communication*, 25(2):75–92.

When objecting to presupposed content comes easily

Alexandra Lorson, Chris Cummins and Hannah Rohde

The University of Edinburgh, EH8 9AD, UK
{a.lorson, ccummins, hannah.rohde}@ed.ac.uk

Abstract

New content can be introduced into dialogue via presupposition as well as by assertion, but on traditional accounts presupposed information is expected to be less addressable in the subsequent dialogue. An alternative approach is to argue that addressability is more closely connected to whether content is at-issue with respect to the current Question Under Discussion. This paper investigates which of these factors is dominant. We report the results of a dialogue-based experiment designed to test whether and how false at-issue content is responded to in an ongoing discourse, and whether this is affected by its status as asserted or presupposed. Our findings suggest that when material is at-issue it can be challenged directly, independently of whether it is presupposed or asserted. However, relevant information introduced by a presupposition was found to be more likely to escape the participants' attention.

1 Introduction

Speakers in dialogue can introduce new information in the form of presupposition: that is, by presenting it as though it were already part of the common ground. In the case of (1), the expression *my car* carries an existential presupposition to the effect that the speaker has a car, but (1) can nevertheless be uttered in a context in which the hearer does not already know this.

Sorry I'm late: my car broke down. (1)

Similarly, in cases such as (2) and (3), the presence of the expressions *quit* and *be happy that* gives rise to presuppositions that might not already be known to the hearer, namely that John used to smoke and that Mary's boss is away.

John is anxious. He quit smoking. (2)

Mary is happy that her boss is away. (3)

This paper focuses on the status of new content that has been introduced via a presupposition. The study we present uses an interactive dialogue paradigm to probe whether and how such content is addressed as a discourse proceeds. The goal is to better understand how a speaker's choice of information packaging strategy within an individual utterance, considered alongside the active Question Under Discussion across the broader discourse context, influences an interlocutor's subsequent discourse continuations and their ability to take up particular content.

2 Background

We can identify the meanings mentioned above (that John quit smoking, and that Mary's boss is away) formally as presuppositions of (2) and (3) on the basis of their ability to project from under the scope of operators such as negation: (4) conveys the same presupposition as (2), and (5) as (3).

John is anxious. He didn't quit smoking. (4)

Mary isn't happy that her boss is away. (5)

Following Lewis (1979) and Von Stechow (2008), the utterance of (1)-(5) to a hearer who lacks the shared knowledge is argued to involve the exploitation of **accommodation**: the speaker acts as though a presupposition is already part of the common ground, and the hearer responds by adjusting their world-view, or situation model, to incorporate that presupposition. However, this relies on the assumption that the presupposition is one that the hearer is willing to entertain (or at least to ignore; see Glanzberg 2005) rather than one that the hearer refuses to accept or wishes to challenge. In normal cooperative conversation this assumption seems generally to be satisfied, but it's

easy to find cases in which it is violated by a speaker deliberately introducing potentially controversial material in the form of a presupposition, as in (6).¹

Everybody knows that Brett Kavanaugh's confirmation was a farce. (6)

Why, then, would a speaker choose to package information in the form of a presupposition rather than as a regular assertion? In the cooperative cases where the information is relevant but uncontroversial, we could see this as arising partly from efficiency considerations – an utterance such as (1), (2) or (3) is more concise than the corresponding version in which the presupposed content is directly asserted (“I have a car and it broke down”, etc.). But independent of efficiency, speakers might also select particular ways of packaging information because they anticipate how the discourse will proceed and what content will (or should) be taken up in subsequent utterances. Presupposed information, unlike asserted information, is typically regarded as difficult to address in the ongoing discourse. This is again connected to the projection behaviour of presuppositions. If a speaker utters (7) in response to (3), they are most naturally taken to be denying the assertion of (3) rather than its presupposition. As shown by (5), if we simply negate (3) we allow the presupposition to stand, because it projects from under the scope of negation. Hence, the speaker who responds to (3) with (7) is most naturally understood to mean (5).

That's not true! (7)

Von Stechow (2008), following Shanon (1976), argues that this offers a convenient diagnostic for presupposition: if we wish to deny a presupposition, we have to use a circumlocution such as “Hey, wait a minute...”, as shown in (8) (again considered as a response to (3)). This is dispreferred as a means of addressing asserted content, as shown by (9).

Hey, wait a minute, her boss isn't away. (8)

?Hey, wait a minute, she's not happy. (9)

Given the relative lack of addressability of presupposed content, we might expect cooperative speakers only to presuppose information that they do not expect to be taken up in the following

discourse. Otherwise, they would risk giving rise to the sense described by Schwarz (2019: 85) that “crucial and important information has been introduced in an inappropriate, underhanded way”. Correspondingly, we might expect a less straightforward and cooperative speaker to be able to sneak controversial information into the discourse without it being questioned, simply by couching that information in terms of presupposition rather than assertion. This assumes that what is paramount for the addressability of the information is its status as presupposed or not – that if material is presupposed, it will automatically be less questionable and addressable than if it had been asserted.

An alternative viewpoint is argued by Simons et al. (2010), who stress the importance of **(not-) at-issueness** in understanding presupposition projection. On their account, the crucial distinction is not that between presupposed and asserted content; rather, it is the distinction between material that is at-issue and that which is not-at-issue, where at-issueness is understood relative to the Question Under Discussion (QUD) in the sense of Roberts (1996). The crucial feature in determining at-issueness is whether the utterance addresses the QUD, which is defined as the accepted question for the interlocutors at that moment – that is, the question for which the interlocutors are presently committed to finding the answer.

As a generalisation, presupposed content tends not to be at-issue, for the obvious reason that material that is already part of the common ground isn't usually a good candidate for settling any open questions. However, in principle, novel presupposed content (for instance, where a speaker expects to exploit accommodation) can be at-issue, as a speaker could use it to answer the QUD. Consider the exchange (10)-(11).

Have you ever worked in Berlin? (10)

I quit my job at the Humboldt University last year. (11)

Taking (10) at face value as the QUD, (11) answers indirectly by (formally) presupposing that the speaker had a job in Berlin. However, this material is clearly at-issue, as it does indeed answer the QUD, which the non-presupposed

¹ <https://ottawacitizen.com/opinion/columnists/cohen-what-everybody-knows-about-america>, retrieved 30 May 2019

content (that the speaker doesn't currently work at the Humboldt University) does not.

In a similar spirit, there are various politeness formulae that can be used to introduce novel content but which do so in a way that is formally presuppositional, as in (12).

Miss Otis regrets she's unable to lunch today. (12)

Uttered by a waiter to someone sitting in a restaurant awaiting their lunch companion (as in the Cole Porter song), the main contribution of (12) is to convey that the person in question will not be attending. Although there is no explicit QUD, the implicit QUD seems more likely to concern whether Miss Otis will attend than whether Miss Otis regrets anything. Hence, the presupposed content of (12) appears to be at-issue. In cases such as (11) and (12), we could hardly say that the speaker is being "inappropriate" or "underhanded" in the way they introduce new content into the discourse, even though they are doing so via clearly presuppositional means, from a formal perspective. Yet it is still possible that using presupposition in this way has consequences for the addressability of the new content in the subsequent discourse, depending on the extent to which it is at-issueness rather than presuppositionality than determines addressability.

We can distinguish two positions on this question that represent the ends of a spectrum of possibilities. If addressability is purely a matter of at-issueness (as the name rather suggests), then whether material was formally asserted or presupposed should be irrelevant to how and whether a subsequent speaker can take it up as a topic of discussion. Note that in these cases asserted content is also present in the discourse turn, and this might still interfere with a subsequent speaker's attempts to address the presupposed content, potentially requiring them to use a "Hey, wait a minute"-style circumlocution. At the other end of the spectrum, addressability might be purely a matter of the status of the material in terms of whether it is asserted or presupposed, with at-issueness being moot as far as subsequent discourse turns are concerned.

In this paper, we tackle the issue of addressability by presenting an experiment designed to tease apart the contributions of these two factors, at-issueness and presuppositional status. We do so by constructing a scenario in

which a (confederate) speaker presents material that is at-issue but which is sometimes couched as assertions and sometimes as presuppositions, and in which the participant is encouraged to identify and rebut the falsehoods in the confederate's utterances. In this way we explore, firstly, whether the participant is equally able and inclined to challenge erroneous material when presented as assertion or presupposition (that is, whether the confederate is able to insert controversial material into the discourse by making it presuppositional, controlling for at-issueness), and secondly, whether the status of the challenged material as assertion or presupposition influences its addressability, as measured by the directness with which the participant is able to challenge it, when they choose to do so.

3 Experiment

In this experiment, participants role-played a dialogue with a confederate. The scenario was a police interrogation, in which the participant played the role of the detective and the confederate played the role of a suspect in a robbery. Participants were instructed to ask the suspect specific questions and identify and challenge lies in the suspect's responses. The aim was to investigate whether participants would respond the same way to false information given in the form of presupposition and in the form of assertion, controlling for QUD by ensuring that the same question was asked and the same answer provided in each case.

3.1 Materials and design

Participants were provided with instructions which included the cover story and a list of 19 questions which they were instructed to ask in sequence. Eight of these questions were target items in which the confederate's response contained false content, packaged either in the form of an assertion (four items) or a presupposition (four items), see Appendix A for the full set of items. Participants were randomly allocated to one of two lists of experimental items, which differed only in how the confederate was instructed to respond to these critical items, e.g. the first question was responded to with an asserted falsehood in version 1 and with a presupposed falsehood in version 2, and so on.

The presupposition triggers used represented a wide range of trigger types (stop, know, regret, discover, return, only, to be annoyed, to be happy), reflecting the variability among triggers documented by much prior research (see Schwarz 2019 for recent discussion), which was not a focus of this study. The confederate's responses to the other 11 filler questions were the same (asserted truths) in both versions of the task. The critical items are included, in both versions, in Appendix A.

Corresponding to each question, the participant had also been provided with a note describing the information currently known to the police, and instructed to challenge any statement that contradicted that information. The confederate's initial responses were scripted; she was instructed to admit the 'truth' if challenged on any point. Participants' responses were audio-recorded and later transcribed and analysed.

3.2 Participants

50 participants (aged 18-39) of which 46% were female were recruited in Edinburgh and paid for their participation. The only criterion was that they should self-identify as native speakers of English.

3.3 Results

Across the critical items, participants objected to the false content in 89% of items in which it was asserted and in 79% of items in which it was presupposed. We conducted a mixed-effects logistic regression, postulating a main effect of content type, to examine whether this difference was significant. The model with maximal random effects structure failed to converge and iterative reduction in RE structure yielded a converging model with only by-subject and by-item random intercepts. The model disclosed a significant effect of content type ($\beta = 0.752$, $SE = 0.297$, $p = 0.012$ by likelihood ratio test), indicating that false asserted content was objected to more often than false presupposed content.

For the cases in which participants did object to the content, the length of their response was measured in two ways: by the number of words uttered, and by the number of hesitations or verbal dysfluencies identified. The former measure was designed directly to investigate the claim that presupposed material would be less addressable in the sense of a speaker requiring more words to object to it (as exemplified by the "Hey, wait a

minute" test). The latter measure aimed to explore whether there was evidence of greater cognitive load in cases where speakers were obliged to respond to less addressable content, building on work by Loy, Rohde and Corley (2018) showing an increase in dysfluencies in scenarios involving deception.

We conducted two mixed-effects linear regressions, taking as dependent variables the number of words and number of dysfluencies produced, and postulating again a main effect of content type in each case. A model with maximal random effects structure was used to predict the number of words uttered, and a model with by-subject random slopes and intercepts was conducted to predict the number of dysfluencies. There were no significant differences in number of words uttered ($\beta = 0.96$, $SE = 1.117$, $p = 0.367$ by likelihood ratio test) or number of hesitations/verbal dysfluencies between conditions ($\beta = -0.037$, $SE = 0.086$, $p = 0.66$ by likelihood ratio test), suggesting that no extra linguistic effort was required to object to presupposed content.

4 Discussion

Our experiment was designed to investigate whether the presentation of controversial content as presupposition rather than assertion influenced how it was responded to, when controlling for at-issueness with respect to the QUD. The results suggested that, across the board, there was indeed a dispreference for objecting to presupposed content – that is, from a speaker's perspective, it is possible to forestall objections to false material to a certain extent by making it presuppositional, even in a context in which such objections are socially sanctioned. However, there was little evidence that speakers had difficulty in formulating objections to presupposed content, when they did choose to engage with it: there was no significant difference between responses to presupposed and asserted content with respect to utterance length and dysfluencies.

With respect to the first result, we must acknowledge that participants were generally effective in identifying and challenging falsehoods throughout the experiment, and that the majority of false presuppositions did elicit challenges. However, some QUD-addressing false presuppositions were nevertheless allowed to stand, suggesting that presuppositions do tend to be

less addressable than assertions per se. One possible explanation for this would be that the presuppositional materials are more complex than their purely assertional counterparts, because they contain asserted content that does not transparently address the QUD as well as presuppositional content that does.

One way of testing such an explanation in future work would be to look for systematic differences between participants' behaviour with different presupposition triggers, because triggers vary in the kind of relationship that they encode between the presupposition and assertion, as discussed by Sudo (2012) and Klinedinst (2012). Compare the exchanges (13)-(14) and (15)-(16).

Did Mary argue with her boss? (13)

She regrets doing so. (14)

Did John use to smoke? (15)

He quit recently. (16)

With the trigger *regret*, as in (14), the presupposition (that Mary argued with her boss) answers the QUD directly, but the assertion (that Mary regrets arguing with her boss) entails the presupposition and hence also answers the QUD. With the trigger *quit*, as in (16), the presupposition (that John used to smoke) answers the QUD, but what is sometimes taken to be the assertion (that he does not currently smoke) does not answer the QUD.

Consequently, in a *regret*-type case, one could argue that the presupposed content is not effectively 'concealed' as it is also entailed by the assertion, and therefore we would expect a high proportion of challenges to false presuppositions in such a case. In a *quit*-type case, the presupposed content is independent of the assertion and therefore potentially less salient, and less addressable. However, our experiment does not license us to explore this question in detail as each trigger occurred in just one sentence, risking confounds with item effects.

With respect to the participants' behaviour in cases where they challenge false material, our results appear to support the at-issueness account of Simons et al. (2010). There is no indication that participants felt obliged to use circumlocutions in order to challenge presupposed but at-issue

content: these materials, at least in this context, did not appear to elicit "Hey, wait a minute"-style behaviour from our participants. This may be illustrated by taking a closer look at participants' objections towards both false presupposed (17)-(21) and false asserted content (22)-(26).²

Condition: Presupposed content

Q: Have you held any other positions? (17)

A: I stopped working for the national gallery in Russia in 2017. (18)

P1: Was that not in Shenzhen China? (19)

P2: That's not true. (20)

P3: Okay um how long were you in Russia for? (21)

Condition: Asserted content

P: Have you held any other positions? (22)

S: I used to work for the national gallery in Russia until 2017. (23)

P4: Russia or Shenzhen in China? (24)

P5: That's not true you were working in China. (25)

P6: Why did you leave? (26)

In both conditions, participants object rather directly to the falsehood of the suspect's claim to have worked in Russia: compare (19)-(20) with (24)-(25). Hence, the "Hey, wait a minute" test may be mainly sensitive to the informational status rather than the presuppositional status of content. Furthermore, from a qualitative point of view, similar objection strategies were used independently of the content's presuppositional status: participants objected by asking follow-up questions that addressed the false content (19)/(24), by raising the issue that the suspect lied (20)/(25), or by asking indirect follow-up questions (22)/(26).

Taking both results into account, it seems that in order to predict whether content is available for subsequent discussion warranting discourse coherence one has to account for both the presuppositional status and the at-issueness of content. The approach of Abrusán (2011), further developed in Abrusán (2016), reconciles these two aspects by claiming that although hearers pay attention to certain aspects of meaning by default, their attention may be shifted by contextual cues.

² The following dialogue examples begin with the experimental item, i.e. the question-answer pair, see (17)-(18)

and (22)-(23), and are followed by objections of specific participants, named P1, P2, P3 etc., see (19)-(21).

Despite being developed for predicting presupposition projection this account seems applicable to our scenario: presupposed content is accommodated by default, but as soon as the hearer's attention is broadened by contextual cues, the content is available for further discussion to the same extent as asserted content. As regards the potential differences between presupposition triggers, Abrusán (2016) claims that the complements of factives like *know* can be brought to the focus of attention more easily than the complements of emotive factives like *regret*, since in the latter case, hearers direct their attention towards the attitude holder instead of the complement's content. In accordance with our reasoning above, the pre-state implicature of the presupposition triggered by *stop* is claimed to be focused even less easily, 'concealing' the presupposed content more effectively. But again, differences between presupposition triggers remain to be investigated in future work.

Clearly we should exercise caution about interpreting these results, in that the use of this novel paradigm gives rise to questions about the naturalness of the participants' elicited behaviour. The kind of objections elicited by the false statements in this paradigm might be atypical for at least two (contradictory) reasons. Firstly, our participants may have been unusually willing to flatly contradict false presuppositions because they were aware that the scenario placed them (playing the police officer) in a position of power relative to the confederate (playing the suspect) and entitled them to change the subject and discuss any issue that they wished to, rather than adhere to the topics foregrounded by the confederate. Secondly, our participants may have been uncomfortable at the task of repeatedly contradicting or challenging their interlocutor and started using circumlocutions and unnecessary politeness formulae when questioning assertions as well as presuppositions. On a similar note, the interrogation setting may have encouraged our participants to actually back up their objections, with concrete evidence which led to longer objections overall, see (27)-(32).

Condition: Presupposed content

Q: Have you held any other positions? (27)

A: I stopped working for the national gallery in Russia in 2017. (28)

P7: That's interesting I have here in my document that you were an employee at the national gallery in Shenzhen in China. (29)

Condition: Asserted content

Q: Have you held any other positions? (30)

A: I used to work for the national gallery in Russia until 2017. (31)

P8: That's interesting cause of right now we have/ at least on my record it says that you were at the national gallery in Shenzhen in China. (32)

Nevertheless, the potential advantage of this paradigm is that it creates a scenario in which repeated false statements are made, each for a clearly-motivated reason, and in which these falsehoods can be challenged naturalistically without violating politeness norms.

5 Conclusion

The experimental results presented in this paper suggest that, when material is at-issue, it can be challenged directly by a subsequent speaker whether it is formally asserted or presupposed. However, expressing at-issue material through presupposition rather than assertion appears to have the effect of reducing the frequency of such challenges. These findings are consistent with a view on which speakers are able to manipulate their interlocutors' ability to address discourse content to some extent through the formal apparatus of presupposition, but where material that is relevant to the Question Under Discussion is usually available for subsequent challenge to quite a pronounced extent. Thus, a speaker-hearer model that predicts what material is eligible to discuss in the subsequent dialogue must account both for interlocutors' expectations about information packaging as well as about the overall discourse topic.

Acknowledgments

This work was supported by the Scottish Graduate School for Arts & Humanities in conjunction with Scottish Funding Council.

References

- Márta Abrusán. 2011. Predicting the presuppositions of soft triggers. *Linguistics & Philosophy*, 34:491-535.
- Márta Abrusán. 2016. Presupposition cancellation: Explaining the 'soft-hard' trigger distinction. *Natural Language Semantics*, 24:165-202.

Kai von Fintel. 2008. What is presupposition accommodation, again? *Philosophical Perspectives*, 22(1):137–170.

Michael Glanzberg. 2005. Presuppositions, truth values and expressing propositions. In Gerhard Preyer and Georg Peter (eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford University Press, Oxford, pages 349–396.

Nathan Klinedinst. 2012. THCSF. Ms., UCL.

David Lewis. 1979. Scorekeeping in a language game. In Rainer Bäuerle, Urs Egli and Arnim von Stechow (eds.), *Semantics from Different Points of View*. Springer, Berlin, pages 172–187.

Jia Loy, Hannah Rohde, and Martin Corley. 2018. Cues to lying may be deceptive: speaker and listener behaviour in an interactive game of deception. *Journal of Cognition*, 1(1):42.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. In Jae-Hak Yoon and Andreas Kathol (eds.), *Ohio State University Working Papers in Linguistics, Volume 49*. Ohio State University Publications, Columbus, OH, pages 91–136.

Florian Schwarz. 2019. Presuppositions, projection and accommodation. In Chris Cummins and Napoleon Katsos (eds.), *Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford University Press, Oxford, pages 83–113.

Benny Shanon. 1976. On the two kinds of presuppositions in natural language. *Foundations of Language*, 14:247–249.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In Nan Li and David Lutz (eds.), *Proceedings of the 20th Conference on Semantics and Linguistic Theory*. CLC Publications, Ithaca, NY, pages 309–327.

Yasutada Sudo. 2012. On the semantics of phi features on pronouns. Massachusetts Institute of Technology PhD thesis, Cambridge, MA.

	first approach you in 2018?	at a gallery opening in November.	gallery opening in November.
to be happy	We are interested in what happened on the 2nd of September. What did you do on that day?	I was happy that I had time to finish a journal article at home.	I had time to finish a journal article at home.
discover	Now, we would like to know more about the forth of September when you went to the locksmith. What happened there?	I discovered that the key to my flat needed replacing.	The key to my flat needed replacing.
regret	How was the race on the 17 th of October?	I regret that I didn't go to that race.	I didn't go to that race.
to be annoyed	Did you meet John Smith on that day?	I was annoyed that I bumped into him unexpectedly.	I bumped into him unexpectedly.
only	Did you stay at the race-course after going to the bar?	I only went to the bathroom at four pm.	I went to the bathroom at four pm.
return	We would like to know more about the 17 th of October. What else did you do that day?	I returned to my office.	Before and after the race I was at the office.

Appendix A. Experimental Items

Trigger	Question	Condition: Presupposed content	Condition: Asserted content
stop	Have you held any other positions?	I stopped working for the national gallery in Russia in 2017.	I used to work for the national gallery in Russia until 2017.
know	When did John Smith	I know that I first saw him	I first saw him at a

Implicatures in continuation-based dynamic semantics

Florrie Verity

Research School of Computer Science
Australian National University
Canberra, Australia
florrie.verity@anu.edu.au

Abstract

Advances have been made towards interpreting context-dependent meaning in a logical form, but treatments of implicatures remain incomplete. This paper captures implicature-related meaning in Lebedeva’s (2012) extension of de Groote’s continuation-based dynamic semantics (2006), exploiting the fact that context is incorporated as a parameter, meaning its structure may be altered while preserving the properties of the framework. The new context structure is a simple logic of common-sense reasoning using Poole’s (1988) framework for classical logic that switches from *reasoning as deduction* to *reasoning as theory formation*. Focusing on *but* and supplementary content, a treatment of implicatures in a compositional framework – using only common tools from logic – is proposed. This is situated within the goal of formally accounting for presupposition, conversational implicature and conventional implicature in a single semantics.

1 Introduction

The *dynamic turn* in natural language semantics, attributed to Heim (1982) and Kamp (1981), relocated the meaning of a sentence from the logical form itself to its *context change potential*, interpreting new sentences in the context of those preceding. This enabled the interpretation of context-dependent meaning such as the referent of a pronoun, which had eluded the prevailing paradigm of Montague semantics (1970a; 1970b; 1973).

The dynamic semantics of de Groote (2006), as extended by Lebedeva (2012), goes further by incorporating *continuations* from programming language semantics (Strachey and Wadsworth, 1974) for a second notion of context as *the future of the discourse*. The result is a dynamic semantics in the style of Montague that firmly separates the context from the content of a sentence, uses only com-

mon mathematical tools – providing more insight than ad hoc definitions – and is entirely compositional – the meaning of a sentence is determined by the meanings of its constituents and its syntactic structure, allowing for the automatic interpretation of complex expressions. Furthermore, since both kinds of context are abstracted over the meaning of the sentence, the structure of the context is flexible – for example, a list of names (de Groote, 2006) or a conjunction of propositions (Lebedeva, 2012).

This paper exploits the flexibility of context by considering not just interaction with the context, but interaction *within* the context, to locate implicatures. Implicatures are situated in a group of meaning classes characterized by existing outside the plain semantic content of an utterance. Also in this group is presupposition – meaning assumed by an utterance for it to be meaningful – as in ‘John quit smoking’, which relies on John having smoked to make sense. If this presupposed information is not in the discourse context, it is accommodated alongside the plain content of the sentence. Implicature refers to meaning outside of what is explicitly said, logically entailed or presupposed by an utterance. It is traced back to Frege (1879) and was brought to prominence by Grice’s (1975) treatment that introduced a provisional division – with prevailing terminology – between *conversational implicature*, governed by principles of cooperative conversation such as utterances being relevant to what has come before, and *conventional implicature*, instead associated with particular words – *but*, for example, is said to implicate a contrast between two clauses, while not explicitly stating this contrast.

If these meaning classes and their distinctions seem murky, it is because they are. Potts’ (2015) survey of these phenomena contends that their definitions are “still hotly contested” and suggests

refocusing towards developing “rich theories of properties... the way those properties interact, and the effects of those interactions on language and cognition.” Lebedeva’s extension of de Groote’s framework goes some way towards this by accounting for presuppositions of referring expression and proposes a mechanism for handling conversational implicatures. Treatment within the same framework allows a preliminary formal distinction between presuppositions and certain kinds of conversational implicatures to be made.

This paper goes further by distilling Lebedeva’s approach of *conversational implicatures by proof-theoretic abduction* to *implicatures by reasoning in the context*. By elaborating the context structure to a logical theory using Poole’s (1988; 1989; 1990) classical logic framework for reasoning as theory formation, meaning associated with *conventional* implicatures is captured while preserving the features of compositionality and the use of common mathematical tools. In this paper, it is used to formalize an intuition about supplementary content, revealing proximity to conversational implicature, and provide a treatment of *but*.

Section 2 proceeds by detailing the problems of capturing implicatures. Section 3 provides the formal background: the continuation-based dynamic semantics in use, the approach to conversational implicatures by proof-theoretic abduction from which this work stems, and Poole’s framework for reasoning as theory formation. Section 4 adapts this framework for natural language interpretation and uses the new context structure to solve the problems from Section 2.

2 Implicatures

This section gives a pre-formal presentation of implicatures and their challenges to formal semantics. Although divided into conversational and conventional varieties, the formalization will approach them in the spirit of Potts’ aforementioned call to move from labels to rich theories of properties. Recall also that the solution we seek to these problems is one that is compositional and uses only common tools from logic, distinguishing it from other approaches.

2.1 Conversational implicature

Consider how A may interpret B’s statement in the following discourse from Grice (1975):

- (1) A: Smith doesn’t seem to have a girlfriend these days.
B: He has been paying a lot of visits to New York lately.

Assuming that B is a cooperative speaker, providing content relevant to A’s statement, B’s response contains meaning outside of Smith’s visits to New York. Suppose A believes having a girlfriend in a different city is a reason for frequently visiting that city, then A takes B to mean Smith has a girlfriend in New York. This is the conversationally implicated meaning Grice associates with (1) and is challenging to capture because it is not associated with a particular lexical item.

2.2 Conventional implicatures

Formalizing conventional implicatures is complicated by the fact that the term is used to refer to a diverse body of lexical items, has at least two very distinct characterizations, and is the subject of prominent claims of non-existence (Bach, 1999). The Gricean conventional implicatures (Grice, 1975) have been expanded to include adverbs *already, only, also, yet*; connectives *but, nevertheless, so, therefore*; implicative verbs *bother, manage, continue, fail*; and subordinating conjunctions *although, despite, even though*. We focus on *but* as a canonical example.

But

But is often thought of as contrasting two clauses, as in the following example from Bach (1999):

- (2) Shaq is huge but he is agile.

Classical treatments of *but* follow a standard template for conventional implicatures observed in (Potts, 2015) of associating independent dimensions of meaning with a word. In the case of *but*, this is the pair $(p \wedge q, R(p, q))$, where R represents a relation of contrast between p and q .

The contrast need not be between the two clauses joined by *but*, however. The contrast in (3) is reasons for and against inviting Robinson:

- (3) A: Robinson always draws large audiences.
B: He always draws large audiences, but he is in America for the year.

Returning to (2), Bach (1999) considers “the most natural way of taking *but* especially out of context” is “as indicating that being huge tends to preclude being agile”. However, it is not clear

whether “out of context” means by the conventional meaning of words alone, or additionally implies some knowledge of the world – that people exist in restricted spaces surrounded by objects that make swift movement easier for smaller people, or that it is biologically the case that great size generally precludes agility. To clarify, consider the following variations:

(4) Shaq is huge but he is rich.

(5) Shaq is huge but he is small.

Utterance (2) is comparable to (4) and different to (5) as only the latter contains a *conventional* contrast – based on the meaning of words alone – but is infelicitous for this very reason. Utterance (4) appears infelicitous “out of context”, unlike (2), but not in a highly specific context: consider a conversation between Shaq’s friends about who to invite on an expensive caving holiday. Speaker B suggests inviting Shaq, to which it is replied:

(6) A: Shaq is huge! He’s too big to go caving.
B: Shaq is huge but he is rich.

The challenge is to account for these context-dependent conditions on felicitousness.

Supplements

The second characterization of conventional implicatures is Potts’ (2005) reformulation, motivated by a dearth of formal treatments and based on Grice’s remarks but divorced from the notion of implicature – enforced by called them ‘CIs’. The formulation of CI as speaker-oriented commitments that are part of the conventional meaning of words and logically independent from at-issue content, is evidenced not by the classical examples above but expressives, such as ‘damn’ and supplemental expressions, underlined in the following example:

(7) Ed’s claim, which is based on extensive research, is highly controversial.

While Potts’ multidimensional logic for handling CIs spurred interest in formalizing this meaning class, it largely did not extend to Gricean conventional implicatures, and relationships to conversational implicatures remain unexplored, as in Potts’ interpretation of (7):

With the CI content expressed by the supplementary relative, I provide a clue

as to how the information should be received. This example is felicitous in a situation in which, for example, I want to convey to my audience that the controversy should not necessarily scare us away from Ed’s proposal – after all, it is extensively researched. (Potts, 2005)

The problem here is in explaining the proximity of Potts’ description of CIs to Grice’s notion of implicature as meaning outside of what is explicitly said, formalizing Potts’ intuition that CIs provide “a clue as to how the information should be received”.

3 Formal background

We proceed by introducing the natural language semantics to be used for these problems, the proposal for capturing conversational implicatures in this semantics by proof-theoretic abduction, and the framework for common-sense reasoning that will be used to generalize this approach.

3.1 Continuation-based semantics $GL\chi$

The continuation-based dynamic semantics $GL\chi$ (Lebedeva, 2012) is a version of de Groote’s Montagovian account of dynamics (2006), enhanced by a systematic translation from static to dynamic interpretations and an exception raising and handling mechanism for capturing presuppositions. The interpretation of a sentence is a logical form in a λ -calculus, built compositionally from individual lexical items, such as the following:

$$\llbracket \overline{\text{loves}} \rrbracket = \lambda Y X. X(\lambda x. Y(\lambda y. \overline{\text{love}} x y)) \quad (8)$$

$$\llbracket \widetilde{\text{John}} \rrbracket = \lambda P. (\text{sel}(\text{named “John”})) \quad (9)$$

$$\llbracket \widetilde{\text{Mary}} \rrbracket = \lambda P. (\text{sel}(\text{named “Mary”})) \quad (10)$$

Term (8) is analogous to the static interpretation of *loves* in Montague semantics, except that $\overline{\text{love}}$ abbreviates a systematic dynamization of *love*:

$$\begin{aligned} \overline{\text{love}} &= \lambda e \phi. \text{love}(xe)(ye) \\ &\quad \wedge \phi(\text{upd}(\text{love}(xe)(ye), e)) \end{aligned}$$

This is dynamic in the sense that it is parameterized by two contexts: e is the *left context*, made of background knowledge and preceding sentences, and ϕ is the *right context*, made of the discourse to come. The right context is formally a *continuation* (Strachey and Wadsworth, 1974), invented

for compositionality problems in the semantics of programming languages. Function **upd** adds new content to the context, while **sel**, in terms (9) and (10), selects a referent from the context satisfying a certain property – such as being named “John”.

With these terms, the sentence *John loves Mary* can be interpreted compositionally by β -reduction of the following term:

$$\begin{aligned} & (\overline{[\![\text{loves}]\!]}) (\overline{[\![\text{Mary}]\!]}) (\overline{[\![\text{John}]\!]}) \rightarrow_{\beta} \lambda e \phi. \\ & \mathbf{love} (\mathbf{sel}(\text{named “John”})e) (\mathbf{sel}(\text{named “Mary”})e) \\ & \wedge \phi(\mathbf{upd}(\mathbf{love} (\mathbf{sel}(\text{named “John”})e) \\ & \quad (\mathbf{sel}(\text{named “Mary”})e), e)) \end{aligned} \quad (11)$$

Dynamic semantics is concerned with discourse, rather than individual sentences. Suppose we have a context containing Mary and John, formally:

$$\mathbf{c} = \exists j. \text{named “John” } j \wedge \exists m. \text{named “Mary” } m$$

Then interpretation in context \mathbf{c} is found by applying the sentence-level interpretation (11) to \mathbf{c} , β -reducing and evaluating the oracle functions to find the referents of Mary and John:

$$\lambda \phi. \mathbf{love } j \ m \wedge \phi(\mathbf{upd}(\mathbf{love } j \ m, \mathbf{c})) \quad (12)$$

If appropriate referents cannot be found, an exception is raising and handled by introducing new individuals to the context (see (Lebedeva, 2012) for further details).

Suppose the discourse continues with the sentence *He smiles at her*. Then it has the following interpretation, found by applying (12) to the sentence-level interpretation of *He smiles at her*:

$$\begin{aligned} & \lambda \phi. \mathbf{love } j \ m \wedge \mathbf{smiles-at } j \ m \\ & \wedge \phi(\mathbf{upd}(\mathbf{smiles-at } j \ m, \mathbf{upd}(\mathbf{love } j \ m, \mathbf{c}))) \end{aligned}$$

Since we will be using a context structure to capture implicatures, we are only interested in the last subterm of this expression – the incremental context update. In $GL\chi$, the context is treated as a conjunction of terms, so $\mathbf{upd}(\mathbf{t}, \mathbf{c})$ simply adds term \mathbf{t} to context \mathbf{c} by conjunction, as in:

$$\mathbf{upd}(\mathbf{love } j \ m, \mathbf{c}) = \mathbf{c} \wedge \mathbf{love } j \ m$$

Since context is defined as a parameter, its structure – and the definition of context update – may be changed while otherwise preserving the properties of the framework, including compositionality.

3.2 Conversational implicatures by proof-theoretic abduction

Our starting point for treating implicatures in framework $GL\chi$ is Lebedeva’s (2012) proposal for conversational implicatures by proof-theoretic abduction. Abductive reasoning is adopting a statement because it provides an explanation for another statement known to be true: where deduction is the conclusion of q from p and $p \Rightarrow q$, abduction is the conclusion of p from q and $p \Rightarrow q$. Such reasoning is *defeasible*, in the sense of being open to revision.

Although logically invalid, abduction is prolific in human reasoning and Hobbs et al. (1993) argue that it is inherent in interpreting discourse, based on the hypothesis that “it is commonplace that people understand discourse so well because they know so much” (Hobbs et al., 1993). To interpret B’s remark in (1) requires not just knowledge of the meaning of words but knowledge of the world – specifically that people spend time with their partners and seeing someone who lives elsewhere requires visiting them. This knowledge means reasoning occurs when new information is encountered, motivating the use of proofs to capture natural language meaning.

This is incorporated into $GL\chi$ via the definition of a handler for an exception raised when a proposition cannot be proved from the context of background knowledge and preceding sentences. This implements – in a compositional framework using only familiar logical tools – the idea from Hobbs (2004) of computing implicatures by attempting to prove the logical form of a sentence, taking as axioms formulae corresponding to the current knowledge base. If no proof is found, the facts necessary to complete the proof are added to the knowledge base via abduction. These abduced facts correspond to the implicatures of the sentence.

We develop this proposal in two ways. Firstly, the approach – left generic to demonstrate a concept – inherits the computational problems of both proof search and abduction, such as monotonicity. An implementation requires choosing a logic for abduction while preserving the original principles of the framework, namely the use of standard logical tools. To this end, we consider abduction *outside* of a proof-theoretic approach, observing that this is not intrinsic to the proposal and has the disadvantage of automatically excluding

other ways of implementing abduction, such as a forward-reasoning system.

The second development is incorporating reasoning more broadly. Once one notion of reasoning has been introduced to the context, it becomes clear that interpretation can depend on deductive inference from content in the context, as well as *induction* – another form of defeasible reasoning. Inductive reasoning takes several cases of p and q occurring together to conclude $p \Rightarrow q$, and can be cast as *default reasoning* – as in ‘ q usually follows from p ’. It is then necessary to account for how defeasible and non-defeasible information interact.

3.3 The Theorist framework

Based on this, we want a logic of defeasible reasoning with good computational properties and using familiar mathematical tools. For this, we choose Poole’s logical framework for default reasoning (Poole, 1988), further developed in (Poole, 1989, 1990) and including an implementation called *Theorist*. It is a semantics for classical logic that considers reasoning not as deduction but as *theory formation*. This is achieved by allowing hypothetical reasoning, and so handles nonmonotonic reasoning in classical logic.

Given a standard first-order language over a countable alphabet, *formula* refers to a well-formed formula over this language and an *instance of a formula* refers to a substitution of free variables in a formula by terms in the language. The following sets are provided: F of closed formulae thought of as ‘facts’, Δ and Γ of (possibly open) formulae constituting the hypotheses – defaults and conjectures respectively – and O of closed formulae of observations about the world.

The semantics has three definitions at its core. A *scenario* of $(F, \Delta \cup \Gamma)$ is a set $D \cup G$, where D and G are ground instances of elements of Δ and Γ respectively, such that $D \cup G \cup F$ is consistent. An *explanation* of a closed formula t from $(F, \Delta \cup \Gamma)$ is a scenario of $(F, \Delta \cup \Gamma)$ that implies t . An *extension* of (F, Δ) is the logical consequences of a maximal (with respect to set inclusion) scenario of (F, Δ) , that is, the closure under modus ponens $F \cup \bar{D}$ for some maximal set D of ground instances of Δ . With these definitions in hand, a *state* of the system is a tuple $\langle F, \Delta, \Gamma, O, \mathcal{E} \rangle$ where \mathcal{E} is the set of explanations of the observations in O .

Note that there can be multiple extensions of a

scenario, and so a formula g is *predicted* by (F, Δ) if g is in every extension of (F, Δ) . See (Poole, 1989) for other possible definitions of prediction and a discussion of different ways of computing explanations; we follow (Poole, 1990) in taking our explanations to be least presumptive (not implying other explanations) and minimal (not containing other hypotheses).

To illustrate, consider the following example from (Poole, 1989) of medical diagnosis. Suppose the starting state is $\langle F, \Delta, \Gamma, \{\}, \{\} \rangle$, with:

$$\begin{aligned} F &= \{\text{broken (tibia)} \Rightarrow \text{broken (leg)}\} \\ \Delta &= \{\text{broken (leg)} \Rightarrow \text{sores (leg)}\} \\ \Gamma &= \{\text{broken (leg)}, \text{broken (tibia)}\} \end{aligned}$$

If **sores (leg)** is observed, the new state is $\langle F, \Delta, \Gamma, \{\text{sores (leg)}\}, \{E_{\text{leg}}\} \rangle$, where:

$$E_{\text{leg}} = \{\text{broken (leg)}, \text{broken (leg)} \Rightarrow \text{sores (leg)}\}$$

Another possible explanation is:

$$\begin{aligned} E_{\text{tibia}} &= \{\text{broken (tibia)}, \\ &\quad \text{broken (leg)} \Rightarrow \text{sores (leg)}\} \end{aligned}$$

This is a minimal explanation, but not least presumptive.

Alternatively, suppose that from the initial state **broken (leg)** is observed. Then the new state is $\langle F, \Delta, \Gamma, \{\text{broken (leg)}\}, \{\} \rangle$, and **sores (leg)** is predicted because it is in every extension.

4 Implicatures by reasoning in the context

With the formal background in place, we proceed by adapting Theorist for reasoning in natural language interpretation and use it to solve the problems from Section 2.

4.1 Theorist for implicatures

Using Theorist for implicatures requires categorizing the information in a discourse context. Defaults and conjectures play the same role in our application, while observations, with their incremental update, correspond naturally to content.

More difficult is the question of what constitutes fact – information that we are not prepared to give up. The intuition in a model-theoretic interpretation of knowledge about the world, such as ‘Canberra is in Australia’, is that it is not necessarily true in every model. In the case of natural language, there is information that must be

true in every model – lexical semantic information. Meaning we are not prepared to give up is the meaning of words and relationships between them, such as antonyms and ‘green is a colour’. Thus we take this to correspond to the facts in Theorist. A new set B is added, corresponding to background knowledge and containing any individuals given a priori, or via $GL\chi$ ’s exception handling mechanism. These entities comprise the domain of the context, assumed to be pairwise distinct.

We can now make the following definitions. A *state* of the discourse context is a tuple

$$\langle L, \Delta, \Gamma, O, B, \mathcal{E}, \mathcal{P} \rangle$$

with sets of closed formulae L of lexical semantic information, B of background information and O of discourse content; sets of open formulae Δ of defaults and Γ of conjectures; and sets of sets of closed formulae \mathcal{E} of explanations and \mathcal{P} of predictions. Sets L , B and Δ are provided by the user, and Γ may be given automatically as the set of antecedents of the implications in Δ .

Given a context $\mathbf{c} = \langle L, \Delta, \Gamma, B, O, \mathcal{E}, \mathcal{P} \rangle$, the context update function upd in $GL\chi$ is defined:

$$\text{upd}(\mathbf{t}, \mathbf{c}) = \langle L, \Delta, \Gamma, B', O', \mathcal{E}', \mathcal{P}' \rangle$$

- The new discourse content \mathbf{t} is added to the set of observations:

$$O' := O \cup \{\mathbf{t}\}$$

- The background information is updated with deductive inference from lexical semantic knowledge and the new content:

$$B' := B \cup (\overline{L \cup O'}) \setminus (L \cup O')$$

- The explanation set contains the least presumptive and minimal explanations of O' from $(L \cup B, \Delta \cup \Gamma)$, which takes the form of instances $D \cup G$ of $\Delta \cup \Gamma$.
- For each explanation E_i there is a corresponding prediction set P_i in \mathcal{P}' defined by:

$$P_i = \overline{S_i} \setminus S_i$$

where S_i is the union of the maximal set of ground instances of Δ over the domain, the new discourse content and background, and explanation E_i :

$$S_i = \max(\Delta) \cup B' \cup O' \cup E_i$$

Note that predictions are not made from the set of lexical semantic information since its consequences are not defeasible. Instead, it is placed in the background. Note also conjectures are used in explanation but not in prediction. We will make reference to the *hypotheses* $H(\mathbf{c})$ of a context theory \mathbf{c} – the union of the explanations and predictions.

We return to the problems from Section 2. The computation of sentence-level interpretations are omitted but can be found compositionally in $GL\chi$.

4.2 Interaction of supplementary content

To answer the questions about the supplement in sentence (7) we want to represent the following information: *a claim can be unresearched, an unresearched claim is typically controversial, a controversial claim is typically rejected*. Let the initial context be given by $\mathbf{c}_0 = \langle L, \Delta, \Gamma, B, \{\}, \{\}, \{\} \rangle$, with L, B, Δ and Γ as follows:

$$L = \{\}$$

$$B = \{\exists \lambda e. (\text{named “Ed”})e, \lambda f. \text{claim } f \wedge \text{poss } e f\}$$

$$\Delta = \{\neg \text{researched } x \Rightarrow \text{controversial } x, \\ \neg \text{researched } x \Rightarrow \text{reject } x\}$$

$$\Gamma = \{\neg \text{researched } x\}$$

Consider the sentence with the supplementary content removed:

(13) Ed’s claim is highly controversial.

The context update term of its interpretation in context \mathbf{c}_0 is:

$$\phi(\text{upd}(\text{controversial } f, \mathbf{c}_0))$$

Computing the upd function call to get \mathbf{c}_1 :

$$\mathbf{c}_1 = \langle L, \Delta, \Gamma, B, \{\text{controversial } f\}, \{E_1\}, \{P_1\} \rangle$$

where $E_1 = \{\neg \text{researched } f, \neg \text{researched } f \Rightarrow \text{controversial } f\}$ and $P_1 = \{\text{reject } f\}$. Interpreting (13) in this context predicts that Ed’s claim should be rejected, and proposes it is controversial because it is not well researched.

Now consider inclusion of the supplement. Suppose *which* is given the same interpretation as the plain discourse connective *and*, differentiated only by its syntax.¹ Then the context update term of its interpretation in context \mathbf{c}_0 is:

$$\phi(\text{upd}(\text{controversial } f, \text{upd}(\text{researched } f, \mathbf{c}_0)))$$

¹This interpretation is not sufficient to capture the projection behaviour of supplements, however this is beyond the scope of this paper.

Let $c_1 = \text{upd}(\text{researched } f, c_0)$. Then:

$$c_1 = \langle L, \Delta, \Gamma, B, \{\text{researched } f\}, \{\emptyset\}, \{\emptyset\} \rangle$$

and there is no explanation or prediction in the theory of context. Computing the second context update:

$$\begin{aligned} c_2 &= \text{upd}(\text{controversial } f, c_1) \\ &= \langle L, \Delta, \Gamma, B, \{\text{researched } f, \\ &\quad \text{controversial } f\}, \{\emptyset\}, \{\emptyset\} \rangle \end{aligned}$$

Again, there is no explanation or prediction.

Potts' meaning – not to dismiss Ed's claim on the basis of being controversial – can be located in the difference between the context with and without the supplementary content. To do this, we expand the notion of context change potential to allow comparison of theories of context, formalizing the meaning of Potts' "clue". Significantly, this meaning is not associated with *which*, and so need not be encoded in a lexical item.² This is consistent with Potts' treatment and diagnosis of CI, however, formalizes the interaction of supplementary content with main content in a way that looks like conversational implicature. Thus the formalism proves valuable in identifying different flavours of implicature at play.

4.3 But

Again, the connective in question is assigned the same interpretation as the plain discourse connective *and*, to demonstrate how the meaning associated with *but* emerges through reasoning in the context.

Interpreting utterance (3) shows how this approach can identify a contrast existing outside of the clauses connected by *but*. We want to represent the following context: *being popular is a reason for inviting someone, not being in Oxford is a reason for not being invited and if someone is in Oxford then they are not in America*. Let $c_0 = \langle L, \Delta, \Gamma, B, \{\}, \{\}, \{\} \rangle$, with sets given as

follows:

$$\begin{aligned} L &= \{\} \\ B &= \{\exists \lambda r. (\text{named "Robinson"})r, \text{male } r, \\ &\quad \text{human } r\} \\ \Delta &= \{\text{popular } x \Rightarrow \text{invite } x, \\ &\quad \text{invite } x \Rightarrow \text{in-oxford } x, \\ &\quad \text{in-oxford } x \Rightarrow \neg \text{in-america } x\} \\ \Gamma &= \{\text{popular } x, \text{in-oxford } x\} \end{aligned}$$

The interpretation of (3) in context c_0 includes the following subterm for updating the context:

$$\phi(\text{upd}(\text{in-america } r, \text{upd}(\text{popular } r, c_0)))$$

Beginning with the innermost context update:

$$\begin{aligned} c_1 &= \text{upd}(\text{popular } r, c_0) \\ &= \langle L, \Delta, \Gamma, B, \{\text{popular } r\}, \{\emptyset\}, \{P_1\} \rangle \end{aligned}$$

There is no explanation, but it is predicted that Robinson should be invited, and that he is in Oxford and not America:

$$P_1 = \{\text{invite } r, \text{in-oxford } r, \neg \text{in-america } r\}$$

Performing the second context update:

$$\begin{aligned} c_2 &= \text{upd}(\text{in-america } r, c_1) \\ &= \langle L, \Delta, \Gamma, B, \{\text{popular } r, \text{in-america } r\}, \\ &\quad \{\emptyset\}, \{\emptyset\} \rangle \end{aligned}$$

There is no explanation for the new content and no predictions, because *invite* r is no longer consistent with the context. The meaning of *but* is located in the context change from c_1 to c_2 . Rather than creating a contradiction in the context, a serious problem in a classical logic, Poole's framework prevents contradiction, preserving monotonicity in the context logic while capturing the occurrence of inconsistencies. By modelling the theory of context this way, hard-coding a contradiction into the interpretation of *but*, as in the classical interpretation, becomes redundant – the inconsistency automatically arise in the context theories joined by *but*.

This model suggests viewing *but* as a pragmatic choice of connective to licence an inconsistency from one context to the next. This is not a new idea: it is compatible with a procedural account of meaning (Blakemore, 1987), in which beyond determining truth conditions, connectives guide the inferences made by the hearer of an utterance.

²Encoding in the interpretation of the lexical item could be useful for the problem of automatically generating context, however.

Based on these observations, we propose the following *pragmatic* definition of *but*, in the sense that it is defined on the level of discourse interpretation, as opposed to the semantic interpretation of a lexical item. Suppose *but* conjoins propositions S_a and S_b , with the following context updates:

$$\begin{aligned} c_n &= \text{upd}(\mathbf{a}, c_{n-1}) \\ c_{n+1} &= \text{upd}(\mathbf{b}, c_n) \end{aligned}$$

Then there exists $p \in H(c_n)$ and $q \in c_{n+1}$ such that $p \wedge q \vdash \perp$, that is, there is a defeasible contradiction.

To test this proposal, consider (2) and variations (5) and (4). The context can be formalised as $c_0 = \langle L, \Delta, \Gamma, B, \{\}, \{\}, \{\} \rangle$, with sets given as follows:

$$\begin{aligned} L &= \{\forall x. \mathbf{huge} x \Leftrightarrow \neg \mathbf{small} x\} \\ B &= \{\exists \lambda s. (\text{named "Shaq"} s, \mathbf{male} s, \mathbf{human} s)\} \\ \Delta &= \{\mathbf{huge} x \Rightarrow \neg \mathbf{agile} x\} \\ \Gamma &= \{\mathbf{huge} x\} \end{aligned}$$

In the interpretation of (2) in context c_0 , the following subterm updates the context:

$$\phi(\text{upd}(\mathbf{agile} s, \text{upd}(\mathbf{huge} s, c_0)))$$

Evaluating the innermost context update:

$$\begin{aligned} c_1 &= \text{upd}(\mathbf{huge} s, c_0) \\ &= \langle L, \Delta, \Gamma, B_1, \{\mathbf{huge} s\}, \{\emptyset\}, \{P_1\} \rangle \end{aligned}$$

There is no explanation, but there is a prediction and the background is updated:

$$\begin{aligned} B_1 &= \{\neg \mathbf{small} s\} \\ P_1 &= \{\neg \mathbf{agile} s\} \end{aligned}$$

The new context theory predicts that Shaq is not agile. Performing the second context update:

$$\begin{aligned} c_2 &= \text{upd}(\mathbf{agile} s, c_1) \\ &= \langle L, \Delta, \Gamma, B_1, \{\mathbf{huge} s, \mathbf{agile} s\}, \{\emptyset\}, \{\emptyset\} \rangle \end{aligned}$$

As in the previous example, there is a contradiction between subsequent contexts, with $\neg \mathbf{agile} s \in H(c_1)$ and $\mathbf{agile} s \in c_2$.

Accounting for the infelicitousness of (5), the update from c_0 to c_1 is the same, but the update c_2 is as follows:

$$\begin{aligned} c_2 &= \text{upd}(\mathbf{small} s, c_1) \\ &= \langle L, \Delta, \Gamma, B_1, \{\mathbf{huge} s, \mathbf{small} s\}, \{\emptyset\}, \{P_1\} \rangle \end{aligned}$$

Since lexical semantic consequence is not defeasible, and so is added to the background rather than predicted, $\neg \mathbf{small} s$ remains in the context from c_1 to c_2 . Rather than having a contradiction between contexts, the contradiction is within the context.

For (4), the update from c_1 to c_2 is:

$$\begin{aligned} c_2 &= \text{upd}(\mathbf{rich} s, c_1) \\ &= \langle L, \Delta, \Gamma, B_1, \{\mathbf{huge} s, \mathbf{rich} s\}, \{\emptyset\}, \{P_1\} \rangle \end{aligned}$$

There is no contradiction between c_2 and c_1 , and so the condition under which *but* is the pragmatic choice of connective is not satisfied. However, the context for discourse (6) could be given as *inviting Shaq is a possibility, caving in a remote area is expensive, being rich is a reason for inviting someone on an expensive trip, being huge tends to make caving difficult, being unable to go caving is a reason for not inviting someone on a caving trip*. Then when *he is rich* is added to the context, there will be an inconsistency between subsequent contexts, between a reason to invite Shaq and a reason against inviting Shaq. This illustrates the context-dependence of *but*, and how *GL χ* with reasoning in the context can account for it.

5 Conclusion

The thesis advanced in this paper is that implicature-related meaning – under various labels – can be located by incorporating reasoning into the discourse context. By elaborating the structure of context in de Groote and Lebedeva’s continuation-based dynamic semantics with Poole’s framework for reasoning as theory formation, implicature-related meaning can be interpreted compositionally and with the use of only standard logical tools.

It remains to continue testing this approach on other instances of implicature and to see how locating this meaning in the context can address the projection problem for implicatures, all towards the goal of formally comparing the properties of meaning labelled presupposition, conventional implicature, and conversational implicature, in a single framework.

Acknowledgments

The author wishes to acknowledge discussions with Ekaterina Lebedeva, Bruno Woltzenlogel Paleo and Daniyar Itegulov, which motivated this

work. This research was supported by an Australian Government Research Training Program Scholarship.

References

- Kent Bach. 1999. The myth of conventional implicature. *Linguistics and Philosophy*, 22(4):327–366.
- Diane Blakemore. 1987. *Semantic Constraints on Relevance*. Blackwell.
- Phillipe de Groote. 2006. Towards a Montagovian account of dynamics. *Semantics and Linguistics Theory*, 16.
- Gottlob Frege. 1879. Begriffsschrift. In M. Beaney, editor, *The Frege Reader*, 1997. Oxford: Blackwell.
- H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics*, volume 3, pages 41–58.
- Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.
- Jerry R. Hobbs. 2004. Abduction in natural language understanding. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*. Blackwell.
- Jerry R. Hobbs, Mark E. Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.
- Hans Kamp. 1981. A theory of truth and semantic representation. In Jeroen Groenendijk, Theo Janssen, and Martin Stokhof, editors, *Formal Methods in the Study of Language, Part 1*, volume 135, pages 277–322. Mathematical Centre Tracts, Amsterdam. Reprinted in Jeroen Groenendijk, Theo Janssen and Martin Stokhof (eds), 1984, *Truth, Interpretation, and Information; Selected Papers from the Third Amsterdam Colloquium*, Foris, Dordrecht, pp. 1–41.
- Ekaterina Lebedeva. 2012. *Expressing discourse dynamics through continuations*. Ph.D. thesis, INRIA, Université de Lorraine.
- Richard Montague. 1970a. English as a formal language. In B. Visentini et. al., editor, *Linguaggi nella e nella Tecnica*, pages 189–224. Edizioni di Comunità, Milan.
- Richard Montague. 1970b. Universal grammar. *Theoria*, 36:373–398.
- Richard Montague. 1973. The proper treatment of quantification in ordinary English. In J. Hintikka, J. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language: proceedings of the 1970 Stanford workshop on Grammar and Semantics*, pages 221–242. Reidel, Dordrecht.
- David Poole. 1988. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–47.
- David Poole. 1989. Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence*, 5(2):97–110.
- David Poole. 1990. A methodology for using a default and abductive reasoning system. *International Journal of Intelligent Systems*, 5(5):521–548.
- Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford University Press.
- Christopher Potts. 2015. Presupposition and implicature. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantics Theory*. Oxford: Wiley-Blackwell.
- Christopher Strachey and Christopher P. Wadsworth. 1974. Continuations: A mathematical semantics for handling full jumps. Technical report, Oxford University, Computing Laboratory.

A Framework for Annotating Co-working Dialogues in Complex Task Settings

Emma Barker

Department of Computer Science
University of Sheffield

E.Barker@sheffield.ac.uk

Robert Gaizauskas

Department of Computer Science
University of Sheffield

R.Gaizauskas@sheffield.ac.uk

Abstract

There is immense potential for applications involving human-machine co-working. Building dialogue agents for co-working is a key challenge to realising this potential. We propose schemes to annotate co-working dialogues at two levels – the task level and the interaction level – in order to better understand the structure and nature of co-working dialogues and to provide the basis for annotating texts for training dialogue agent models. Our work is grounded in an analysis of part of the Apollo 11 air-to-ground mission transcripts.

1 Introduction

Many of human-kind’s most impressive accomplishments – from building the Pyramids to landing a man on the moon or photographing a black hole – are the result of *co-working*, the coordinated activity of two or more human agents working together to achieve a common goal. Communication between co-workers is an essential part of co-working and in general the most efficient and effective form of communication for co-working is spoken dialogue. We refer to linguistic interaction between co-workers whose function is to directly facilitate or enable progress towards the common goal as *co-working dialogue*.¹

Why do co-working agents A and B need to communicate? Reasons to do so include: to co-ordinate actions that need to be carried out in a certain order or at a certain time (e.g. A and B both need to push together to jump start the car); to provide or query prior knowledge that A has about

the task and B does not (e.g. expert and novice mechanics working together on a car engine); to provide or check the perspective that A has and B does not (e.g. A is above the engine looking down, B underneath the engine looking up); to divide the roles in a multi-agent task or divide the tasks in a parallelisable task; to schedule tasks over a coming work period. Dialogue can also play an important role in maintaining social relations between co-workers, building trust, camaraderie, and so on, and therefore contributes indirectly to task completion; but here we focus on the task-oriented aspects of dialogue in co-working.

Why study co-working dialogues? As an important and ubiquitous sub-type of dialogue, co-working dialogues are, of course, worthy of study in their own right. However, there are also important practical reasons for studying them. Currently there is considerable excitement around the potential for human-machine co-working, where the machine may be a robot or a disembodied intelligent agent (cf. the Industry 4.0 vision (Hermann et al., 2016)). For example, a human and robot might work together in a manufacturing setting, where the robot is doing the heavy lifting under direction of a human; or, a human might be repairing a complex electrical or mechanical fault while an agent provides relevant information, e.g. schematic plans and instructions. In both these scenarios spoken dialogue would significantly increase the ease and effectiveness of the interaction. With advances in speech recognition spoken language interfaces are now becoming possible, but limited understanding of how to design intelligent co-working dialogue agents remains a major obstacle.

There has been substantial prior work on collecting and analysing extended human-human co-working dialogues – we review this in Section 5. However, this work has significant limita-

¹A more common term is *task-oriented dialogue*. We view co-working dialogue as a sub-type of task-oriented dialogue, which includes only genuinely collaborative, task-focussed dialogue, excluding cases that could be deemed task-oriented but which are not genuinely collaborative, e.g. certain types of negotiation and debate, where one participant’s gain is typically the other participant’s loss.

tions with respect to the challenge of fully understanding co-working dialogue. First, the task settings studied are generally artificial and/or very restricted, in particular: (1) they typically involve single tasks, unlike many real world workflow settings where co-workers are involved in multiple, overlapping tasks and must switch between them (Lemon et al., 2002); (2) they are typically static, not dynamic, i.e the world does not change independently of the participants’ actions during the dialogue, requiring an unanticipated shift of focus in the dialogue. Second, the analytical schemes developed to study these dialogues are limited in that they: (1) are typically designed for single task settings and do not distinguish between tasks in complex multi-task settings; (2) do not take into account linking to external task specification or domain ontology resources that are frequently available in complex real world task settings; (3) often focus on generic “dialogue acts”, leaving interpretation of the content of utterances to a task-specific module, hence missing potential generalizations across interaction types that recur in many co-working settings.

In this paper we report our initial efforts to address these issues. First (Section 2) we identify a very substantial, publicly available real world corpus of co-working dialogues – the NASA manned space flight mission dialogues – in a setting where (1) there are multiple tasks to be carried out by multiple actors that may be sequential, concurrent or partially overlapping (2) tasks are co-ordinated in accordance with a high level pre-specified plan, and (3) the task environment is dynamic and only partially known, potentially throwing up unforeseen events or outcomes that may need to be dealt with immediately by unplanned activity and may require task rescheduling. Second (Section 3) we show how, by aligning dialogue transcripts from the corpus with an external task specification or plan, multi-task dialogues can be segmented into interleaved task-related chunks. We illustrate this through a case study in which two annotators separately annotate a 3 hour chunk of co-working dialogue and achieve high accuracy in both segment boundary identification and aligning tasks with the external pre-specified task plan. Third (Section 4) we propose an initial set of dialogue move or interaction types that capture not only the broad communicative function of utterances (e.g. “inform”, “query”, etc.) but also aspects of the semantics of

utterances in co-working dialogues that we claim are generic across co-working settings. We illustrate these interaction types by means of examples taken from the corpus. Our motivating hypothesis here is that a generic co-working dialogue agent can be constructed that can interpret these interaction types in conjunction with external domain- and task-specific knowledge resources, such as ontologies and task or workflow specifications.

Together our proposals for task segmentation and interaction types form the basis of a novel annotation scheme for co-working dialogues. Applied at scale to real world co-working dialogue corpora this scheme can yield both data for training dialogue agents for complex co-working scenarios as well as deeper insights into co-working dialogue itself.

2 The NASA Manned Space Flight Program Data Resources

The US National Aeronautics and Space Administration (NASA), via the Johnson Space Center’s History Portal, has made available audio recordings and transcripts of its entire manned space flight programme in the period 1961-1972, including air-to-ground and onboard conversations for all of the Mercury, Gemini and Apollo missions.² This is an incredible data resource, especially for investigating co-working dialogue, and much understudied in the computational linguistics community. The only prior work on this data by members of the CL community that we are aware of is Mann (2002), who considered a small excerpt from the Apollo 13 mission transcripts to illustrate his dialogue “macrogame” theory, but did not consider the resource more broadly or from the specific perspective of co-working. Clancey (2004) used a portion of the Apollo 17 mission transcripts (~ 1.5 hours) to investigate interactions between the ground-based NASA flight controller (CapCom) and the mission crew that took place during a series of lunar surface activities. He argued that the coordination role of the CapCom provided a model for future disembodied agent assistants working to support humans in similar remote working scenarios on Earth or in space. In particular, he identified various CapCom services that could be automated, such as taking logs, an-

²historycollection.jsc.nasa.gov/JSCHistoryPortal/history/mission_trans/all_transcripts.htm.

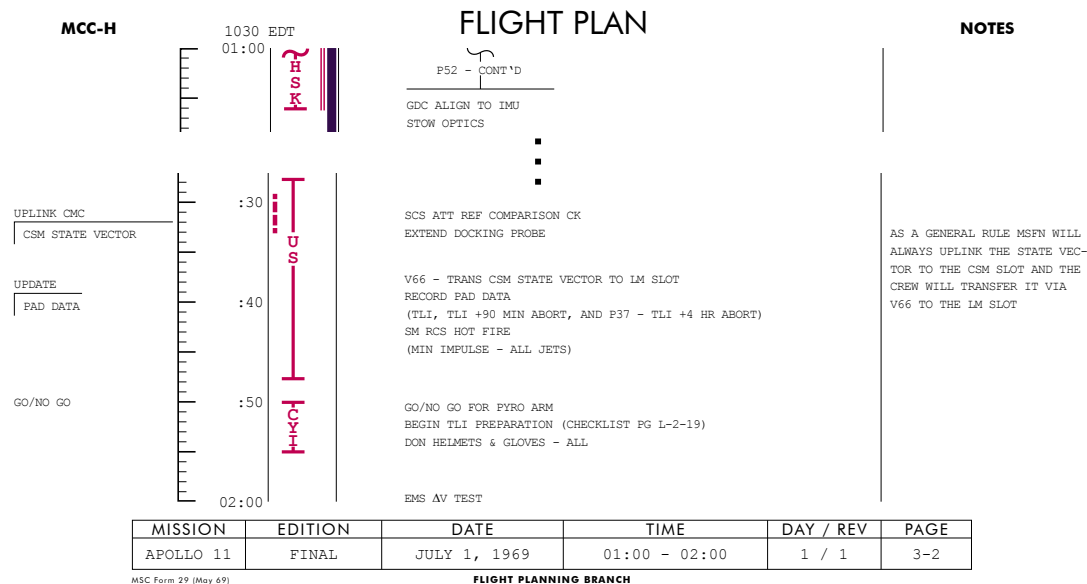


Figure 1: An excerpt from the Apollo 11 Flight Plan. Rule on left shows elapsed time in hours:minutes from launch. Next column to right (in red) shows which ground-based comms station will be handling communications with the spacecraft at that time. Middle column shows activities scheduled to be carried out at the specified time.

swering queries, and managing lunar surface activities and plans. This was a qualitative study, resulting in a fairly coarse grained analysis, and to the best of our knowledge was not developed more fully into a linguistic annotation scheme.

While not optimally curated for CL research, the total volume of dialogue data available far exceeds all other co-working dialogue resources put together and has the further advantage of being spontaneous and from a real task setting. Of course space flight is not a day-to-day experience, but our claim is that the patterns of co-working and co-working dialogue we see here are common to a multitude of other co-working settings.

The dialogues we have chosen to analyse and report on here are from Day 1 of the Apollo 11 mission. They are part of the air-to-ground interaction (onboard interactions are available separately) and have been transcribed and made available in both html and pdf form³ as part of the Apollo 11 Lunar Surface Journal.⁴ To help identify and label task threads (see Section 3) we used the NASA flight plan for the mission⁵ a sample

³hq.nasa.gov/alsj/a11/a11trans.html. Here we have used the html version.

⁴hq.nasa.gov/alsj/a11/a11.html This integrates all of the air-to-ground and onboard interactions, the NASA Public Affairs Officer commentary, including both transcripts and audio, and also includes additional helpful commentary and images. It is invaluable in providing context and background necessary to understand the dialogues.

⁵<https://www.hq.nasa.gov/alsj/a11/>

of which is shown in Figure 1. The flight plan is similar to workflow specifications found in many domains and task settings.

The air-to-ground transcripts provide a good example of remote co-working, where the parties in the dialogue are at different locations, have access to different sources of information and have different capabilities in terms of the task-related actions they can carry out.

We chose to analyse in detail an approximately 3 hour portion of the air-to-ground mission transcript from 00:01:34:33 to 00:04:28:49 (i.e from 0 days, 1 hour, 34 minutes and 33 seconds into the mission until just under 3 hours later). We refer to this corpus as the A11-MCC – Apollo 11 Mini Co-working Corpus. Each utterance in the transcripts is labelled with a time stamp and a speaker id. An example dialogue segment from the A11-MCC is shown in Figure 2.

3 Task Threads

We pursue a two-stage approach to analysing the extended NASA co-working dialogues. The first is to segment the dialogue into task-specific chunks. As is common in many real world task settings, our co-workers engage in multiple tasks in an overlapping fashion (e.g. a process may be set going, then something else done, then the process returned to for monitoring or adjustment). Unsur-

[a11fltpn_final_reformat.pdf](#)

00 01 39 54	CDR	Houston, Apollo 11 is ready to go ahead with the - extend the docking probe, and ready to go with the RCS hot fire when you're ready to monitor. Over.	EDP
00 01 40 06	CC	Roger. Go ahead with the probe, now.	HF
00 01 40 13	CDR	Roger.	EDP
00 01 41 33	CDR	Okay. We're ready to - for the hot fire check when you're ready.	HF
00 01 41 39	CC	Roger. We're ready 11. Go ahead.	HF
00 01 41 48	CDR	Roger. Here's the pitch.	HF
00 01 42 13	CC	Apollo 11, this is Houston. We are seeing the pitch hot firing and it looks good.	HF
00 01 42 18	CDR	Roger. Be advised that we are unable to hear them.	HF
00 01 42 22	CC	Roger. We copy.	HF
00 01 42 24	CDR	Have you seen all three axes fire?	HF
00 01 42 31	CC	We've seen pitch and yaw; we've not seen roll to date.	HF
00 01 42 36	CDR	Okay. I'll put in a couple more rolls.	HF
00 01 42 42	CC	Okay. We've got the roll impulses, and you're looking good here.	HF
00 01 42 48	CDR	Roger. Houston, Apollo 11. We're standing by for a GO for sequence logic ON.	PA
00 01 43 03	CC	Apollo 11, this is Houston. Go ahead and we'll watch you on TM.	PA
00 01 43 07	CDR	Okay. Sequence logic, two of them. Sequence logic 1 and 2 coming up and ON.	PA
00 01 43 36	CC	Apollo 11, this Houston. You are GO for PRYRO ARM.	PA
00 01 43 40	CDR	Roger. Thank you.	PA
	:		
00 01 47 06	CC	... Would you verify that you have extended the probe? Over.	EDP
00 01 47 16	CDR	Roger. That's verified; the probe is extended.	EDP

Figure 2: A Short Sample of the Apollo 11 Air-to-Ground flight transcript. CDR = Commander (Armstrong); CC = Capsule Communicator (Mission Control). Final column is our addition and shows our mapping to activities in flight plan. EDP = Extend Docking Probe; HF = SM RCS Hot Fire; PA = Go/No Go for Pyro Arm

prisingly we find the dialogue pertaining to these overlapping tasks also overlaps. Therefore the task of separating the dialogue into task-specific chunks is not one of simple segmentation but one of identifying task-specific *threads*.

The second stage is to identify recurring *interaction types* in the dialogues we analyse. Here our methodology is one of iteratively analysing the interactions within a mission transcript, hypothesising interaction types with a view to them generalising across other domains, testing the hypotheses against the corpus and refining them to fit. We discuss this process further in Section 4.

3.1 Identifying and Annotating Task Threads

The flight plan (Figure 1) shows a list of tasks the astronauts are meant to be carrying out at each point in the mission. Of course in the event they are not able to stick exactly to schedule; also, some tasks get dynamically rescheduled by ground control. But the flight plan serves as a good guide to what is going on and provides labels for the tasks.

Two annotators (the authors) independently carried out the task segmenting the utterances in the A11-MCC into threads corresponding to a named task in the flight plan. At first glance the sample dialogue in Figure 2 appears to be an undifferentiated stream of mission-related conversation. But on more careful inspection and cross checking with the flight plan, sequences of turns can be

aligned with activities in the flight plan (final column in Figure 2). Note the threaded nature of the task discussions: e.g., first, second and third turns mention the “Extend Docking Probe” task, which is then not mentioned again until the last two turns in the Figure, seven minutes later in the dialogue.

To date we have used an informal annotation scheme to mark up task threads. The key idea is to introduce an abstract “task” element that is realised by one or more “task segments” – sequences of turns where each utterance in the sequence pertains exclusively to a single task. This picture is complicated by the fact that some turns may refer to more than one task (e.g. the first turns in Figure 2). Thus, task-turn relation is many-to-many.

We plan to develop a concrete XML-based syntax consistent with other dialogue annotation formalisms, e.g. Bunt et al. (2012).

3.2 Results and Discussion

Following our double annotation of task threads in the A11-MCC corpus, we discussed divergences on a case-by-case basis and produced a consensus annotation.⁶ Some summary statistics on the consensus data set are presented in Table 1.

As can be seen from the table there were 243 turns across the 3 hour period examined in which

⁶This consensus version is available via the DOI: [10.5281/zenodo.3364099](https://doi.org/10.5281/zenodo.3364099)

Turns	Tasks	Segs Per Task			Turns Per Seg		
		Avg	Min	Max	Avg	Min	Max
243	23	1.52	1	3	5.51	1	26

Table 1: Task Threading in A11-MCC

23 tasks were discussed.⁷ Additionally there was what we called a “COMS” task, which had to do with checking and assuring radio connectivity with various receiving sites on the Earth. Since a COMS-related task is not scheduled in the flight plan but is assumed ongoing across the whole mission⁸, we did not count turns, or parts of turns, relating to COMS as a separate task or in computing segments per task or turns per segment. Such turns comprised 67 of the 243 turns in our corpus.

Of the 23 tasks identified 6 were deemed to be “Unscheduled”, i.e. we could not confidently associate them with any task in the flight plan. Inter-annotator agreement was high, though we do not have precise quantitative agreement figures to report as the annotation exercise was a preliminary investigation of the feasibility of the scheme. There are two distinct tasks that can be assessed: one is determining the boundaries of the task segments and the other is the mapping from task segments to named tasks in the flight plan. Comparing the two annotators to the consensus “gold standard” we found that annotator1 correctly identified the boundaries for 44 out of 44 segments (including the COMS segments), while proposing 2 non-matching segments, for a recall and precision of 100%, while annotator 2 correctly identified boundaries for 42 out of 44 segments, while proposing 5 non-matching segments, for a recall of 95.5% and a precision of 89.4% (for each of the 2 missed segments the annotator proposed finer grained segmentation).

Considering the correct segments only, annotator 1 made 3 labelling errors for a labelling accuracy of 93.2%, while annotator 2 made 5 errors, for a labelling accuracy of 88.1%.

Thus, we are confident that task threads can be identified with high accuracy, especially the boundaries of task segments. Mapping these segments to the flight plan is a somewhat harder task

⁷Note that one turn may discuss more than one task, though in practice no turn ever contributed to more than two task segments. In counting turns per task segment, if a turn contributed to more than one task segment it was counted for each segment to which it contributed.

⁸As noted in the caption to Figure 1, the red vertical bars in the flight plan show through which terrestrial receiving site communications are meant to be passing at any given time in the mission, e.g. CYI = Canary Islands.

as some technical knowledge in the domain is needed to understand, for instance, which particular parts of the spacecraft or particular readings, which may be the subject of conversation, are related to which tasks in the flight plan.

4 Interaction Types

Components of dialogue turns that have a specific task-related function in the interaction we refer to as *interaction elements*. Like some before us (see Section 5) we propose these interaction elements can be grouped into *interaction types*. However, our primary interest is not to categorise interaction elements by broad communicative intent (*inform, query*, etc.) but to type them according to the broad class of task activity to which they relate. Our hypothesis is that a general set of interaction types can be defined that reflect both the types of actions (e.g. assemble, check, configure) that are typically carried out in complex physical co-working contexts, such as manufacturing or space flight, and the meta-actions involved in their realisation (e.g. schedule, co-ordinate, check task status). If such a set of interaction types can be defined, then a generic co-working dialogue agent could be defined that could be readily specialised into a task-specific agent by coupling it with a domain-specific ontology and a task-specific workflow specification.

4.1 Task and Domain Modelling

To describe our proposed set of interaction types we presuppose the existence of a task and domain model, i.e. a model of the world in which the co-workers carry out their actions. We do not here want to articulate in detail such a model or to propose a preferred formal representation language for doing so. However, we do need to identify the principal types of components that domain and task models must contain, as our interaction types will be defined in terms of them. Specifically domain and task models must be able to represent:

1. *Objects, Attributes and Relations:* Objects are things that act and are acted upon in the task domain. They have attributes and stand in relations, which change over time. Agents are one type of object, as are docking probes, O₂ valves, etc. It is useful to be able to distinguish object types and instances and to allow for the hierarchical arrangement of object types within a taxonomy.

Generic Conversational Interaction Types		
Interaction Type	Function	Example
Hail(R,S)	Sender S attempts to attract Receiver R's attention	"Apollo 11, this is Houston."
Acknowledge(R,S)	Receiver R confirms receipt of message to Sender S	"Roger."
Over(R,S)	Sender S informs receiver R that his transmission is complete	"Over."
Co-working Interaction Types		
Execute(G1,G2,Act,T*)	Agent G1 instructs agent G2 to execute activity Act at time T	"Go ahead with the probe, now." "You can start PTC at your convenience"
Configure(G1,G2,<O,A>,V,T*)	Agent G1 instructs Agent G2 to set the attribute A of object O to value V at time T	"We'd like at this time for you to place all four CRYO heaters to AUTO"
CoordinateActivity(G1,G2,Act1,Act2,T*)	Agent G1 requests Agent G2 to carry out activity Act2 at time T so that G1 can carry out Act1	"If you will give us P00 and ACCEPT, we have a state vector update for you." "When you are ready to copy, I have your TLI PAD."
AskPermission(G1,G2,Act,T)	Agent G1 asks Agent G2 for permission to do activity Act at time T	"We'd like to arm our logic switches."
ReportStatus(G1,G2,Act <O,A>,T*)	Agent G1 reports to Agent G2 the status of activity Act or the value of attribute A of object O at time T	"We have the PYRO's armed." "The REPRESS package valve is now in the OFF position".
ReportPlan(G1,G2, Act, T*)	Agent G1 informs Agent G2 that they are going to do activity Act at time T	"And, Buzz, we'll be terminating the battery charge in about a half hour."
QueryStatus(G1,G2,Act <O,A>,T*)	Agent G1 asks Agent G2 to report the status of activity Act or the value of attribute A of object O at time T	"What have you been reading for O2 flow on your onboard gauge?"
CheckStatus(G1,G2,Act <O,A>,RV,T*)	Agent G1 asks Agent G2 to confirm that the status of activity Act or the value of attribute A of object O at time T matches reference value RV	"Would you verify that you have extended the probe? "Would you confirm that your RCS heater switch for quad Bravo is in PRIMARY?"
Ready(G1,G2,Act,T*)	Agent G1 informs agent G2 that G1 is ready to begin activity Act at time T	"I am ready with your TLI-plus-90-minute abort PAD."
VoiceData(G1,G2,D)	Agent G1 reads out a block of data D to agent G2 (typically for G2 to copy down)	"P37 format, TLI plus 5: 00744 6485, minus 165, 02506."
ComparePerspective(G1,G2,<O,A>,V,T)	Agent G1 reports the value V of attribute A for object O at time T and invites Agent G2 to report the value he perceives	"And, Houston, looked like we saw about 87 or 88 psi on chamber pressure that time. I'd like you to look at that on the ground."

Table 2: Basic Interaction Types. *'ed arguments are optional with a default assumed if absent.

2. *Actions (or Activities)* In classical planning models (Fikes and Nilsson, 1971; Ghallab et al., 2016), actions have associated *preconditions* and *effects* and are specified in terms of the change they effect in the world, given that world is in a certain state when the action is performed and that state meets the action's preconditions. Actions may either be primitive or may specify a set of *sub-actions*, which must be performed for the higher level action to be accomplished. This recursive structure of actions is something we need for our account of co-working dialogues. As with objects, we need to type actions and distinguish action types from instances.

3. *Goals* Goals are distinguished states to be achieved or actions to be completed.
4. *Plans* Plans are sequences of actions, or partially ordered set of actions, which lead to a goal state or the completion of a goal action.
5. *Time* We require a model of actions, plans and goals in which time and temporal relations figure explicitly, since in many co-working situations scheduling of activities both relative to clock time and to each other is an essential part of what gets discussed.

As noted above, for current purposes we do not need to chose a particular formalism for representing task and domain models. There are,

however, several to choose from. These have emerged from the automatic planning community, which needs models of the world and of the tasks to be performed as input to the planning process (Fikes and Nilsson, 1971; Fox and Long, 2003; Gil, 2005; Ghallab et al., 2016)) and from the community focussed on exchange formats or standards for describing plans and activities in various real world domains, such as NIST’s Process Specification Language (PSL) for manufacturing (Grüninger and Menzel, 2003) and the Shared Planning and Activity Representation (SPAR), sponsored by DARPA and the US Air Force for military planning⁹.

4.2 Identifying and Annotating Interaction Types

Table 2 summarises our proposed set of co-working interaction types. It is divided into a list of generic conversational interaction types and a list of co-working interaction types. In the first column we give the label and associated argument structure for that particular interaction type. In the second column, we explain this notation. The final column shows a dialogue segment which would be classified by this interaction type. So, for example, in the first row of the Co-working interaction types, we find “Instruct(G1,G2,Act,T)”, a label that is applied to dialogue which communicates an instruction from Agent G1 to Agent G2, for G2 to do activity Act at time T. The example text includes “Go ahead with the probe now”; G1 is the CC; G2 is the CDR; Act is “the probe”; T is “now”; (note that the local dialogue context, as shown in Figure 1, reveals that the action referred to is “extend the docking probe”). The list of types we show in table 2 is not exhaustive. The co-working types were identified to accommodate the majority of dialogue in the A11-MCC (excluding the COMS segments); nonetheless our preliminary qualitative analysis of the entire Apollo 11 transcript suggests that these co-working types are applicable throughout the mission dialogue. In future work we plan to annotate a larger sample of dialogue taken from across the 8 day mission, and to extend this list where the data suggests there is a requirement for further co-working types. However, we believe the original list in Table 2 will form the majority of an extended co-working type set. Moreover, our belief is that these types will be

applicable in other domains, such as automotive maintenance and cooking. For example, a Configure(G1, G2, <O,A>, V,T) could apply to a request from a cook to an assistant to “now set the oven to 200” or a CheckStatus(G1,G2,Act,<O,A>,RV,T) could describe a request from a mechanic for an apprentice to check that the clearance for a piston intake valve is within the range of 0.18-0.22mm.

Table 2 also lists a few examples of generic interaction types, e.g. “Hail”, “Acknowledge”. In future work we plan to extend this list, drawing from the extensive list of communicative functional types in the ISO 24617-2 dialogue act annotation standard in order to annotate more general features of task oriented dialogue such as communication management, feedback, turn taking, etc.

5 Related Work

In this section we review prior work on co-working dialogue corpora and on analytical frameworks for describing them. This review is not exhaustive but highlights key related work.

5.1 Previous Co-working Dialogue Corpora

An extensive review of dialogue corpora can be found at Serban et al. (2015). Here we focus solely on corpora of co-working dialogues.

Several dialogue corpora have been built with a view to studying dialogue in co-working settings. These include: the Map Task Corpus (Anderson et al., 1991), in which pairs of participants collaborate via spoken dialogue to reproduce a route drawn on one map on another map; the TRAINS project co-working dialogue corpora, human-human conversations about managing the shipment of goods and movement of trains around a rail network (Allen et al., 1995; Heeman and Allen, 1995); and the AMI corpus of dialogues arising from, primarily, design team meetings (Carletta, 2007). In all these cases the corpora possess one or more of the shortcomings noted in Section 1: the task is artificial; the setting is static; in the case of AMI, dialogue arising from collective deliberation in meetings is very different from the sort of co-working dialogue that is the focus of work here, i.e. dialogue in settings where agents strive in real time to bring about a state of affairs in the physical world.

5.2 Analytical Frameworks for Dialogue

Task-based Dialogue Segmentation Grosz and Sidner (1986) propose segmenting dialogues ac-

⁹See www.aiai.ed.ac.uk/project/spar/.

cording to their *intentional* structure. In the examples they give, segments are recognised and labelled with intentions by human analysts. By contrast, in our case segments are determined by reference to an external task specification or plan. However, by adopting a plan an agent can be seen as forming an intention to execute each of the steps in the plan. Isard and Carletta (1995) segmented the Map Task dialogues into transactions by identifying sequences of dialogue that corresponded to the communication of a particular section of the route (i.e. a sub-task of the high level map task). While this work is similar to our approach, the resulting dialogue segments correspond to a single, artificial task type; we address multiple tasks as specified in a real world plan. Finally, in the AMI corpus dialogues are segmented by topic¹⁰, using a set of domain-specific topics pre-specified by the corpus designers. This contrasts strongly with our task-based segmentation, where the tasks underlying the segmentation are provided in a plan devised not by corpus designers but originating in the real world context in which the dialogues occur.

Games and Moves Starting with Power (1979), there is a tradition of analysing dialogues in terms of *games* and *moves* (Kowtko et al., 1997; Lewin, 2000; Mann, 2002). Kowtko et al. (1997), for example, present a framework for analysing task-oriented dialogues which involves a two-level analysis in terms of *conversational games*, sequences of turns required to achieve a conversational sub-goal, and, at a lower level, *moves*, which are single, multiple or partial utterances that convey a single, specific intent. In annotation of the Map Task dialogues, they used six “initiating moves”: *Instruct*, *Check*, *Align*, *Query y-n*, *Query-w*, *Explain* and six “response and feedback” moves: *Clarify*, *Acknowledge*, *Ready*, *Reply-Y*, *Reply-N*, *Reply-W*. There are some similarities between this work and our own, e.g. their *Instruct* move – “a direct or indirect request or instruction, to be done immediately or shortly” – is similar to our *Execute* interaction type. However, our moves are more grounded in the task – the arguments in our *Execute* serve to link to an external task model; our *Ready* interaction type is about communicating readiness to start a task while their *Ready* move is about conversants signalling readiness to take part in a conversation or game; their

Check is to check a participants understanding of the communication, our *Checkstatus* is about checking that something in the external world is as it should be.

Dialogue Acts Much previous work has focussed on defining a set or hierarchy of dialogue acts, which are like moves as discussed above, but express a more finely nuanced descriptive framework for characterising different functional aspects of elements of dialogue (see, e.g., the Damsel dialogue act markup scheme (Allen and Core, 1997), the Switchboard dialogue act tagset (Stolcke et al., 2000) and the ISO 24617-2 dialogue annotation standard (Bunt et al., 2010, 2012)).

This work, particularly the ISO 24617-2 standard, proposes a rich, multi-dimensional approach to functional segment classification in dialogue. The co-working interaction types we propose fall within the task dimension in the ISO standard. However our types provide a more detailed view of the communicative function of dialogue units, capturing the task semantics in a way that would allow an agent to interpret them in relation to an externally supplied model of the task and domain.

6 Conclusion and Future Work

In this paper we have taken initial steps towards defining a novel two level framework for analysing and annotating co-working dialogues. Key aspects of the framework are (1) the identification and annotation of task-specific threads within extended real world dialogues, which can be linked to external task specifications, and (2) the definition of a set of “interaction types”, which recur across co-working dialogues and serve to identify both the communicative function of the linguistic unit in the co-working context and the elements within it which refer to objects, entities and activities in the task world. We illustrated both levels of the framework by reference to dialogues in the Apollo 11 air-to-ground mission transcripts, an invaluable source of real world co-working dialogues.

Going forward we intend first to validate the generality of our framework by applying it to a co-working corpus in another domain. We then plan to manually annotate a sufficient quantity of dialogue to train automatic annotators. Starting by modelling basic maintenance, repair and overhaul tasks in limited domains, we also intend to implement a co-working dialogue agent based on the framework put forward in this paper.

¹⁰<http://groups.inf.ed.ac.uk/ami/corpus/annotation.shtml>

References

- James Allen and Mark Core. 1997. [Draft of damsl: Dialog act markup in several layers contents](#).
- James F. Allen, Lenhart Schubert, George Ferguson, Peter Heeman, Chung Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, and David Traum. 1995. [The trains project: a case study in building a conversational planning agent](#). *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.
- Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. [The hcrc map task corpus](#). *Language and Speech*, 34(4):351–366.
- Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. [Towards an ISO standard for dialogue act annotation](#). In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).
- Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Andrei Petukhova, Volha an Popescu-Belis, and David 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)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).
- Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.
- W. J. Clancey. 2004. [Roles for agent assistants in field science: Understanding personal projects and collaboration](#). *Trans. Sys. Man Cyber Part C*, 34(2):125–137.
- Richard E. Fikes and Nils J. Nilsson. 1971. [Strips: A new approach to the application of theorem proving to problem solving](#). *Artificial Intelligence*, 2(3):189 – 208.
- Maria Fox and Derek Long. 2003. Pddl2.1: An extension to pddl for expressing temporal planning domains. *J. Artif. Int. Res.*, 20(1):61–124.
- Malik Ghallab, Dana Nau, and Paolo Traverso. 2016. [Automated Planning and Acting](#). Cambridge University Press.
- Yolanda Gil. 2005. [Description logics and planning](#). *AI Magazine*, 26(2):73–84.
- B. J. Grosz and C. L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Michael Grüninger and Christopher Menzel. 2003. The process specification language (psl) theory and applications. *AI Mag.*, 24(3):63–74.
- Peter A. Heeman and James Allen. 1995. [The trains 93 dialogues](#). Technical Report 94-2, Computer Science Dept., University of Rochester.
- M. Hermann, T. Pentek, and B. Otto. 2016. [Design principles for industrie 4.0 scenarios](#). In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 3928–3937.
- Amy Isard and Jean Carletta. 1995. Replicability of transaction and action coding in the map task corpus. In *In AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, pages 60–66.
- Jacqueline C. Kowtko, Stephen D. Isard, and Gwyneth M. Doherty. 1997. Conversational games within dialogue. Technical Report HCRC/RP-31, University of Edinburgh.
- Oliver Lemon, Alexander Gruenstein, Alexis Battle, and Stanley Peters. 2002. [Multi-tasking and collaborative activities in dialogue systems](#). In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue - Volume 2*, SIGDIAL '02, pages 113–124, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ian Lewin. 2000. A formal model of conversational game theory. In *Proc. Gotalog-00, 4th Workshop on the Semantics and Pragmatics of Dialogue*.
- William C. Mann. 2002. [Dialogue macrogame theory](#). In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue - Volume 2*, SIGDIAL '02, pages 129–141, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Richard Power. 1979. [The organisation of purposeful dialogues](#). *Linguistics*, 17:107–152.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. [A survey of available corpora for building data-driven dialogue systems](#). *CoRR*, abs/1512.05742.
- Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol VanEss-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *Comput. Linguist.*, 26(3):339–373.

Good call!

Grounding in a Directory Enquiries Corpus

Christine Howes Anastasia Bondarenko Staffan Larsson

Centre for Linguistic Theory and Studies in Probability (CLASP)

Department of Philosophy, Linguistics and Theory of Science

University of Gothenburg, Sweden

christine.howes@gu.se

Abstract

This paper describes the collection of a corpus of telephone directory enquiries conversations. We analyse the feedback used in the corpus and discuss implications for dialogue systems.

1 Introduction

Effective communication requires collaboration between all participants, with dialogue co-constructed by speakers and hearers. Even in contexts such as lectures or storytelling, which are largely monological (Rühlemann, 2007), listeners provide frequent feedback. This feedback demonstrates whether or not they have *grounded* the conversation thus far (Clark, 1996), i.e. whether something said can be taken to be understood, and comes in the form of relevant next turns, or backchannels (e.g. ‘yes’, ‘yeah’, Example 1; lines 6 and 8¹ or ‘mm’).² Other responses, such as clarification requests (e.g. Example 1; lines 10 and 17) indicate processing difficulties or lack of coordination and signal a need for repair (Purver, 2004; Bavelas et al., 2012).

These communicative grounding strategies (Clark and Brennan, 1991; Traum, 1994) enable dialogue participants to manage the characteristic divergence and convergence that is key to moving dialogue forward (Clark and Schaefer, 1987, 1989), and are therefore crucial for dialogue agents. Importantly, feedback is known to occur subsententially (Howes and Eshghi, 2017), but most dialogue models do not operate in an incremental fashion that would allow them to produce or interpret feedback in a timely fashion.

¹Examples are all taken from our Directory Enquiries Corpus (DEC), described below.

²In face-to-face dialogue this includes non-linguistic cues (e.g. nods), but as our corpus is telephone conversations, we do not consider these here.

(1) DEC07:1–32

1	Caller	hello
2	Operator	hello
3	Caller	hello
4	Operator	how may i help you?
5	Caller	oh hi i'm uh looking for some phone numbers
6	Operator	yes
7	Caller	er here in london
8	Operator	yeah
9	Caller	and the first
10		one is rowans tenpin bowl
11	Operator	can you repeat that for me?
12	Caller	rowans tenpin bowl
13		so it's rowan
14		R O W A N S
15	Operator	yes
16	Caller	tenpin
17	Operator	tenpin?
18	Caller	yeah
19	Operator	the number ten
20	Operator	and pin?
21	Caller	yes
22	Caller	yes
23	Operator	tenpin
24	Operator	road?
25	Caller	bowl
26	Operator	th- like the bird?
27	Caller	uh like bowling
28	Operator	uh bowling
29	Caller	bowl
30	Operator	yes
31		the thing you eat from right?
32		okay here we go

While it is difficult to compare corpus studies of feedback, as terms such as backchannels and repair have not been used consistently in the literature (see Fujimoto, 2007, for review), there are a number of quantificational studies of feedback that bear mentioning. One of the earliest is that described in Duncan (1972, 1974), which presents a detailed multimodal annotation of backchannel responses, and finds that in 885 ‘units’ (roughly corresponding to utterances) there are a total of 71 instances of feedback (8%). Corpus studies that cover aspects of feedback include (Fernández, 2006), whose annotations of non-sentential utterances (NSUs) in a subcorpus of the British

National Corpus (BNC; [Burnard, 2000](#)) include the classes ‘acknowledgements’ (5% of all utterances), and ‘clarification ellipsis’ (1%). However, as her focus is on NSUs, [Fernández \(2006\)](#) deliberately excludes cases in overlap, which means many genuine feedback utterances will be missed ([Rühlemann, 2007](#)). For clarification requests, the numbers reported in ([Fernández, 2006](#)) are also an underestimate, as she is not concerned with sentential cases (e.g. “what do you mean?”). In another BNC study, [Purver \(2004\)](#) found that CRs made up just under 3% of utterances, whilst [Colman and Healey \(2011\)](#) found different levels of CRs in different dialogue domains, with more in the task oriented Map Task ([Anderson et al., 1991](#)). Interestingly, this varied significantly depending on role; route followers produced significantly more CRs than route givers. Additionally, and importantly for phone conversations, participants in the Map Task also produce more backchannels when they are not visible to one another ([Boyle et al., 1994](#)).

Although using low-level features ([Cathcart et al., 2003](#); [Gravano and Hirschberg, 2009](#)) may allow a dialogue model to sound ‘more human’, it can’t provide any insight into why feedback occurs where it does, or whether there are different appropriate responses to feedback dependent on its positioning and other characteristics. It is also unclear whether models in which feedback incorporates reasoning about the intentions or goals of one’s interlocutor ([Visser et al., 2014](#); [Buschmeier and Kopp, 2013](#); [Wang et al., 2011](#)) presuppose a level of complexity that is unnecessary in natural conversation ([Gregoromichelaki et al., 2011](#)).

Here, we focus on feedback in an extremely restricted domain – that of telephone directory enquiries (see also [Clark and Schaefer, 1987](#); [Bangerter et al., 2004](#)), which can be seen as a good test case for dialogue systems. Directory enquiries is a real world application for dialogue systems (e.g. [Chang, 2007](#)) that has particular features that can be problematic for a speech recogniser, such as understanding names which are not present in an existing lexicon over a noisy channel. As we argue below, this is a particularly good domain for studying feedback, as feedback should be more frequent and necessary than in less restricted domains. The reasons for this are two-fold. Firstly, in task-oriented dialogue, where information transfer is crucial for success, and

avoiding miscommunication is vital, feedback is more common than in less goal-directed conversations ([Colman and Healey, 2011](#)). Secondly, verbal feedback is more frequent in dialogues where participants cannot see each other, and therefore do not have the ability to employ non-verbal feedback ([Boyle et al., 1994](#)), such as telephone conversations. In addition, the specific task of a directory enquiries call is less asymmetric than many tasks used to study dialogue, such as the Map Task ([Anderson et al., 1991](#)), because both participants act as ‘information giver’ (caller for name to be looked up; operator for phone number) and ‘information receiver’ (the reverse) at different stages in the dialogue. Additionally, in contrast to corpora which have similar features (such as SRI’s Amex Travel Agent Data, [Kowtko and Price, 1989](#)), relevant parts of the dialogue (names and numbers, see below) do not require anonymisation.

In this paper, we present a new corpus of human-human telephone directory enquiries dialogues, and explore the strategies for feedback that human participants use, especially in cases where misunderstandings arise. We suggest that dialogue models need to be able to perform incremental grounding, particularly in the context of spelling out words and dictating number sequences, with a number of increasingly specific strategies available for both acknowledgements and clarifications. The complete corpus (transcriptions, audio and annotations) is freely available on the Open Science Framework (osf.io/2vjkh; [Bondarenko et al., 2019](#)) thus aiding in the development of spoken dialogue systems that need to both acquire and offer accurate information to the user (e.g. directory enquiries, travel agents etc).

2 Method

2.1 Data collection

The data was collected with the help of 14 volunteers who were paired up for each recording session. Eight of the volunteers were male and six were female. The participants were native speakers of a number of different languages and had various levels of English proficiency.

Each pair of participants was instructed that they were to take turns playing the roles of a directory service enquiries caller and operator. Each caller was provided with a list of three businesses located in London, and told that their task was to find out the phone numbers of the businesses on

their list through a telephone conversation with the operator. The operators task in turn was to provide the caller with the phone numbers using the on-line Phone Book service (thephonebook.bt.com). Each caller made two calls to the operator who was situated in the studio. The recording sessions resulted in 4 dialogues per pair (28 in total) with the shortest dialogue duration being 2 minutes 31 seconds and the longest one being 10 minutes 46 seconds.

2.2 Transcription

The audio recordings were transcribed using ELAN (Brugman and Russel, 2004).

2.3 Annotation

All of the transcripts were manually annotated, with the overview of annotations used shown in Table 1. Two dialogues (281 utterances) were annotated by two coders to ensure inter-rater reliability. Cohen’s kappa tests showed good agreement for all tags: *turn-type* (*ack/CR/C*) $\kappa = 0.635$; *AckType* $\kappa = 0.625$; *CRType* $\kappa = 0.689$.

2.4 Feedback subtypes annotation

Following observations of the data, we further annotated our feedback utterances into subtype. For acknowledgements these are:

- Continuer** acknowledgement/backchannel words like “okay”, “yeah”, “yes”, “mmhm” (e.g. Example 1; line 8).
- Verbatim** verbatim repetitions of (parts of) previous utterances (e.g. Example 1; line 27)
- Paraphrase** paraphrased repetitions of (parts of) previous utterances
- Confirm** confirmation phrases like “correct”, “exactly”, “thats correct”
- Appreciate** appreciative response to the previous utterance: “great”, “good”, “perfect”.

For clarification requests these are:³

- General request** indicates a non-specific lack of perception/understanding of other speaker’s previous utterance (e.g. “sorry?”, “what?”)

³As pointed out by an anonymous reviewer, the categories for acknowledgements may conflate form and function, whilst those for CRs do not consider the form. This may mean that we miss important parallels or differences between acknowledgements and clarification requests and we intend to address this in future work.

- Repeat request** asks other speaker to repeat a previous utterance (e.g. Example 1; line 11)

- Confirmation request** asks other speaker to provide a confirmation (e.g. Example 1; line 17)

- Spelling request** asks other speaker to spell out the name of the queried business or its address (e.g. “could you spell that for me please?”, “is that a W?”)

2.5 Content annotation

Since the main purpose of the data collection was to investigate the domain of telephone directory enquiries each of the the utterances was also labelled according to its content: namely, whether it includes any information about the names, addresses and phone numbers of businesses. Each utterance labelled with any of these was then labelled according to the form such information was conveyed in:

- Word (part)** speaker mentions the name of a business or its address in full or in part
- Spelling installment (part)** speaker provides a spelling for the name or the address of a business in full or in part, usually in installments of one or more letters
- Dictation installment (part)** speaker dictates a phone number in full or in part, usually in installments of one or more digits
- PreviousWord/spelling/dictation, PreviousContent** each utterance is also annotated with the content and form labels of the previous utterance.

3 Results

In our 28 dialogues, there were a total of 4165 utterances, or 3002 speaker turns (for our purposes a turn constitutes multiple consecutive utterances by the same speaker with no intervening material from the other participant). The shortest dialogue consists of 64 utterances (48 turns) and the longest consists of 246 utterances (190 turns). 1285 of these utterances are acknowledgements, which constitutes 31% of utterances or 43% of turns. There are also 277 clarification requests, i.e. 7% of utterances and 9% of turns.⁴ This is higher than found in previous studies (Purver,

⁴As the pattern of results is consistent over turns or utterances, for the remainder of this paper we focus on the by utterance numbers.

Tag	Value	Explanation
acknowledge (Ack)	y/n	For all utterances: does this sentence contain a backchannel (e.g. ‘yeah’, ‘mhm’, ‘right’) or a repeated word or phrase acknowledging the proposition or speech act of a previous utterance? (Note this category does not include direct answers to yes/no questions)
clarification request (CR)	y/n	For all utterances: does this utterance contain a clarification request, indicating misunderstanding of the proposition or speech act of a previous utterance
clarify (C)	y/n	For utterances following a clarification request: does this utterance contain a response to a clarification request, clarifying the proposition or speech act of a previous utterance?

Table 1: Annotation Tags

2004; Fernández, 2006; Boyle et al., 1994, a.o.), and, as discussed in the introduction, is probably due to the nature of the task.

As shown in Table 2, operators produce more acknowledgements and clarification requests than callers (Acks: 36% vs 26% $\chi^2_1 = 48.466, p < 0.001$; CRs: 9% vs 4% $\chi^2_1 = 36.961, p < 0.001$). This result stems from the greater possibility for error in the understanding of names compared to numbers (see section 3.1 below).

	Role					
	Caller		Operator		Total	
Ack	559	26%	726	36%	1285	31%
C	189	9%	64	3%	253	6%
CR	94	4%	183	9%	277	7%
(blank)	1306	61%	1044	52%	2350	56%
Total	2148	100%	2017	100%	4165	100%

Table 2: Summary of results by speaker role

3.1 Asymmetry of information

As shown in Tables 3 and 4, as in Colman and Healey (2011), the pattern of feedback mirrors the asymmetry of roles, with information receiver (i.e. operator for the business name, and the caller for the phone number) providing the majority of acknowledgements and clarification requests.

	Role					
	Caller		Operator		Total	
Ack	50	11%	441	68%	491	44%
C	78	16%	1	0%	79	7%
CR	3	1%	100	15%	103	9%
(blank)	342	72%	105	16%	447	40%
Total	473	100%	647	100%	1120	100%

Table 3: Results by speaker role where the previous utterance is about a business name

	Role					
	Caller		Operator		Total	
Ack	364	73%	92	28%	456	55%
C	0	0%	30	9%	30	4%
CR	60	12%	0	0%	60	7%
(blank)	75	15%	210	63%	285	34%
Total	499	100%	332	100%	831	100%

Table 4: Results by speaker role where the previous utterance is about a business phone number

3.2 Feedback subtypes

As shown in Table 5, most of the acknowledgements in our corpus consist of continuers, with 772 (60%) acknowledgements containing at least one continuer. The next most common type of acknowledgement is a verbatim repeat of material from a prior utterance, with 492 (38%) acknowledgements. For a dialogue system, this is good news: simple utterances of just a continuer or repeated material accounts for 91% of all acknowledgements, suggesting that these may be the only two strategies that need to be implemented for both production and comprehension.

For clarification requests (Table 6), the majority (48%) are confirmation requests – checking that something has been understood by offering a provisional interpretation. These serve to pinpoint the (potential) source of miscommunication in a way that the more general types do not (see also Ginzburg, 2012). In practice, they are very similar to the verbatim acknowledgements, as in example 1 line 17, but with questioning intonation suggesting that they are more tentative. These ought to therefore be generatable in the same way as verbatim acknowledgements. The data suggest a scale of feedback, analogous to Clark and colleagues’ levels of evidence of understanding

(Clark and Brennan, 1991; Clark and Schaefer, 1989; Clark, 1996), with listener confidence being a key component of which type of feedback is appropriate.

Type(s)	Number	%
Appreciate	5	0.4%
Confirm	21	1.6%
Confirm, Continuer	1	0.1%
Continuer	718	55.9%
Continuer, Appreciate	9	0.7%
Continuer, Appreciate, Continuer	1	0.1%
Continuer, Confirm	9	0.7%
Continuer, Paraphrase	2	0.2%
Continuer, Verbatim	3	0.2%
Paraphrase	25	1.9%
Paraphrase, Continuer	2	0.2%
Verbatim	456	35.5%
Verbatim, Appreciate	1	0.1%
Verbatim, Continuer	25	1.9%
Verbatim, Continuer, Appreciate	2	0.2%
Verbatim, Paraphrase	1	0.1%
Verbatim, Verbatim	4	0.3%
Total	1285	100%

Table 5: Types of acknowledgement

Type	Number	%
Confirmation request	134	48.4%
General request	28	10.1%
Repeat request	64	23.1%
Spelling request	51	18.4%
Total	277	100%

Table 6: Types of clarification request

3.3 Strategies

As there is greater scope for miscommunication in the transmission of names than numbers, we now focus on the examples where the feedback follows an utterance whose content is about a name.⁵ For these cases, there is large variability in how easily the names are conveyed, with the number of turns taken from the first mention of any part of the name to the operator confirming that they have found the number ranging from 2 utterances to 82 utterances, with 3 (of 84) cases unresolved.

Table 7 shows that of the turns following an utterance about a business name, 45% contain a spelling installment, or part of one, with similar proportions for acknowledgements (36%) and clarification requests (41%), with only 15% (acks 12%, CRs 21%) relating to the word level. This

⁵Note that row totals in Tables 7, 8 and 9 do not add up to 100% as some turns contain more than one strategy.

shows that models of dialogue need to be able to produce and interpret increments of different sizes – potentially of a single letter, as people do when they are pinpointing sources of (potential) trouble within an unfamiliar name.

Tables 8 and 9 demonstrate that feedback strategies are highly dependent on the information giving strategy employed in the preceding utterance. While generic strategies (continuers or non-specific repairs such as “what?”) are common and always available, participants are also likely to match the prior strategy used in their feedback – it is, for example, rare to acknowledge or clarify a spelling installment with a word, and vice versa.

3.4 Qualitative results

Examples 2–9 show a variety of these strategies in action. In Example 2, the Operator relies on continuer acknowledgements, which, according to Clark and colleagues’ model of levels of evidence of understanding, are weaker signals of understanding than e.g. verbatim repeats and might be therefore more likely to allow misunderstandings to occur. Example 3 from another pair shows the same business name split into different increments (with the first half of the name “bistro” treated as an independent word and the rest spelled out in increments of 3 letters; see also section 3.5, below), with different feedback techniques for different subparts of the utterance – a continuer at line 126, a verbatim acknowledgement at line 128.

(2) DEC11:88–98

88	Operator	er can you spell bistrotheque for me?
89	Caller	abs-
90	Caller	sure er it’s
91	Caller	B I S
92	Operator	yes
93	Caller	T R O
94	Operator	mmhm
95	Caller	T H E
96	Operator	okay
97	Caller	Q U E
98	Operator	er yes i have it here for you

(3) DEC3:123–128

123	Caller	so bistro
124	Caller	T
125	Caller	H E
126	Operator	yeah
127	Caller	Q U E
128	Operator	Q U E

Example 4 splits the business name into two increments of 3 and 4 letters respectively, and is acknowledged by verbatim repeats in each case.

	Ack		CR		Total	
Spelling installment	137	28%	31	30%	394	35%
Spelling installment part	41	8%	11	11%	107	10%
Word	21	4%	5	5%	47	4%
Word part	40	8%	16	16%	127	11%
Other	253	52%	42	41%	452	40%
Total	491	100%	103	100%	1120	100%

Table 7: Strategies for feedback following an utterance about a business name

	Previous utterance content type								Total
	Spelling installment		Spelling instmt part		Word		Word part		
Spelling installment	127	40%	9	20%	0	0%	1	1%	137
Spelling installment part	23	7%	18	39%	0	0%	4	6%	41
Word	3	1%	2	4%	10	20%	6	9%	21
Word part	3	1%	0	0%	15	30%	22	32%	40
(continuer/confirm/appreciate)	171	54%	18	39%	25	50%	42	62%	253
Total	319	100%	46	100%	50	100%	68	100%	491

Table 8: Strategies for acknowledgements about a business name by previous utterance content type

A common strategy for avoiding miscommunications in spellings is developed in Example 5: namely using unambiguous words which start with the same letter. This strategy is prompted by the operator’s clarification request in line 19. Note that the acknowledgements provided by the operator here are sometimes only the word (e.g. line 23 “america”) but sometimes include the letter in a direct repeat of the whole utterance (e.g. line 35 “R for Russia”). In our corpus, different pairs come up with different sets of words for spelling out the letters (e.g. country/city names, as here, or people’s first names – note that this choice can also be the source of miscommunication, as in Example 12). This strategy can be initiated by either participant, or in co-constructions (as in Example 7), and, after repeated interactions, participants may use this strategy productively – even dropping the letter with the country name standing in for the whole, as in Example 6 (this mirrors the way participants strategically align in tasks such as the Maze Game; Mills and Healey, 2006).

(4) DEC16:54–61

54 Caller the next place i’m looking for is called
55 Caller er tayyabs which is spelled
56 Caller T A Y
57 Operator T A Y
58 Caller Y A B S
59 Operator Y A B S
60 Caller it’s a restaurant
61 Operator okay

(5) DEC28:17–35

17 Caller okay so it starts with a
18 Caller L
19 Operator L?
20 Caller as in london
21 Operator yes
22 Caller A as in america
23 Operator america
24 Caller er U
25 Caller as in er
26 Caller er under
27 Caller <laugh>
28 Operator under yes
29 Caller er D as in denmark
30 Operator denmark
31 Caller E as in england
32 Operator england
33 Caller and R
34 Caller for russia
35 Operator R for russia

(6) DEC26:61–69

61 Caller it’s it’s a restaurant by name tayyabs
62 Operator okay can you spell that for me please?
63 Caller should i
64 Caller yes it’s a thailand
65 Operator yes
66 Caller america
67 Operator yes
68 Caller yugoslavia
69 Operator yes
: : :

(7) DEC28:138–141 Co-construction

138 Caller and K for er
139 Caller <laugh>
140 Operator as in king?
141 Caller k- king <laugh> yeah

3.5 Increments

People often break the names into increments to aid understanding, but what counts as an incre-

	Previous utterance content type								Total
	Spelling installment		Spelling instmt part		Word		Word part		
Spelling installment	24	52%	3	43%	1	4%	4	19%	31
Spelling installment part	8	17%	3	43%		0%		0%	11
Word		0%		0%	4	16%		0%	5
Word part	2	4%	1	14%	5	20%	10	48%	16
(generic repair)	17	37%		0%	16	64%	12	57%	42
Total	46	100%	7	100%	25	100%	21	100%	103

Table 9: Strategies for clarification requests about a business name by previous utterance content type

ment is not fixed, and may be further subdivided in case of failure. Examples 8 and 9 show two different ways in which the same name was divided into increments, with Example 9 having many more utterances, including several verbatim acknowledgements to convey the same information.

(8) DEC7:89–98

89 Caller phoenicia mediterranean food
90 Operator can you repeat that for me?
91 Operator tenicia?
92 Caller yeah
93 Caller it's P H
94 Caller O E N
95 Operator mmhm
96 Operator co- continue please
97 Caller I C I A
98 Operator I C I A

(9) DEC23:101–117

101 Caller yeah it's phoenicia
102 Operator clomissia?
103 Caller mediterranean food
104 Caller yes you spell it with a P
105 Operator P
106 Caller H
107 Caller O
108 Operator H O
109 Caller E
110 Operator yes P H O E
111 Caller E N
112 Operator N
113 Caller A C
114 Operator A C
115 Caller A-
116 Caller I A
117 Operator I A

3.6 Repair Strategies

In our data there is some indication that participants are generally good at predicting potentially problematic elements and further specifying those before they lead to miscommunication, such as non-conventional spellings of words as in Examples 10 and 11.

(10) DEC20:4–9

4 Caller the first one being first one being one
called cittie of yorke which is C I T T
I E of
5 Caller yorke spelled with an E at the end
6 Operator cittie of yorke with two Ts?
7 Caller cittie of yorke where cittie isn't
8 Caller C I T Y it's C I T T I E
9 Operator yeah

(11) DEC10:59–9

59 Caller it's called lyle's
60 Caller with a Y
61 Operator lyle's

In general, misunderstandings are resolved quickly and locally, however, there are also interesting cases where misunderstandings persist, such as Example 12, with the specific problematic letter in the name taking 57 utterances to resolve. In this case, as in 13, the participants started by trying to just spell out the names (which can be ambiguous, especially in noisy settings) and then switch strategy to a more specific method (here using the initial letter of a name or place) when the initial strategy fails.

(12) DEC22:82–139

82 Caller with a - filip with an F
83 Operator filip
84 Operator yeah
: :
107 Caller er
108 Operator pilip
109 Caller fanny
110 Operator mmhm
111 Caller fanny
: :
113 Operator P
114 Operator P as in panda
115 Operator right?
116 Caller sorry i didn't hear you
117 Operator P
118 Operator the next one is a P
119 Operator as in panda
120 Caller P?

121 Operator or okay
 122 Operator then
 123 Caller no
 124 Caller it's er
 : :
 133 Caller uh fanny
 134 Operator <unclear> I don't know that name
 funny?
 135 Caller yeah or like filip but with an F
 136 Caller or if you say fruits
 137 Operator with an F?
 138 Operator okay
 139 Caller F yeah

(13) DEC25:67–112 Change of strategy

67 Caller yes and the business i was looking
 for hot- it's a hotel it's called hotel
 wardonia
 : : <lines 68–94 spell out the name >
 95 Operator er i'm sorry i couldn't find any re-
 sult for
 96 Operator hotel swarbonia maybe i spelled
 97 Operator wrong
 98 Caller yes i can spell that once again
 99 Operator yes please
 100 Caller it's er W for wales
 101 Operator er so it's hotel first?
 102 Caller yes it's hotel and W for washington
 yeah
 103 Operator W for washington
 104 Caller yeah then A for er
 105 Caller atlanta
 106 Operator yeah

In Example 14, one of the few cases where misunderstandings did not get resolved, it is clear that the participants are unable to align due to the similarity in sound of a 'B' and a 'V' (especially for the native Spanish caller). Note that this pair did not manage to ascertain the source of the trouble, which a letter + name using the initial letter strategy may have resolved. A dialogue model should therefore be able to generate this type of strategy for disambiguating letter sounds, even where the human user does not do so.

(14) DEC14:4–112 Complete failure

4 Caller er one is a pub
 5 Caller it's called the star tavern
 6 Operator can you repeat please?
 7 Caller the star
 8 Caller tavern
 : :
 16 Caller yeah the well the place is called
 the star tavern
 17 Operator the star
 18 Caller tavern
 19 Caller yeah
 : :
 29 Operator i'm not sure if i heard the name
 of the place correctly

30 Operator can you repeat?
 31 Caller yeah the the name of the place
 the
 32 Operator yes
 33 Caller the tavern it's the star
 34 Caller star like a star in the sky you
 know <laugh>
 35 Operator yes
 36 Caller the night
 37 Operator mmhm
 38 Caller er tavern
 39 Operator can you spell it er please ta-?
 40 Caller the address you say?
 41 Operator er the star ta- what?
 42 Caller the star tavern
 : :
 58 Caller and it's tavern it's T A
 59 Operator and then
 60 Caller V E er <R> un <N>
 61 Caller N
 62 Caller sorry
 : :
 72 Operator T A B E R N
 73 Operator is that correct?
 74 Caller yeah
 : :
 94 Caller okay you have the name of the
 place correct?
 95 Caller right?
 96 Operator star tabern right?
 97 Caller yeah
 : :
 112 Operator website still says we're sorry we
 co- couldn't find any results

4 Discussion and future work

We have presented a new corpus of telephone directory enquiries that is freely available, and a preliminary exploration of the feedback used in these dialogues.

In future work, we hope to provide a formal model of incremental grounding incorporating the phenomena observed in our corpus including spelling and dictation installments, as well as a comparison with previous work (e.g. Purver, 2004; Fernández, 2006; Rieser and Moore, 2005). Work on formal modelling of grounding (e.g. Traum, 1994; Larsson, 2002; Visser et al., 2014) has often assumed that the minimal units being grounded are words. In a complete model, this needs to be complemented by the grounding of subparts of words, including single letters. Work in this direction includes Skantze and Schlagen (2009), where dictation of number sequences is used as a test case “micro-domain” for an implemented model of incremental grounding. However, this system works exclusively on the level of single digits (or sequences thereof). A challenge for a general model of grounding is to combine grounding of whole words/utterances with grounding of sub-parts of words, using the many strategies that people do.

Acknowledgements

This work was supported by two grants from the Swedish Research Council (VR): 2016-0116 – Incremental Reasoning in Dialogue (IncReD) and 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We are also grateful to the three anonymous reviewers for their helpful comments.

References

- Anne Anderson, Miles Bader, Ellen Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task data. *Language and Speech*, 34(4):351–366.
- Adrian Bangerter, Herbert H Clark, and Anna R Katz. 2004. Navigating joint projects in telephone conversations. *Discourse Processes*, 37(1):1–23.
- Janet Beavin Bavelas, Peter De Jong, Harry Korman, and Sara Smock Jordan. 2012. Beyond backchannels: A three-step model of grounding in face-to-face dialogue. In *Proceedings of Interdisciplinary Workshop on Feedback Behaviors in Dialog*.
- Anastasia Bondarenko, Christine Howes, and Staffan Larsson. 2019. [Directory enquiries corpus](https://osf.io/2vjkh). Available at osf.io/2vjkh.
- Elizabeth A Boyle, Anne H Anderson, and Alison Newlands. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and speech*, 37(1):1–20.
- Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2065–2068. European Language Resources Association.
- Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.
- Hendrik Buschmeier and Stefan Kopp. 2013. Co-constructing grounded symbols–feedback and incremental adaptation in human-agent dialogue. *KI-Künstliche Intelligenz*, 27(2):137–143.
- Nicola Cathcart, Jean Carletta, and Ewan Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the tenth EACL conference*, pages 51–58. Association for Computational Linguistics.
- Harry M Chang. 2007. Comparing machine and human performance for callers directory assistance requests. *International Journal of Speech Technology*, 10(2-3):75–87.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.
- Herbert H. Clark and Susan A. Brennan. 1991. *Grounding in communication*, pages 127–149. Washington: APA Books.
- Herbert H. Clark and Edward A. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.
- Herbert H Clark and Edward F Schaefer. 1987. Collaborating on contributions to conversations. *Language and cognitive processes*, 2(1):19–41.
- Marcus Colman and Patrick G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.
- Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283 – 292.
- Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns. *Language in society*, 3(2):161–180.
- Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King’s College London, University of London.
- Donna T Fujimoto. 2007. Listener responses in interaction: A case for abandoning the term, backchannel. *Journal of Osaka Jogakuin College*, 37:35–54.
- Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.
- Agustín Gravano and Julia Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *INTERSPEECH*, pages 1019–22.
- Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, Greg J. Mills, Ronnie Cann, Wilfried Meyer-Viol, and Patrick G. T. Healey. 2011. Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1):199–233.
- Christine Howes and Arash Eshghi. 2017. Feedback relevance spaces: The organisation of increments in conversation. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*. Association for Computational Linguistics.
- Jacqueline C Kowtko and Patti J Price. 1989. Data collection and analysis in the air travel planning domain. In *Proceedings of the workshop on Speech and Natural Language*, pages 119–125. Association for Computational Linguistics.
- Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University. Also published as Gothenburg Monographs in Linguistics 21.

- Gregory Mills and Patrick G. T. Healey. 2006. Clarifying spatial descriptions: Local and global effects on semantic co-ordination. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, Potsdam, Germany.
- Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.
- Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 239–246, Ann Arbor. Association for Computational Linguistics.
- Christoph Rühlemann. 2007. *Conversation in Context: A Corpus-Driven Approach*. Continuum.
- Gabriel Skantze and David Schlangen. 2009. [Incremental dialogue processing in a micro-domain](#). In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09*, pages 745–753, Stroudsburg, PA, USA. Association for Computational Linguistics.
- David Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.
- Thomas Visser, David Traum, David DeVault, and Rieks op den Akker. 2014. A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8(1):61–73.
- Zhiyang Wang, Jina Lee, and Stacy Marsella. 2011. Towards more comprehensive listening behavior: beyond the bobble head. In *Intelligent Virtual Agents*, pages 216–227. Springer.

A Corpus Study on Questions, Responses and Misunderstanding Signals in Conversations with Alzheimer’s Patients*

Shamila Nasreen, Matthew Purver, Julian Hough

Cognitive Science Group / Computational Linguistics Lab

School of Electronic Engineering and Computer Science

Queen Mary University of London Mile End Road,

London E1 4NS, UK

{shamila.nasreen,m.purver,j.hough}@qmul.ac.uk

Abstract

This paper describes an initial corpus study of question-answer pairs in the Carolina Conversations Collection corpus of conversational interviews with older people. Our aim is to compare the behaviour of patients with and without Alzheimer’s Disease (AD) on the basis of types of question asked and their responses in dialogue. It has been suggested that questions present an interesting and useful phenomenon for exploring the quality of communication between patients and their interlocutors, and this study confirms this: questions are common, making up almost 14% of utterances from AD and Non-AD patients; and type distributions vary, interviewers asking many Yes-No questions (nearly 6%) from AD patients while more Wh-questions (5.4%) from Non-AD patients. We also find that processes of clarification and coordination (e.g. asking clarification questions, signalling non-understanding) are more common in dialogue with AD patients.

1 Introduction

Alzheimer’s Disease (AD) is an irreversible, progressive deterioration of the brain that slowly destroys memory, language and thinking abilities, and eventually the ability to carry out the simplest tasks in patients’ daily lives. AD is the most prevalent form of dementia, contributing to 60%-70% among all types of dementia (Tsoi et al., 2018). The most common symptoms of AD are memory lapses, difficulty in recalling recent events, struggling to follow a conversation, repeating the

conversation, delayed responses, difficulty finding words for talk, and orientation problems (e.g. confusion and inability to track daily activities).

Diagnosis can be based on clinical interpretation of patients’ history complemented by brain scanning (MRI); but this is time-consuming, stressful, costly and often cannot be offered to all patients complaining about functional memory. Instead, the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and Alzheimer’s Association established criteria for AD diagnosis require the presence of cognitive impairment to be confirmed by neuropsychological testing for a clinical diagnosis of possible or probable AD (McKhann et al., 1984). Suitable neuropsychological tests include the Mini-Mental Status Examination (MMSE; Folstein et al., 1975, one of the most commonly used tests), Mini-Cog (Rosen et al., 1984), Addenbrooke’s Cognitive Examination Revised (ACE-R; Noone, 2015), Hopkins Verbal Learning Test (HVL; Brandt, 1991) and DemTect (Kalbe et al., 2004).

However, these tests require medical experts to interpret the results, and are performed in medical clinics which patients must visit for diagnosis. Currently, researchers are therefore investigating the impact of neurodegenerative impairment on patients’ speech and language, with the hope of deriving tests which are easier to administer and automate via natural language processing techniques (see e.g. Fraser et al., 2016a).

In this paper, we focus on language in conversational interaction. We explore this as a diagnostically relevant resource to differentiate patients with and without Alzheimer’s Disease (AD vs. Non-AD), using the Carolina Conversations Collection data in which patients interact with researchers and community persons on different but not prefixed topics like discussion about breakfast, lunch, special occasions (thanksgiving, Christ-

*This research was partially supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBED-DIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors’ views and the Commission is not responsible for any use that may be made of the information it contains.

mas) etc. We particularly focused on the types of questions asked from both groups, how they are responded to, and whether there are any significant patterns that appear to differentiate the groups.

The remainder of this paper is organized as follows. In the next section, we describe earlier work on analyzing conversational profiles of AD and particularly on the types of questions they focused on. In Section 3 we give details about our new corpus study and annotation scheme. We then present and discuss the results in Section 4: in particular, how the distributions of different types of questions, and frequencies of signals of non-understanding, clarification questions and repeat questions, differ between AD patients and Non-AD. We close with a discussion of the overall result, and of possible further work.

2 Related Work

Recent years have seen an increasing amount of research in NLP for dementia diagnosis. Much of this work has looked at properties of an individual’s language in isolation: acoustic and lexical features of speech (Jarrold et al., 2014), or syntactic complexity, fluency and information content (Fraser et al., 2016b,a). However, this is usually studied within particular language tasks, often within specific domains (e.g. the Cookie Theft picture description task of the DementiaBank Pitt Corpus¹); however, conversational dialogue is the primary area of human natural language use, and studying the effects of AD on dialogue and interaction — and particularly more open-domain dialogue — might therefore provide more generally applicable insights.

Recent statistical modelling work shows that AD has characteristic effects on dialogue. Luz et al. (2018) extract features like speech rate, dialogue duration and turn taking measures, using the Carolina Conversations Collection corpus (Pope and Davis, 2011) of patient interview dialogues, and show that this can build a predictive statistical model for the presence of AD.

Work in the conversation analysis (CA) tradition has looked in more detail at what characteristics of dialogue with dementia might be important. Jones et al. (2016) present a CA study of dyadic communication between clinicians and patients during initial specialist clinic visits, while Elsey et al. (2015) highlighted the role of carer,

looking at triadic interactions among a doctor, a patient and a companion. They establish differential conversational profiles which distinguish between non-progressive functional memory disorder (FMD) and progressive neuro-degenerative Disorder (ND), based on the interactional behavior of patients responding to neurologists’ questions about their memory problems. Features include difficulties responding to compound questions, providing specific and elaborated examples and answering questions about personal information, time taken to respond and frequent “I don’t know” responses.

Questions present an interesting testing ground when exploring the quality of communication between caregivers and persons with AD. Question-answer sequences have long been seen as a fundamental building block of conversation; Sacks et al. (1978) formalized these as a type of adjacency pair in which the first utterance represents the question and the second one is an answer. Hamilton (2005) explored the use of questions in conversation with a patient of AD over a period of four years, finding that Yes-No questions are responded to much more frequently than open-ended question i.e Wh-questions. Gottlieb-Tanaka et al. (2003) used a similar approach, examining Yes-No and open-ended questions in a conversation between family caregivers and their spouse with AD during different activities of daily life. They reported that caregivers used YesNo questions much more frequently than open-ended questions (66% vs. 34%, respectively) and there are fewer communication breakdowns with Yes-No Questions.

Varela Suárez (2018) worked specifically to observe dementia patients’ ability to respond to different types of questions including close-ended questions, open-ended questions, and multiple choice questions. The objective of this study was to verify a) if the ability to answer questions persists until the final stages of dementia b), check if the number of preferred and relevant answers decreases progressively. The interviewers had a list of questions about patients memories, experiences, and daily routine, and were told to talk on the topics introduced by the patients, and only ask the questions from the list when patients are silent. The basic Question-Answer adjacency pair is preserved until the severe stage of the disease; however, the number of answered questions, preferred and relevant answers starts to decrease.

¹<http://talkbank.org/DementiaBank/>

These studies show that the presence of AD affects the production of questions, their use and their responses, but all focus on specific types of question including Yes-No, Wh-questions, and Multiple choice questions. As far as we are aware, none of these studies have extended this approach to look into specific aspects of non-understanding or inability to respond: e.g. non-understanding signals, clarification requests and repetition of questions.

Dialogue Act Models

The ability to model and detect discourse structure is an important step toward working spontaneous dialogue and the first analysis step involves the identification of Dialogue Acts (DAs). DAs represent the meaning of utterances at the level of illocutionary force (Stolcke et al., 2000). Classifying utterances and assigning DAs is very useful in many applications including answering questions in conversational agents, summarizing meeting minutes, and assigning proper DAs in dialogue based games. DAs tagsets classify dialogue utterances based on the syntactic, semantic and pragmatic structure of the utterance.

The most widely used dataset and tagset in DA tagging is the Switchboard corpus, consisting of 1155 annotated conversations containing 205K utterances, 1.4 million words from 5 minute recorded telephonic conversations. The DA types and complete tagset can be seen in (Jurafsky et al., 1997). The corpus is annotated using a variant of the DAMSL tagset (Core and Allen, 1997) with approximately 60 basic tags/classes which combines to produce 220 distinct labels. Jurafsky et al. (1997) then combine these 220 labels into 43 major classes including *Statements*, *Backchannels*, *Questions*, *Agreements*, *Apology* etc.

3 Material and Methods

3.1 Research Questions

This study is a part of a larger project where we analyze what are the significant key indicators in the language and speech of AD patients that can be used as Bio-Markers in the early diagnosis process of Alzheimer’s Disease. The focus of the initial and current study is on the interaction of AD patients and Non-AD patients with interviewers.

Our account suggests these interactions are based on what is being asked from the AD and Non-AD sufferers. We hypothesize that the distri-

bution of questions being asked and the responses generated are not same for both the groups. We hypothesize that the use of different question types such as binary yes-no questions (in interrogative or declarative form), tag questions, and alternative (‘or’) questions will differ between groups; and the signals of non-understanding, back-channels in question form and clarification requests should be more common with AD patients.

In more detail, we are conducting this corpus study to answer the following research questions:

Q1 *Is the distribution of question types asked by the patient and interviewer different when the patient is an AD sufferer?*

Our first interest is in the general statistics regarding what types of questions are asked of the AD and non-AD group. How often does each type occur, and what is the balance between the two groups? What types of questions are more frequently asked from Alzheimer’s patients?

Q2 *How often do signals of non-understanding, clarification requests and back-channel questions occur in dialogues with an AD sufferer compared to those without one?*

We hypothesize that due to the nature of AD, there will be more non-understanding signals and clarification questions in response to questions and statements.

Q3 *Is the distribution of simple-repeat and reformulation questions different for conversations with an AD sufferer compared to those without one?*

We hypothesize that there will be more repeated questions for the AD group from the interviewer, as AD patients find it difficult to follow a conversation.

3.2 Corpus

Our intention was to investigate the behavior of AD patients on the basis of questions and responses observed in a corpus of dialogue. For this purpose, we used the Carolina Conversation Collection (CCC), collected by the Medical University of South Carolina (MUSC)² (Pope and Davis, 2011). This dataset comprises of two cohorts: cohort one contains 125 unimpaired persons of 65

²<https://carolinaconversations.musc.edu/>

years and older with 12 chronic diseases with a total of 200 conversations. Cohort two includes 400 natural conversations of 125 persons having dementia including Alzheimer’s of age 65 and above who spoke at least twice annually with linguistic students. The demographic and clinical variables include: age range, gender, occupation prior to retirement, diseases diagnosed, and level of education (in years) are available. As this dataset includes only older patients with diagnosed dementia, it can only allow us to observe patterns associated with AD at a relatively advanced stage, and not directly tell us whether these extend to early stage diagnosis. However, it has the advantage of containing relatively free conversational interaction, rather than the more formulaic tasks in e.g. DementiaBank. Work in progress is collecting a dataset of conversational language including early-stage and un-diagnosed cases; until then we believe this to be the most relevant corpus for our purposes.

The dataset consists of audio, video and transcripts that are time aligned. The identity of patients and interviewer is anonymized keeping in mind security and privacy concerns. Online access to the dataset was obtained after gaining ethical approval from Queen Mary University of London (hosting the project) and Medical University of South Carolina (MUSC, hosting the dataset), and complying with MUSC’s requirements for data handling and storage.

For our corpus analysis here, we used dialogue data from 10 randomly sampled patients with AD (7 females, 3 males) and 10 patients with other diseases including diabetes, heart problems, arthritis, high cholesterol, cancer, leukemia and breathing problems but not AD (8 females, 2 males). These groups are selected to match age range, to compare the different patterns of interaction and to avoid statistical bias. This portion comprises of 2554 utterances for the AD group and 1439 utterances for the Non-AD group, with a total of 3993 utterances from 20 patients with 23 dialogue conversations.

The CCC transcripts are already segmented at the utterance (turn) level and the word level, and annotated for speaker identity (patient vs. interviewer); however, no DA information is available. We used only the utterance level layers; transcripts were available in ELAN format and we converted them to CSV format. We then manually annotate the transcripts at the utterance level with DA in-

formation.

3.3 Terminology

Throughout this paper, we use specific terms for particular question types and response types, and use these in our annotation procedure. Following Switchboard’s SWBD-DAMSL terminology (Jurafsky et al., 1997), we use **qy** for **Yes-No** questions, and **qy^d** for **Declarative Yes-No** questions. Declarative questions (^d) are utterances which function pragmatically as questions but which do not have “question form” in their syntax. We use **qw** for **Wh-questions** which includes words like *what, how, when, etc.* and **qw^d** for **Declarative Wh-questions**. Yes-No or Wh-questions are questions which do not have only pragmatic force but have a syntactic and prosodic marking of questions or interrogative in nature. We used **g** for **Tag questions**, which are simply confirming questions that have auxiliary inversion at the end of statement e.g. (*But they’re pretty, aren’t they?*). For **Or questions** which are simply choice question and aids in answering the question by giving choices to the patients are represented by **qr** e.g. (*- did he um, keep him or did he throw him back?*).

We used term **Clarification question** for questions that are asked in response to a partial understanding of a question/statement and are specific in nature. These clarification questions are represented by **qc**. **Signal non-understanding** is generated by a person in response to a question that they have not understood and are represented by **br**. **Back-channel Question (bh)** is a continuer which takes the form of question and have question intonation in it. Back-channels are more generic than clarification questions and often occur in many types (*e.g really? Yeah? do you? is that right? etc.*).

When the response to a Yes-No question is just a yes including variations (e.g. *yeah, yes, huh, yes, Yes I do etc.*), it will be represented by **ny** and when there is a yes plus some explanation, it will be represented by **ny^e**.

(1) A: Do you have children?

B: Yeah, but they’re big children now. Grown.

[CCC Mason_Davis_001 28-29]

na is an affirmative answer that gives an explanation without the yes or its variation. **nn** is used for

No-answers and **nn^e** is used for an explanation with No answer (see Appendix A for Examples).

3.4 Annotation Scheme

The original SWBD-DAMSL tagset for the Switchboard Corpus contains 43 DA tags (Jurafsky et al., 1997). Our initial manual includes DA tags from SWBD-DAMSL and our own specific new DA tags with a total of 35 tags. For different types of questions and their possible responses, 14 DA tags are taken from SWBD-DAMSL and 2 new tags are introduced. These new tags are for clarification questions (**qc**) and for answers to Wh-Questions (**sd-qw**), and were required to distinguish key response types.³

The ability to tag specific clarification questions is important for our study, as questions asked by the interviewer can be followed by a clarification which indicates partial understanding while requesting specific clarifying information (SWBD-DAMSL only provides the **br** tag for complete non-understanding). The distinction between answers to Wh-Questions and other, unrelated statements is also important (in order to capture whether the response is relevant: a relevant answer should be different from simple general statement), but SWBD-DAMSL provides only a single **sd** tag for statements. Different types of question and their tags are given with examples in Table 1; a list of response types is given in Table 2.

Another new addition is the tagging of *repetition* of questions, with or without reformulation. We marked repeat questions as simple repeats or reformulations, and tagged with the index of the dialogue act (utterance number) they were repeating or reformulating.

Similarly, clarification questions can signal non-understanding with two main distinct CR forms, and this distinction is tagged: pure repeats and reformulated repeated questions that are slightly changed syntactically but the context remains the same – see Table 3 with utterance 144.

3.5 Inter-Annotator Agreement

To check inter-annotator agreement, three annotators annotated one conversation of an AD patient and Non-AD interviewer of 192 utterances. All

³Some other DA tagging schemes provide categories for these and more; however, we chose to begin with SWBD-DAMSL given its prevalence in DA tagging work, and extend it only as necessary. In future work we plan to examine multi-dimensional schemes (e.g. Core and Allen, 1997; Bunt et al., 2010) to see if they provide benefits in this setting.

annotators had a good knowledge of linguistics and were familiar with both the SWBD-DAMSL tagset and the additions as specified above and in the manual. First, all three annotators annotated the dialogue independently by assigning DA tags to all utterances with the 17 tags of interest for this paper as shown in Table 4 (‘other’ means the annotator judged another SWBD-DAMSL act tag could be appropriate apart from the 16 tags in focus). We use a multi-rater version of Cohen’s κ (Cohen, 1960) as described by (Siegel and Castellan, 1988) to establish the agreement of annotators for all tags and also 1-vs-the-rest as shown in Table 4 below.⁴

As can be seen, an overall agreement was good ($\kappa=0.844$) for all tags and the majority of tags which were tagged by any annotator in the dialogue have $\kappa > 0.67$, with only ‘no’ getting beneath $\kappa < 0.5$. We judged this test to be indicative of a reliable annotation scheme for our purposes.

4 Results and Discussion

From the CCC transcripts, we selected 23 conversations, which when annotated yield 3993 utterances. All utterances were tagged with one of the 16 dialogue act tags relating to all question categories and their possible answers as described above, plus an ‘other’ tag. In addition to the dialogue act tag, utterances deemed to be responses (tags in Table 2) were tagged with the index of the utterance being responded to. Repeat questions were also marked as *simple repeats* or *reformulations*, and tagged with the index of the dialogue act they were repeating or reformulating.

Is the distribution of question types asked by the patient and interviewer different when the patient is an AD sufferer?

To investigate the distribution of dialogue acts, we calculated the relative frequency of each question and response type separately for AD and Non-AD group, and for the patient and interviewer within those groups. A comprehensive analysis of particular types and their distribution between AD and Non-AD patient with their interviewer is shown in Table 5. More yes-no questions (qy) are asked by the interviewer from AD Patients than Non-AD patients (6% vs 3.7%) and fewer wh-questions

⁴The annotation results and scripts are available from https://github.com/julianhough/inter_annotator_agreement.

Type	Tag	Example
Yes-No Question	qy	Did you go anywhere today?
Wh-Question	qw	When do you have any time to do your homework?
Declarative Yes-No Question	qy^d	You have two kids?
Declarative Wh-Question	qw^d	Doing what?
Or Question	qr	Did he um, keep him or did he throw him back?
Tag Question	^g	But they're pretty aren't they?
Clarification Question	qc	Next Tuesday?
Signal Non-understanding	br	Pardon?
Backchannel in question form	bh	Really?

Table 1: Question Types for CCC

Type	Tag	Example
Yes answer	ny	Yeah.
Yes- plus expansion	ny^e	Yeah, but they're
Affirmative non-yes answer	na	Oh I think so. [laughs]?
No answer	nn	No
Negative non-no answers	nn^e	No, I belonged to the Methodist church.
Other answer	no	I, I don't know.
Declarative statement wh-answer	sd-qw	Popcorn shrimp and it was leftover from yesterday.

Table 2: Answer Types for CCC

Tag	Speaker:Utterance	Text	Repeat Question?
qw	A:15	-Where's she been?	15
br	B:16	-Pardon?	
qw	A:17	-Where is she been?	
qy	A:142	-Well, are you, are you restricted from certain foods?	142-reformulation
br	B:143	-What?	
qy	A:144	-Like, do they, do they make you eat certain foods because your medication?	

Table 3: Examples of Repeated questions

(qw) are asked in the AD group compared to the non-AD group (4% vs 5.4%). Choice questions (qr) are also asked more from AD patients compared to non-AD patients (2% vs 0.3%). These results suggest there is a systematic difference in question distributions; one plausible explanation for this is that AD patients find it easier to answer a simple Yes-No question or a choice question compared to a wh-question. It is also obvious from the results that AD patients are also asking more questions than Non-AD patient during their conversation with the interviewer (*qy*: 1% vs 0.3%), (*qw*: 1% vs 0.3%), (*^g*: 0.2% vs 0.1%), (*br*: 3% vs 0.4%), and (*qc*: 2% vs 0.1%).

We also compared the distribution of these tags with the Switchboard SWDA corpus, as shown in Table 6. As the CCC is a set of clinical interviews, the percentage of tags which are questions is higher in this corpus compared to Switchboard. Although simple yes-no questions have almost identical frequencies in both corpora, declarative yes-no, wh-questions, declarative wh-questions, tag questions, and signals of non-understanding are higher in the CCC than Switchboard. Our new clarification question (*qc*) tag accounts for 1% for both AD group and Non-AD group tags but is not annotated in SWDA.

Tag	# times annotated	κ
qy	26	0.758
qw	30	0.895
qy^d	12	0.660
qw^d	3	1.000
^g	2	0.498
br	22	0.953
bh	0	0
qc	15	0.795
qr	0	0
ny	12	1.000
ny^e	11	0.907
na	8	0.873
nn	1	0
nn^e	6	0.663
no	4	0.497
sd-qw	26	0.637
other	398	0.902
all tags	576	0.844

Table 4: Multi-rater Cohen’s κ statistics for one-vs-rest and overall agreement score for one dialogue.

DA tag	AD		Non-AD	
	Pat	Int	Pat	Int
qy	1%	6%	0.3%	3.7%
qy^d	1%	6%	0.1%	5%
qw	1%	4%	0.1%	5.4%
qw^d	0.4%	1%	0.5%	0
^g	0.2%	2%	0.1%	0.7%
qr	0.1%	2%	0	0.3%
br	3%	0.1%	0.4%	0
bh	1%	1%	1%	1%
qc	2%	1%	0.1	1%
simple-Repeat	0	1%	0	0
reformulation	0	2%	0	0

Table 5: Distribution of DA question tags among the AD group and Non-AD group

How often do signals of non-understanding, clarification requests and back-channel questions occur in dialogues with an AD sufferer compared to those without one?

An examination of signals of non-understanding, clarification requests and back-channel requests reveals that the ability to follow and understand questions decrease for AD patients so they produce more signals of non-understanding (e.g *sorry* *Maam?*, *Pardon?*, *huh?*, *eh?*), when questions are posed to them. On the other hand, signals of non-

DA Tag	CCC-AD	CCC-Non-AD	SWDA
qy	3%	2%	2%
qy^d	4%	2%	1%
qw	3%	3%	1%
qw^d	1%	0.3%	<.1%
^g	1%	0.5%	<.1%
br	1%	0.2%	0.1%
bh	1%	1%	1%
qc	1%	1%	-
qr	1%	0.2%	0.1%
ny	3%	1%	1%
ny^e	2%	2%	0.4%
na	3%	3%	1%
nn	0.4%	0.4%	1%
nn^e	1%	1%	0.1%
no	0.4%	0.3%	1%
sd-qw	4%	6%	-

Table 6: Comparison of relative frequency of DA tags in the AD group, Non-AD group of the CCC and SWDA corpora

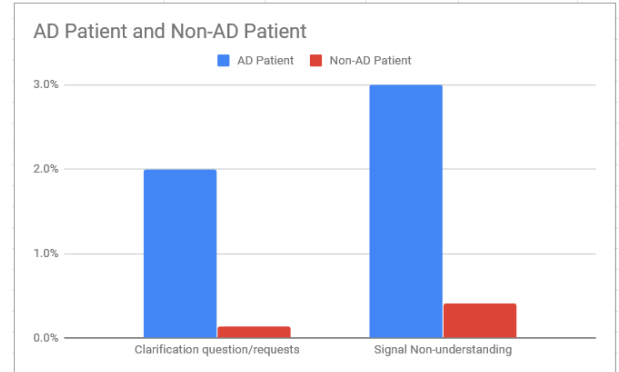


Figure 1: Clarification questions and Signal Non-understanding

understanding from Non-AD patients are much less frequent as shown in Figure 1. The overall frequency of clarification questions (qc) between the two conversation groups was not systematically different as shown in Table 6 when utterances from both patient and interviewer are combined, but dealing with them separately, AD patients produce more clarification requests than non-AD patients (2% vs 0.1%) – see Table 5 and Fig. 1.

We further examine how often signals of non-understanding and clarification requests are issued in response to questions rather than statements/answers. Examination of the data shows that clarification requests are more often gener-

	AD Group	Non-AD Group
Question followed by Signal of Non-understanding	24 (35)	2 (3)
Statements followed by Signal of Non-understanding	11 (35)	1 (3)
Question followed by Clarification Question	8 (34)	1 (11)
Statement followed by Clarification Question	26 (34)	10 (11)

Table 7: Occurrences of signal non-understanding and clarification question followed by question/statements

ated in response to statements, and less often after questions are raised; but signal non-understanding happen more often after questions. Out of total 35 signal non-understanding, 24 are generated in response to a question of AD Group as shown in Table 7. However, only 8 clarification questions are asked in response to questions, with 26 asked in response to declarative statements – (see Appendix A for more examples and context).

Is the distribution of simple-repeat and reformulation questions different for conversations with an AD sufferer compared to those without one?

Many questions are followed by clarification questions or signal non-understanding, so there will be more repetition of a similar type of question in case of the AD patients. Repeated questions are asked in two variations; either repeated simply or reformulated so that the patient can understand the question properly as in (4). In the AD group 4.7% questions are simple-repeat questions and 6.7% are reformulated as shown in Table 8 while for the non-AD group only 2.4% are reformulated questions and there were no repeated questions.

- (4) A: Your dad worked for who was it? Swisten
A: and that's why you went up to Baltimore?.
B: Huh?
A: Your dad went to –worked at – worked for Swisten?
B: My Father?
A: Yeah. Is that why you guys went to Baltimore?

[CCC Tappan_Patte_001 37-43]

5 Conclusion and Future work

Our study provides the first statistical analysis of different types of question asked in conversations

Repeat Type	AD Group	Non-AD Group
Total Question	313	127
Simple-Repeat Question	15 (4.7%)	0
Reformulated Question	21 (6.7%)	3 (2.4%)

Table 8: Repetition and reformulation of questions for AD group and Non-AD group

with AD patients in the Carolina Conversation Collection (CCC) Corpus. We found that yes-no questions were asked more frequently in the AD sufferer conversations than the Non-AD conversations (6% vs 3.7% of all dialogue acts) and less Wh-questions were asked in AD sufferer conversations compared to Non-AD ones (4% vs 5.4%). While our newly introduced tags were not frequent, they are significant in AD sufferer conversations, with 2% of all dialogue acts by AD sufferers being clarification questions and 3% being signals of non-understanding.

In future work, we plan to work on the CCC corpus conversations of both AD and Non-AD conversations to build an automatic dialogue act tagger for the tagset we used in this study. We will also explore more complex questions including compound questions and questions that relate to semantic memory and episodic memory. We also plan to look into disfluency and repairs in this data collection which could further aid interpretation and automatic diagnosis.

References

- Jason Brandt. 1991. The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5(2):125–142.
- Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC 2010, the Seventh International Conference on Language Resources and Evaluation*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56. Boston, MA.
- Christopher Elsey, Paul Drew, Danielle Jones, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, Annalena Venneri, and Markus Reuber. 2015. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, 98(9):1071–1077.
- M F Folstein, S E Folstein, and P R McHugh. 1975. Mini-mental status. a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.
- Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016a. Linguistic features identify Alzheimer’s disease in narrative speech. *Journal of Alzheimer’s Disease*, 49(2):407–422.
- Kathleen C. Fraser, Frank Rudzicz, and Graeme Hirst. 2016b. [Detecting late-life depression in Alzheimer’s disease through analysis of speech and language](#). In *Proc. CLPsych*, pages 1–11, San Diego, CA, USA. Association for Computational Linguistics.
- Dalia Gottlieb-Tanaka, Jeff Small, and Annalee Yassi. 2003. A programme of creative expression activities for seniors with dementia. *Dementia*, 2(1):127–133.
- Heidi Ehernberger Hamilton. 2005. *Conversations with an Alzheimer’s patient: An interactional sociolinguistic study*. Cambridge University Press.
- William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. [Aided diagnosis of dementia type through computer-based analysis of spontaneous speech](#). In *Proc. CLPsych*, pages 27–37, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Danielle Jones, Paul Drew, Christopher Elsey, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, and Markus Reuber. 2016. Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders. *Aging & Mental Health*, 20(5):500–509.
- Daniel Jurafsky, Elizabeth Shriberg, and Debra Binasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual.
- Elke Kalbe, Josef Kessler, Pasquale Calabrese, R Smith, AP Passmore, Met al Brand, and R Bullock. 2004. DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *International journal of geriatric psychiatry*, 19(2):136–143.
- Saturnino Luz, Sofia de la Fuente, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. In *Proceedings of the LREC 2018 Workshop Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)*.
- Guy McKhann, David Drachman, and Marshall Folstein. 1984. [Clinical diagnosis of Alzheimer’s disease](#). *Neurology*, 34(7):939—944. Views & Reviews.
- Peter Noone. 2015. Addenbrooke’s Cognitive Examination-III. *Occupational Medicine*, 65:418–420.
- Charlene Pope and Boyd H Davis. 2011. Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.
- W G Rosen, R C Mohs, and K L Davis. 1984. A new rating scale for Alzheimer’s disease. *American Journal of Psychiatry*, 141(11):1356–1364.
- Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. Elsevier.
- Sidney Siegel and NJ Castellan. 1988. Measures of association and their tests of significance. *Nonparametric Statistics for the Behavioral Sciences*, pages 224–312.
- 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.
- Kelvin K F Tsoi, Lingling Zhang, Nicholas B Chan, Felix C H Chan, Hoyee W Hirai, and Helen M L Meng. 2018. Social Media as a Tool to Look for People with Dementia Who Become Lost : Factors

That Matter. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 9:3355–3364.

Ana Varela Suárez. 2018. The question-answer adjacency pair in dementia discourse. *International Journal of Applied Linguistics*, 28(1):86–101.

A Examples from Carolinas Conversation Collection

Yes-No Question followed by no plus expansion answer:

Tag	Text
qy nn ^e	A: were you Primitive Baptist? B: — no, I belonged to the Methodist church.

[CCC Mason_Davis_001 92-93]

Yes-No Question followed by other answer:

Tag	Text
qy no	A: are you going to go with them to see the Christmas Lights? B: Oh, I, I dont know.

[CCC Wakefield_Brock_001 51-52]

Two Wh-Questions followed by declarative statements wh-answer:

Tag	Text
qw sd-qw	A: - what does he preach about? B: – hell hot and heaven beautiful.
qw sd-qw +	C: what types of food do you like the best? D – vegetables, meat, - and desserts.

[CCC Mason_Davis_001 31-32]

[CCC Wakeman_Rhyne_001 6-7]

Wh-question followed by a clarification question(*qc*) and a wh-question followed by a statement and then a clarification(*qc*):

Tag	Text
qw qy br qw qc qw sd-qw qc	A: where is Jerusalem Primitive Baptist Church? - is that near Fountain Hill? B: - m'am? A: where is that church? B: Fountain Hill? A: what do you do? B - I'm a teacher. A: Preacher?-

[CCC Mason_Davis_001 83-86,64-66]

Declarative wh-question followed by signal non-understanding(*br*) and then by reformulated-repeat wh-question:

Tag	Text
qw ^d br qw	A: You were married for– B: Huh? A: How long– have you been married? (reformulated-repeat)

[CCC Tappan_Patte_001 7-9]

Declarative statement followed by back-channel question(*bh*) and then by yes answer:

Tag	Text
sd bh ny	A: huh, it used to be something special. it used to be my Mother's birthday. B: Really ? A: Yeah

[Wheaden_Lee_001 52-54]

How to Reject *What* in Dialogue

Julian J. Schlöder and Raquel Fernández

Institute for Logic, Language and Computation

University of Amsterdam

julian.schloeder@gmail.com raquel.fernandez@uva.nl

Abstract

The ability to identify and understand *rejection moves* in dialogue is vital for successful linguistic interaction. In this paper, we investigate the different linguistic strategies available to express rejection and categorise them in a two-dimensional taxonomy. To wit, we categorise rejections by *what* aspect of their target utterance they reject and by *how* that rejection is expressed. Using this taxonomy, we annotate a set of 400 natural instances of rejection moves. From this data, we draw some tentative conclusions about the role of certain linguistic markers (such as polarity particles) with respect to the different strategies for expressing rejection.

1 Introduction

Partaking in a dialogue requires all interlocutors to coordinate on what they mutually take for granted, i.e. their *common ground* (Stalnaker, 1978) or their *shared commitments* (Asher and Lascarides, 2008). That is, dialogue proceeds (at least in part) by the making and accepting of proposals to update the shared information through the collaborative process of *grounding* (Clark, 1996; Poesio and Traum, 1997; Ginzburg, 2012).

However, the process of grounding can fail. A substantial part of prior research focuses on failures resulting from various kinds of misunderstandings (e.g. issues related to the acoustic channel, parsing, reference retrieval) and the mechanisms to repair such misunderstandings (e.g. clarification requests) (Schegloff et al., 1977; Purver, 2004; Ginzburg and Cooper, 2004; Schlagen, 2004). Moreover, it is evidently the case that not every proposal made in a dialogue is acceptable to all participants. Hence, even in the absence of misunderstandings, grounding can fail because one participant in the conversation *rejects* the proposal to update the common ground. As we point out

in earlier work (Schlöder and Fernández, 2015a), there is a continuity between rejections and other failures to ground. Notably, the repair mechanisms associated with rejections are clarification requests like *Why not?*.

Hence, to maintain coordination on what is mutually supposed, it is incumbent on any participant in a conversation to keep track of which proposals have been rejected. This issue also arises in some practical applications, e.g. summarisation tasks, for which one needs to compute which issues have been raised in a dialogue and which of these issues have been accepted (Galley et al., 2004; Wang et al., 2011; Misra and Walker, 2013).

It is however far from trivial to determine whether some response to a proposal constitutes a rejection (Horn, 1989; Walker, 1996; Lascarides and Asher, 2009). Compare for example (1b) and (2b), taken from Schlöder and Fernández (2014). Both have the same surface form, but when considered in context the former is an acceptance move whereas the latter is a rejection (also see Farkas and Bruce, 2010, Roelofsen and Farkas, 2015 for formal takes on the ambiguity of such responses).

- (1) a. A: But its uh yeah its uh original idea.
b. B: Yes it is.
- (2) a. A: the shape of a banana is not its not really handy.
b. B: Yes it is.

Comparing (3), (4) and (5) reveals another interesting contrast. The utterance (3b) rejects by making a counterproposal, i.e. by making a proposal that is incompatible with the proposal that is being rejected. This is not so in (4) and (5), where the second utterance rejects the first, but the propositional contents of proposal and response are compatible. This can be seen by observing that

the contents of (4a) and (5b), respectively, entail the contents of (4b) and (5a).

(3) a. B: Yes, a one.

b. A: I say a two.

(4) a. B: No that's for the trendy uh feel and look.

b. C: Yeah but everything is.

(5) a. A: It's your job.

b. B: It's our job.

The rejecting force of (4b) and (5b) can instead be appreciated as follows. (4b) rejects (4a) by implicating *normal* \leadsto *not interesting* whereas (5b) is rejecting the implicature of (5a) that *your job* \leadsto *not my job* (Schlöder and Fernández, 2015b).

In this paper, we aim to get a more comprehensive and systematic picture of the different ways to express a rejection. We consider three dialogue corpora that are annotated with categories that allow us to identify rejection moves: The AMI Meeting Corpus (Carletta, 2007), the ICSI Meeting Corpus (Janin et al., 2003) and the Switchboard Corpus (Godfrey et al., 1992). We survey the rejection moves found in these corpora and develop a taxonomy that classifies them along two dimension: *what* they reject, and *how* this rejection is expressed. To see how these dimensions interact, we annotate a substantial fragment of the rejection moves in these corpora.

In the following section we outline some previous theoretical work about rejecting speech acts, noting that some substantial assumptions go into our working definition of *rejection move*. In Section 3 we present our taxonomy, including multiple examples from our corpora for each category. We describe our annotation procedure in Section 4 and summarise our results in Section 5.

2 Theoretical Background

To investigate the notion of *rejecting force* in dialogue requires making some theoretical choices. One tradition, going back to Frege (1919), sees a rejection of a content *p* as equivalent to the assertion that *not p*. Another tradition, where this is not so, may be traced back to Austin (1962). Austin talks about *cancellations* of arbitrary speech act, which amount to making it so that the effects of the cancelled speech act do not obtain. This latter, Austinian notion seems to be more appropriate for the study of dialogue.

When we talk about grounding a dialogue act, we mean that the act is *taken up* such that a certain,

essential effect of that act obtains (Clark, 1996). In the context of assertion, that effect would be that the assertion's content becomes common ground (Stalnaker, 1978). Cancellation (or, rejection) of that effect means that the content does *not* become common ground—but not that the negation of that content becomes common ground (which would be the essential effect of a Fregean rejection). Indeed, Stalnaker (1978) himself appears to espouse the Austinian view:

“It should be made clear that to reject an assertion is ... to refuse to accept the assertion. If an assertion is rejected, the context [common ground] remains the same as it was.” (Stalnaker, 1978, p.87).

Sometimes, Stalnakerian models are associated with the idea that the essential effect of an assertion—i.e. addition to common ground—is achieved immediately after the assertion has been made or understood (e.g. Murray, 2009). Taking rejection seriously reveals this to be a simplification. The actual picture is more complicated: the essential effect obtains *only if* the assertion has not been rejected. This means that one may view assertions as *proposals* to achieve their essential effect. That proposal is up for discussion and may be cancelled (Incurvati and Schlöder, 2017).

Note, however, that after an assertion is understood, something *is* immediately added to common ground: that the proposal to update the common ground with the assertion's content has been made. Stalnaker (1978) calls this the *first effect* (to be distinguished from the second, essential effect). This effect “cannot be blocked” (p. 87). Thus what is up for rejection is exactly the essential effect.

So far, this applies only to rejections of assertions, but as pointed out by Schlöder et al. (2018) one may associate *every* dialogue act with a proposal to achieve some essential effect that characterises what happens upon successful grounding of that act. They identify this effect as the *speech-act related goal* of the dialogue act, in the sense of Asher and Lascarides (2003). One significant consequence of this view, noted by Schlöder et al., is that one may reject rejections. To wit, a rejection proposes to achieve the effect of leaving some prior dialogue act ungrounded—this itself is up for acceptance or cancellation.

Thus, following Stalnaker and these additional considerations, we say that an utterance has *rejecting force* if it is interpreted as a proposal to *not*

achieve the essential effect of an earlier utterance. For example, assertions that *p* are rejected by utterances that propose to *not* add *p* to common ground. This may be achieved by asserting *not p*, but not necessarily (Khoo, 2015). Questions are rejected by dialogue acts that propose to not make any answer common ground; Commands are rejected by dialogue acts that propose to not create the obligations proposed by the command. Etc.

Furthermore, the essential effect of a dialogue act may be pragmatically enriched (Lascarides and Asher, 2009). That is, for example, an assertion proposes to make common ground not just its literal content, but also all of its implicatures. Hence rejections of implicatures, as seen in example (5) are rejections. Similarly, a dialogue act may implicate a rejection, as seen in example (4).

In what follows, we adopt the following terminology: an utterance is a rejection if it is about the essential effect of a prior utterance (the *rejection target*) and if it proposes to *not fully* achieve the (pragmatically determined) essential effect of that utterance. We refer to the part of the rejection target’s essential effect that is proposed to remain unachieved as *what is rejected*.

3 How To Reject What: A Taxonomy

To create a taxonomy of the different ways in which rejections may be expressed, we surveyed a fragment of 250 utterances annotated as rejection in the AMI Meeting Corpus (Carletta, 2007) and identified commonalities. In this section, we discuss these categories—and their relevant subcategories—in turn.¹ In these descriptions, we mention examples from our data set which we have edited for readability by removing speech disfluencies.

- *What* is rejected in the target utterance:
 - (Some of) its content.
 - (Some of) its implicatures.
 - (Some of) its preconditions.
- *How* the rejection obtains its rejecting force.
 - By having content that is contrary to what is rejected.
 - By conversationally implicating content that is contrary to what is rejected.

¹Geurts (1998) already notes out that one can negatively respond to contents, implicatures, preconditions and metalinguistic content. His analysis of what he calls *denials* is however restricted to uses of the word *not*, whereas we consider a broader variety in *how* one can negatively respond.

- By conventional implicature.
- By expressing disbelief.
- By irony.

In earlier work (Schlöder and Fernández, 2015b), we identified the additional theoretical option of rejecting *by* having a presupposition that is contrary to what is rejected, e.g. as in the constructed example (6), where *Frank stopped smoking* presupposes *Frank used to smoke*, which contradicts the content of (6a).²

- (6) a. A: Frank never smoked.
 b. B: He stopped before you met him.

We did not, however, find any example of such a rejection move in our initial sampling or our annotation study. Similarly, it may be theoretically useful to separate rejections of conversational implicatures from rejections of conventional implicatures, but we did not find any examples of the latter in our data.

3.1 What

Content. We identify a rejection move as rejecting the *content* of its target if one interprets it as rejecting the semantic (as opposed to pragmatically enriched) contribution of the target. The principal members of this category (a) use propositional anaphora to select the content of the rejection target, as in (7), or (b) repeat the target content with an inserted or removed negation, as in (8) and (9), respectively.

- (7) a. A: We can’t make a docking station anyway.
 b. D: That’s not true.
 what: content, how: contradiction

- (8) a. B: It’s a fat cat.
 b. C: It is not a fat cat.
 what: content, how: contradiction

- (9) a. B: No, not everything.
 b. C: Yeah, everything.
 what: content, how: contradiction

²More generally, one may say that an utterance can add multiple *discourse units* to the discourse (what is asserted and what is presupposed may be treated as different units) and that a rejection can attach to any such unit by different discourse relations (Lascarides and Asher, 2009). Categorising rejections by *what* they attach to by *which relation* may make up a more fine-grained taxonomy of the *what* and *how* of rejection. We thank an anonymous reviewer for observing this.

Implicated content. We identify a rejection move as rejecting an *implicature* of its target if one interprets it as rejecting part of the pragmatic content of the target. For example, in (10), *A* does not explicitly assert that *rubber is too soft* in (10a), but *B* takes *A* to implicate this and rejects it.

- (10) a. *A*: Rubber is kind of soft.
 b. *B*: Yeah, but not too soft we have decided.
what: implicature, how: contradiction

We include in this category *rejections of rhetorical questions* like (11), where *C* conveys *nobody is gonna buy a remote just for the TV unless they've lost theirs* in a rhetorical questions, which *A* rejects by asserting a contrary content.

- (11) *C*: I was like who's gonna buy a remote just for the TV unless they've lost theirs.
A: Fashionable chic people will.
what: implicature, how: contradiction

Precondition. We identify a rejection move as rejecting a *presupposition* or *precondition* if one interprets it as pointing out that some requirement for the rejection target fails. In (12), *A* does not assert that *they have not redesigned the product*, but *D* recognises this to be a precondition of *A*'s contribution and points out that it does not obtain. In (13), *A* points out that an expression in *B*'s utterance does not refer, and in (14), that a presupposition triggered by *know* fails.

- (12) *A*: So I don't think we need to redesign the product .
D: Uh that's what we've just done .
what: precondition, how: contradiction
- (13) a. *B*: you just rub on the cover, so you rub on the painting.
 b. *A*: No no, there's no painting
what: precondition, how: contradiction
- (14) a. *B*: I didn't know there was such a thing.
 b. *A*: No, there isn't.
what: precondition, how: contradiction

We include in this category rejections that challenge the felicity condition of their rejection target. For example, it seems to be the case that knowledge is required for felicitous assertion (Williamson, 2000). In (15), *C* challenges *A*'s assertion on the grounds of this condition.

- (15) *A*: but we did we didn't get that.
C: You don't know that.
what: precondition, how: contradiction

One important felicity condition is that a contribution must be *relevant* or *on topic* (Asher and Lascarides, 2003; Roberts, 2012). In (16), *A* rejects *C*'s utterance for being off topic.

- (16) a. *C*: Yes, two, but only when you compare it with elderly.
 b. *A*: Uh, that is not the question.
what: precondition, how: contradiction

3.2 How

Propositional content. We identify a rejection move as rejecting by *contradiction* if the semantic content of the rejection is incompatible with what it rejects. There are two principal options: (i) By making a claim that is incompatible with what is rejected, as in (17) or (2); (ii) by asserting the falsity of what is rejected, as in (7).

- (17) *C*: And they're double curved .
A: Single-curved .
what: content, how: contradiction

Conversational implicature. We identify a rejection move as rejecting by *conversational implicature* if its semantic content is compatible with what it rejects, but implicates something that is incompatible.³

- (18) a. *C*: This is a very interesting design .
 b. *D*: It's just the same as normal .
what: content, how: conversational impl.

Prima facie, something can be both *normal* and *interesting*, so the content of (18b) does not outright contradict the content of (18a). However, in this context, (18b) can be read as a rejection move by pragmatically enriching it with the scalar implicature that *normal* \sim *not interesting*.

For the purposes of this study, we do not wish to commit to any particular theory of conversational implicature. Therefore, we include as *rejections by conversational implicature* also the following special cases that, depending on one's preferred theory, may or may not be classified differently.

We include in this category those rejection moves that point out *counterevidence* to what is rejected, i.e. information that is not outright contradictory, but entails that what is rejected is unlikely (Asher and Lascarides, 2003). One example is (19). While *not having the slogan* and *the slogan*

³On some theories of implicature, this would not be possible as the prior context would be considered as cancelling the contradictory implicature See Walker, 1996, Schlöder and Fernández, 2015b for discussion on how to resolve this.

being obvious are not contraries, the latter constitutes counterevidence to the former.

- (19) a. B: We don't have the slogan though.
 b. A: slogan is quite obvious.
what: content, how: conversational impl.

A special kind of counterevidence are unwelcome consequences of a proposed course of action. In (20), *D* rejects a proposal by *A* by pointing out a drawback that would follow from implementing *A*'s suggestion. However, since one may follow *A*'s suggestion and accept the drawback, the content of (20b) is not contrary to (20a), which is why we categorise this as a rejection by implicature.

- (20) a. A: Drop the special colour.
 b. D: Well. That would make it less appealing. So that's no option.
what: content, how: conversational impl.

We include as rejections by implicature also utterances that express a *negative evaluation* of the rejection target, as in (21) where *A* negatively evaluates *D*'s proposal using the negative sentiment term *weird*.

- (21) D: but not important is the channel selection,
 A: That's a little weird.
what: content, how: conversational impl.

Note that one can express negative evaluations by using vocabulary that expresses a *positive* sentiment, if the rejection target has negative polarity. In (22), *A* uses the positive term *better* to reject *C*'s proposal to *use a ball instead of a wheel*.

- (22) a. C: not a wheel but a ball,
 b. A: No, a wheel is better.
what: content, how: conversational impl.

Finally, we also include rejections that make their point using a rhetorical question, like (23).

- (23) a. A: with some kind of cutting edge battery technology
 b. D: For twelve Euros?
what: content, how: conversational impl.

In (23), *D* rejects the proposal to use *cutting edge battery technology* by using a rhetorical question that implicates that this is impossible to achieve *for twelve Euros* (which is determinable from context to be a constraint on the task *A* and *D* are working on).

Conventional implicature. We identify a rejection move as rejecting *by conventional implicature* if it uses an idiomatic fixed phrase to express rejection, as in the following examples:

- (24) a. A: we should get him to do that.
 b. B: I disagree.
what: content, how: conventional impl.
 (25) a. D: That's stupid.
 b. B: We'll see.
what: content, how: conventional impl.
 (26) a. D: Look at it. That is a piece of work.
 b. C: You're kidding.
what: content, how: conventional impl.

Expression of disbelief. We identify a rejection move as rejecting *by expressing disbelief* if it expresses that the speaker does not believe what is rejected (without having content that is outright incompatible with the target). First, one may directly state *I don't know* (27) or *I'm not sure* (28).

- (27) a. A: maybe I can learn something.
 b. B: Well, I don't know how much you can learn.
what: content, how: expr. disbelief
 (28) a. B: but then you buy a new cover.
 b. A: I'm not sure if it's the it's the entire cover you change.
what: content, how: expr. disbelief

Second, we include a rejection move in this category when it expresses hesitation to accept the rejection target. One example are *Why*-questions as in (29); another are hedging phrases like *maybe not* (30) or *I guess* (31).⁴

- (29) A: Yeah, or just different colours would be uh I don't know if people also wanna spend more money on fronts for their uh remote control.
 B: Why not?
what: implicature, how: expr. disbelief
 (30) a. A: I need to get started on that.
 b. B: Well, maybe not.
what: content, how: expr. disbelief
 (31) a. A: that's not the first question.
 b. B: well - well i guess.
what: content, how: expr. disbelief

⁴Such expressions of disbelief have also been called *resistance moves* by Bledin and Rawlins (2016), as a category separate from rejection. However, according to our theoretical framework—where rejecting force means non-acceptance into common ground—resistance moves are just a special kind of rejection.

Irony. Finally, we identify a rejection move as rejecting *by irony* if it would be read as an acceptance move, save for the fact that it is best read ironically (e.g. because it is exaggerated). Two vivid examples are (32) and (33).

- (32) a. C: I want gold plating.
b. D: Yeah right.
what: content, how: irony
- (33) a. C: it's a normal colour,
b. A: Yellow rubber.
c. A: Yeah, normal.
what: content, how: irony

4 Corpus Study

4.1 Data

We collected all utterances from the AMI Corpus (Carletta, 2007), the ICSI Corpus (Janin et al., 2003) and the Switchboard Corpus (Godfrey et al., 1992) that are annotated as rejection moves. ICSI and Switchboard follow the DAMSL definition of rejection moves (Core and Allen, 1997), whereas AMI uses an idiosyncratic scheme for dialogue acts. In particular, the AMI scheme annotates some adjacency pairs as the second part being an *objection or negative assessment* of the first part, which we take to contain the class of rejections. In total, we found 929 such utterances (697 from AMI, 157 from Switchboard, and 75 from ICSI) from which we selected a random sample of 400 to annotate (317 AMI, 63 Switchboard, 20 ICSI).

However, not all these data correspond to our theoretical definition from Section 2. In case of the AMI corpus, there is a systematic reason: the class *objection or negative assessment* also contains adjacency pairs like (34b)–(34c).

- (34) a. B: Are you left-handed?
b. C: No.
c. B: Oh, pity.

Clearly, (34c) does not cancel any essential effect of (34b): the latter utterance is an answer to the question in (34a) and its essential effect—that the answer *C is not left handed* becomes common ground—is achieved. We therefore instructed our annotators to not take for granted that any item in the data set is a rejection and mark any cases that do not fit the theoretical definition.

To facilitate annotation, we displayed to the annotators the rejecting utterance and its rejection target within context. Specifically, we displayed

dimension	initial set	after refinement
what	0.35	0.68
how	0.56	0.76

Table 1: Inter-annotator agreement (Cohen’s κ) before ($n = 99$) and after ($n = 50$) refinement of the annotation manual.

the full turn⁵ containing the rejection target, the full turn containing the rejecting utterance and any other utterances in between these turns. For the AMI and ICSI corpora we also added the two utterances preceding the rejection target and the two utterances succeeding the rejecting utterance.

4.2 Annotation procedure

The data was annotated with the two-dimensional taxonomy outlined in Section 3 by two expert annotators who are versed in the theoretical background given in Section 2.

This is a difficult annotation task, in particular in cases where *what* is rejected is not “content” and simultaneously *how* it is rejected is not “contradiction”. For example, both annotators agreed that in the following example, *C* uses an implicature to reject an implicature of *B*’s utterance.

- (35) a. B: I don’t see why we should use the flipping mechanism.
b. C: I thought it would be cool.
what: implicature, how: conversational impl.

The interpretation of (35) is that *B* implicates, by way of an embedded question, that *they should not use the flipping mechanism*, which is what is rejected by *C*’s utterance in that (35b) positively evaluates that *they should use the flipping mechanism*. Although the annotators were provided with much more context than we display here, this interpretation requires careful and complex reasoning that would be difficult to achieve with naive or crowd-sourced annotators.

We pursued the following strategy. The annotators were first given a shared set of 99 items. They then compared their disagreement, agreed on a gold standard on that set, and proposed refinements to the annotation manual that follow the

⁵In the Switchboard corpus, we use the preexisting segmentation into turns. In the AMI and ICSI corpora we define the turn an utterance *u* is contained in to be the maximum sequence of utterances by the same speaker that contains *u* and that is only interrupted by other speakers with backchannel and fragmented/aborted contributions (where the classification of an utterance as backchannel and fragment follows the preexisting annotation in these corpora).

gold standard. To track the progress made by this refinement, they then annotated another shared set of 50 items. Their inter-annotator agreement (measured in Cohen’s κ , [Cohen, 1960](#)) before and after the refinement is displayed in Table 1. The inter-annotator agreement after refinement is substantial given the complexity of this task. The remaining 251 items were then annotated by a single annotator using the refined manual.

One result of this intermediate step is that sometimes even substantial context was insufficient to determine the nature of an utterance that was annotated as a rejection. For example, the annotators agreed that example (36) is of this kind; we display this example here with the full context available to the annotators with the rejecting utterance and the rejection target, as previously annotated in the ICSI corpus, in italics.

- (36) a. A: right?
 b. A: i mean you scan - i mean if you have a display of the waveform.
 c. B: *oh you’re talking about visually.*
 d. C: yeah.
 e. B: i just don’t think ==
 f. C: *w- - well | the other problem is the breaths.*
 g. C: cuz you also see the breaths on the waveform.
 h. C: i’ve - i’ve looked at the int- - uh - s- - i’ve tried to do that with a single channel.
 i. C: and - and you do see all sorts of other stuff besides just the voice.

One may read (36c) as *B* offering an interpretation of what *A* is suggesting and (36f) as implicating counterevidence to that interpretation. But other readings are possible, e.g. that (36f) points out a problem that neither *A* nor *B* have identified.

We instructed the annotators to mark such cases—where one needs to speculate about what *might* be meant, due to the absence of a clearer interpretation—as *insufficient context*. In total, 48 utterances from the 400 selected for annotation were annotated as either being determinately not a rejection (like (34)) or being unclear (like (36)).

4.3 Results

The results of the annotation are displayed in Table 2. Perhaps not unexpectedly, the vast majority of rejecting utterances are interpreted as rejecting the content of their target. Additionally, the majority of rejections are outright rejections by contradic-

how	what		
	content	impl.	precon.
contradiction	142	14	17
convers. impl.	111	34	0
convent. impl.	5	0	0
disbelief	26	0	0
irony	3	0	0

Table 2: Distribution of rejection types.

how	what		
	content	impl.	precon.
contradiction	.85	.86	.82
convers. impl.	.49	.47	-
convent. impl.	0	-	-
disbelief	.65	-	-
irony	.33	-	-

Table 3: % of polarity particles in rejections.

tion (most of them using a polarity particle like *no*, see below). This seems to be somewhat in tension with politeness theory ([Brown and Levinson, 1978](#)) that predicts that indirect ways of expressing disagreement are preferred.

Rejections of implicatures and of preconditions have previously been noted to be rather rare ([Walker, 1996](#); [Schlöder and Fernández, 2015b](#)). We did, however, find enough of them to make some noteworthy observations.

All rejections of preconditions we found are rejected by outright contradiction. This matches the theoretical claim that utterances that respond to a presuppositions (or not-at-issue content in general) are highly marked and that one needs to be explicit when responding to them ([Geurts, 1998](#); [Tonhauser, 2012](#)). Moreover, although there were only a few items annotated as rejections by conventional implicature, expression of disbelief, or irony, these were all annotated as rejecting content. It stands to reason that a conventional implicature rejection also conventionally is about the content of its rejection target. But it is unexpected that there is no expression of disbelief about non-explicit content; we cannot think of a theoretical reason for this. Finally, that rejections by irony only occur as rejecting content in our dataset may be simply due to the sparsity of ironic utterances.

To gain some insight on the use of polarity particles, we computed how many rejecting utterances in each category contain a polarity particle (i.e. one of *yes*, *no* or the more informal vari-

ants *nope, yeah, nah, nee, nay, yea*). These results are displayed in Table 3. Interestingly, while polarity particles seem to appear somewhat more commonly with utterances that contradict outright (many of these are just bare *no*), they do appear fairly frequently in rejections that reject by conversational implicature and by expression of disbelief as well. This confirms an empirical claim made by Dickie (2010) and Incurvati and Schlöder (2017) that *no* does not always express that the rejection target expressed a falsity.

That no conventional implicature occurred with a polarity particle, however, seems to be an artifact of the sparsity of conventional implicatures, as *No, I disagree* seems intuitively possible as a rejection move. The single rejection by irony that contains a polarity particle is *Yeah right* from example (32); (33) was not counted here because the polarity particle only occurs in the utterance that follows the one annotated as a rejection (per the existing annotation in the AMI corpus).

4.4 Interesting cases

In our annotated data, we find some rejections that deserve more fine-grained attention than captured by our annotation scheme. We close our analysis by discussing two such cases in depth.

First, it seems that rejections of rhetorical questions take the form of an answer to the question interpreted non-rhetorically, as in the example (11) from Section 3. One may be inclined to conclude that rhetorical questions are only interpreted as making claims when they are not rejected. This would complicate the theoretical analysis of such rhetorical questions (see, e.g., Biezma and Rawlins, 2017). However, we found one rhetorical question in our data that is rejected by an utterance that does not have the form of an answer.

- (37) a. B: How many people would notice that, though?
 b. A: But they'll notice it after like a year, what: implicature, how: conversational impl

The analysis of this example is rather complex. The rhetorical question (37a) is interpreted as the claim that *few people would notice that*, which in turn implicates that *that does not matter*. The speaker of (37b) seems to grant that *few people would notice that*, but rejects that *that does not matter* by providing counterevidence (*they'll notice, hence it does matter*), making (37b) the rejection by implicature of an implicature of (37).

So, it would be incorrect to conclude that rhetorical questions are rejected by answering them as questions. However, it may still be the case that one answers a rhetorical question (i.e. treats it as a genuine question) to reject its core proposition (which the rhetorical question is interpreted to assert). We do not have enough rhetorical questions in our data to settle this matter definitively.

Second, utterances like (38b) seem to offer *refinements* of a previous utterance.

- (38) a. A: um - even though there is probably no train from here to new york.
 b. B not direct.

The interpretation of (38) seems to be this. The utterance (38a) is ambiguous between the claim that *there is no direct train from here to NY* and *there is no train at all from here to NY*. B makes clear that she is only willing to agree to the former.

The preexisting annotation of the ICSI corpus identifies (38b) as a rejection of (38a). It is not clear whether (38b) counts as a rejection in the sense of our definition from Section 2. It is oftentimes incorrect to say that an utterance makes a single, unambiguous proposal to update common ground—rather, what precisely is proposed is the subject of a collaborative negotiation process (Clark, 1996). By specifying which possible proposal she is willing to accept, B seems to be contributing to this process, but not to be rejecting any proposal; unless, that is, we count the exclusion of one possible proposal as such a rejection. Clearly, we conclude, our theoretical picture is still too coarse to fully capture how speakers negotiate what becomes common ground. For now, we have annotated (38) as not being a rejection move.

5 Conclusion

We have presented a fine-grained taxonomy for categorising rejection moves that is both theoretically motivated and driven by actual dialogue data. We classified rejections along two dimensions—*what* aspect of the target utterance is being rejected and *how* the rejection is realised—and used this scheme to annotate rejection moves from three different dialogue corpora: AMI, ICSI, and Switchboard. We expect the taxonomy and the annotated dataset to be a useful resource for further studies on the linguistic strategies available to express rejection in English conversation.⁶

⁶Data available at: https://uvaauas.figshare.com/articles/Taxonomy_of_Rejection/8870615

Acknowledgements

This work has received funding from the Netherlands Organisation for Scientific Research (NWO) under VIDI grant no. 276-89-008, *Asymmetry in Conversation* and from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 758540) within the project *From the Expression of Disagreement to New Foundations for Expressivist Semantics*.

References

- Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.
- Nicholas Asher and Alex Lascarides. 2008. Commitments, beliefs and intentions in dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue*, pages 29–36.
- John L. Austin. 1962. *How to do Things with Words. The William James lectures delivered at Harvard University in 1955*. Clarendon Press.
- Maria Biezma and Kyle Rawlins. 2017. Rhetorical questions: Severing asking from questioning. In *Semantics and Linguistic Theory*, volume 27, pages 302–322.
- Justin Bledin and Kyle Rawlins. 2016. Epistemic resistance moves. In *Proceedings of Semantics and Linguistic Theory 26*, volume 26, pages 620–640.
- Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.
- Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.
- Herbert H. Clark. 1996. *Using language*. Cambridge University Press.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1(20):37–46.
- Mark Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, pages 28–35. Boston, MA.
- Imogen Dickie. 2010. Negation, anti-realism, and the denial defence. *Philosophical Studies*, 150(2):161–185.
- Donka F Farkas and Kim B Bruce. 2010. On reacting to assertions and polar questions. *Journal of semantics*, 27(1):81–118.
- Gottlob Frege. 1919. Die Verneinung: Eine logische Untersuchung. *Beiträge zur Philosophie des deutschen Idealismus*, 1:143–157.
- Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL’04)*, pages 669–676.
- Bart Geurts. 1998. The mechanisms of denial. *Language*, 74(2):274–307.
- Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.
- Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–365.
- John J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. *IEEE Conference on Acoustics, Speech, and Signal Processing*, 1:517–520.
- Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.
- Luca Incurvati and Julian J Schlöder. 2017. Weak rejection. *Australasian Journal of Philosophy*, 95(4):741–760.
- Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elisabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of ICASSP’03*, pages 364–367.
- Justin Khoo. 2015. Modal disagreements. *Inquiry*, 58(5):511–534.
- Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2):109–158.
- Amita Misra and Marilyn Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France. Association for Computational Linguistics.
- Sarah E Murray. 2009. A hamblin semantics for evidentials. In *Semantics and linguistic theory*, volume 19, pages 324–341.
- Massimo Poesio and David Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

- Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London.
- Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.
- Floris Roelofsen and Donka F Farkas. 2015. Polarity particle responses as a window onto the interpretation of questions and assertions. *Language*, 91(2):359–414.
- Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, 53:361–382.
- David Schlagen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*.
- Julian Schlöder and Raquel Fernández. 2014. The role of polarity in inferring acceptance and rejection in dialogue. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 151–160.
- Julian J Schlöder and Raquel Fernández. 2015a. Clarifying intentions in dialogue: A corpus study. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 46–51.
- Julian J Schlöder and Raquel Fernández. 2015b. Pragmatic rejection. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 250–260.
- Julian J Schlöder, Antoine Venant, and Nicholas Asher. 2018. Aligning intentions: Acceptance and rejection in dialogue. In *Proceedings of Sinn und Bedeutung 21*, pages 1073–1089.
- Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics (Syntax and Semantics 9)*. Academic Press.
- Judith Tonhauser. 2012. Diagnosing (not-) at-issue content. *Proceedings of Semantics of Underrepresented Languages of the Americas (SULA)*, 6:239–254.
- Marilyn A. Walker. 1996. Inferring acceptance and rejection in dialogue by default rules of inference. *Language and Speech*, 39(2-3):265–304.
- Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of ACL*, pages 374–378.
- Timothy Williamson. 2000. *Knowledge and its Limits*. Oxford University Press.

The Status of Main Point Complement Clauses

Mandy Simons

Department of Philosophy
Carnegie Mellon University
mandysimons@cmu.edu

Abstract

This paper provides support for the analysis of clausal complement sentences as consisting of two discourse units, defending the view against an alternative according to which the embedded content is communicated as a conversational implicature. The argument is based on two MTurk studies of the availability of embedded content for conversational continuations. Further consequences of these findings for modeling discourse are considered in the concluding sections.

1 Uses of Clausal Complement Sentences

What I will here call a *clausal complement sentence* (cl-comp) is any sentence whose main predicate takes as complement a full tensed clause, such as the sentences in 1:

1. Jane thinks / heard / said / is glad that it's raining.

Sentences of this form have an interesting property (one which they share with other sentences with embedded finite clauses): they express two distinct propositional contents. The *matrix content* is the content of the sentence as a whole, typically an evidential, reportative or attitude claim. The *embedded content* is the content of the complement clause.¹

Several researchers from different traditions have observed that these sentences can be used in two different ways (Urmson 1952, Hooper 1975, Simons 2007, Hunter 2016). In one use, the *matrix main point use* (MMPU), the matrix content is what, informally speaking, we would call the main

point of the utterance. In the other use, the *embedded main point use* (EMPU), the embedded content is the main point, while the matrix content serves some kind of secondary discourse function, often evidential. These two uses can easily be illustrated in Q/A pairs (cf. Simons 2007):²

2. A: What did Jane say?
B: She said that it's raining.
3. A: What's the weather like?
B: Jane said that it's raining.

In 2., the matrix content is the answer to the question, so this is an MMPU. But in 3., the answer is expressed by the embedded content. We naturally understand speaker B as intending to provide that answer – that it is raining – but also to be indicating the source of her information. This is an EMPU.

In this paper, we explore the following question: What is the status of the content that we identify as “main point content” in embedded main point uses of clausal complement sentences? In particular, we will try to adjudicate between two positions on this question, both of which are articulated in prior work. The first is that the embedded clause is an independent discourse unit, which, despite syntactic embedding, makes an independent contribution to discourse content (Hunter 2016). The competing position is that EMPUs involve a conversational implicature which happens to be similar or identical in content to the content of the complement clause (Simons 2007).

Now, if the latter position is correct, we would expect the main point in EMPUs to behave in similar ways to other types of implicature, such as Relevance implicature. One of the central features

¹ When the complement clause contains an expression bound in the matrix, as in *Every linguist thinks they have the most interesting data*, the cl-comp does not express an independent proposition. As far as I can determine, these cannot have the embedded main point uses that are the focus of this paper.

² The examples in this paper are all constructed by the author. See Hunter 2016 for a slew of naturally occurring examples of EMPUs, although restricted to reportatives; and Simons 2007 for additional naturally occurring cases.

of EMPUs is that the embedded content becomes highly available for uptake in conversational continuations, as in 4., where C responds to B with a denial of the content of the embedded clause.

4. A: What's the weather like?
 B: Jane said that it's raining.
 C: But it's not, I can see the sunshine.

To evaluate the proposal, we will explore the degree to which this feature differentiates main-point embedded content from Relevance implicatures. As we will see, embedded content in fact seems to behave differently; and this behavior is not even restricted to EMPU cases.

In the next section, I will explain in more detail the two positions on the status of embedded content. In section 3, I'll evaluate the implicature proposal, presenting results from two MTurk elicitation experiments, concluding that the data support a slightly modified version of the Hunter analysis. In sections 4 and 5, I will briefly discuss two important distinctions that the data reveal: the distinction between rhetorical structure and the intentional structure of a discourse, which reflects the commitments of speakers to propositions (section 4); and the distinction between main point status of a proposition, and the simple fact of a proposition having been expressed, which, as we will see, has a significant impact on its discourse status (section 5).

2 Two approaches to EMPUs

2.1 Hunter 2016: Embedded clause as independent discourse unit

Hunter, summarizing approaches to EMP uses of cl-comp sentences, says:³

"The treatment of discourse parenthetical reports in the [Penn Discourse Tree Bank], the [Copenhagen Dependency Tree Bank] and Hunter et al. 2006 [SDRT] all have in common the idea that **discourse parenthetical reports are best modeled by attaching the embedded clause directly to the incoming discourse and that this attachment pattern distinguishes them from**

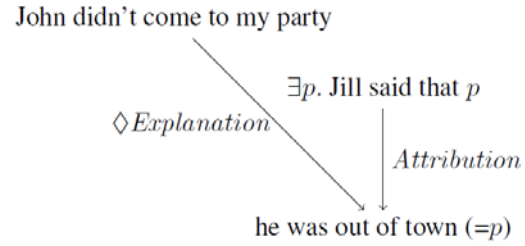


Figure 1: Illustration of Hunter 2016 analysis

non-parenthetical reports, in which it is the attribution predicate that is attached to the incoming discourse." (7, emphasis added).

In other words, on this view, the embedded clause is treated as making its own, independent contribution to the discourse structure. Hunter's 2016 analysis continues this approach.

For our purposes, Hunter's analysis involves two central claims.⁴ First, all cl-comp sentences are segmented into two discourse units: the contribution of the "attribution predicate" itself; and the contribution of the embedded clause. (See Figure 1.) Second, each of these units can participate independently in rhetorical structure. In EMPUs, the embedded content stands in some rhetorical relation to a previously introduced discourse unit, while in MMPUs, only the attribution clause is related to prior discourse. In both uses, the embedded content is obligatorily related to the attribution predicate via the *attribution* relation.

2.2 Simons 2007: Main point as implicature

Simons 2007 suggests that EMP interpretations arise through Gricean conversational reasoning.⁵ For example 3. above, the following sort of reasoning is suggested: *Information about what Jane said does not directly answer the question; but I expect B's contribution to be a cooperative response; the content of what Jane said would, if true, be an answer to my question; so plausibly B intends me to consider this reported content as an answer.* Further conversational reasoning can lead

³ Hunter limits her discussion to cl-comp sentences whose main predicate is a reportative, focussing on the issue of parenthetical reports. I assume here that Hunter's analysis can be generally extended to all cl-comp sentence. Hunter herself does not make this claim.

⁴ Hunter's analysis also involves substantive semantic claims; discussion of these is outside the scope of the current paper.

⁵ In that paper, I suggested *both* that the embedded clause has an independent discourse function, and also provided a Gricean account of how the EMP interpretation arises. In later work, (Simons 2013, 2016) I continued the argument that embedded clauses can make an independent contribution to discourse. I subsequently realized that in the 2007 paper, I had failed to establish a clear position on the status of the main point content in EMPU cases. This work is an attempt to remedy that situation.

to effects of hedging or of strengthening. For example, if Jane is a very reliable source of information, identifying her as the source might be a way for the speaker to enhance the reliability of the content.

The central claims of this analysis, which distinguish it from Hunter's, are these: First, the content of the matrix clause is asserted, and no other discourse contributions are directly made. The utterance implicates that the speaker has the more complex conversational intention described above, resulting in an implicature whose content is closely related to that of the embedded clause. In cases like 3. above, the implicated content is plausibly identical to that of the embedded clause. But as just noted, EMPUs often involve a degree of *hedging* of the main point content. Answering in 3. with *Jane thinks that it's raining* would, on the Simons 2007 view, generate a relatively weak implicature, along the lines of the modal *It's possible that it's raining*.⁶

3 Adjudicating between the approaches

The two approaches just outlined differ in their predictions in testable ways. First, if the Simons 2007 implicature analysis is correct, then we would expect other main-point implicatures to behave in relevant respects like the embedded content of cl-comp sentences in EMP uses. To illustrate a case of a main point implicature, consider example 5.:

5. A: Is Helen in her office?
B: The light's on.

The structure of this question/answer sequence is parallel to that of 3. above. B's utterance does not directly answer A's question; but assuming that B intends to be cooperative, A can conclude that B intends her to consider the light being on as evidence that Helen is in her office (just as the fact that someone *said* that Helen is in her office would provide such evidence). The implied answer, then, is that Helen (probably) is in her office. If the implicature analysis of how the main point of EMPUs arises is correct, then the implied answers in these two cases should have similar properties.

The second point of difference concerns the question of when the embedded content should be accessible for conversational uptake. On Hunter's view, cl-comp sentences *always* make available two distinct discourse units, regardless of whether

they are used in an EMPU or an MMPU. On the implicature view, in contrast, embedded content becomes independently available only in EMPU cases. When the matrix content coheres fully with the prior discourse, as in MMPU cases, no implicature is generated, and hence the embedded content should simply remain embedded: it is not present as an independent discourse contribution. In the next sections, I discuss the results of two small scale Mechanical Turk studies which provide evidence in favor of Hunter's analysis.

3.1 Embedded main point vs. standard implicature: conversational uptake

As noted above, the embedded content in a cl-comp sentence not only determines the relevance of the utterance to the prior discourse, but can also be the target of conversational continuations: an interlocutor can respond directly to the embedded content, as illustrated in 4. above and 6. below.

6. A: What's the weather going to be like?
B: Jane thinks it's going to rain.
A: I'd better wear my raincoat then.

If the implicature analysis of EMPUs is correct, then we should expect that dialogues like 5. should also allow conversational uptake of the implicated main point content. And at first pass, it appears that it can. Speaker A might respond: *Good, because I have this form I need her to sign*, a response to the information that Helen is in her office, and not to the light being on.

To explore this issue more carefully, I conducted a small scale study on Amazon Mechanical Turk. Participants saw text of a sequence of three-segment dialogs; Examples 1 and 2 from the experiment are shown in 7-8:

7. A: Will Henry be here for the start of the meeting?
B: [Emb-Cond] Jane said that he won't be
B: [Imp-Cond] He missed the bus.
C: Yes that's right / I'm surprised.
8. A: Is Lili coming to the movie?
B: [Emb-Cond] Jeff said she's not coming out tonight
B: [Imp-Cond] She's working.
C: That's too bad.

⁶ Hunter deals with hedging by positing modalized rhetorical relations, as in Fig. 1.

Embedded Condition					
	Ex1	Ex2	Ex3	Ex4	Ex5
IE	18	19	2	16	20
LM	0	0	14	2	0

Implicature Condition					
	Ex1	Ex2	Ex3	Ex4	Ex5
IE	1	4	0	1	9
LM	18	7	16	18	10

Table 1: Counts per condition by example. Numbers may not sum to 20 due to uncodable items.



Figure 2: Relative proportion of responses per condition, by example

The A utterance in each case is a yes/no question. The B utterance is either a cl-comp sentence whose embedded content directly answers the question (Embedded Condition), or an atomic sentence from which an answer is inferable (Implicature Condition). The C utterance in each case is one of *yes*, *that's right*, *I'm surprised* or *That's too bad*. These responses are anaphoric, interpretable either as referring to the matrix content / literal meaning, or to the embedded content / implicature. (Each of the participants saw one version of each question+responses sequence, plus at least one filler, used to check for competence in the task.) A total of 20 responses was collected for each dialog in each condition.⁷

Immediately below the dialog, participants were given a write-in box, and the prompt: “Write in the box below what Cate [name used for C] agrees with / finds surprising / thinks is too bad”. Responses were then hand-coded by the author for whether the participant understood the C utterance as referring to the matrix content or to the embedded content (in the Embedded Condition) or as referring to the literal content or to the implied

content (in the Implicature Condition). Answers not clearly falling into one of these categories were treated as uncodable. In the Implicature Condition, some answers mentioned both the literal content and the implicature (e.g., in response to Ex.2 shown in 8 above: “Cate thinks it's too bad that Lili has to work and will miss the movie.”) These were coded as *both*, but excluded from the data.

If it is correct that in EMP uses of cl-complement sentences, the main point is conveyed as an implicature, then, in this experiment, responses to the Embedded Condition and the Implicature Condition should not differ: participants should be just as likely to select the implicature as the target of a response in the Implicature Condition as they are to select the embedded content as the target in the Embedded Condition. For purposes of analysis, implicature responses in the Implicature Condition and embedded clause responses in the Embedded Condition were identified as a single value, IE. Similarly, literal meaning responses in the Implicature Condition and main clause responses in the Embedded Condition were identified as a single value, LM.

⁷ There were a total of 5 dialogs. These were run in two separate iterations of the experiment. The first iteration,

conducted in February 2018, used Exs 4 and 5. The second iteration, conducted in May 2019, used Exs 1-3.

Table 1 above shows the raw counts of each response type per condition, by example. In the Embedded Condition, a total of 9 responses were uncodable. In the Implicature Condition, 8 were uncodable. An additional 8 were coded as *both* (as explained above) and excluded from the data.

Figure 2 is a mosaic plot that shows the relative proportion of responses of each type per condition, by example. Width of the bars reflects the number of coded examples; note that Ex.2 is particularly narrow. This is due to the fact that 7 of the 8 *both* responses were elicited by this example.

In order to test independence of condition (Embedded vs. Implicature) from the understood target of the C utterance (IE vs. LM), a Chi-square test of independence was performed.⁸ The relation between these variables was significant ($\chi^2(1) = 70.322$, $p < 0.0001$), showing them to be highly correlated.

Although clearly all contents in almost all examples are construable as the antecedent of the C utterance, there is a robust difference between the Embedded Condition and the Implicature Condition. Overall, as shown by the statistical test, the two conditions give rise to clearly distinct patterns of response. A closer look at the data shows that in all but one example, there is a strong preference to treat the embedded content as the antecedent in the Embedded Condition.⁹ In contrast, in the Implicature Condition, the literal content is far more likely to be chosen as antecedent than the implicated content. In summary: an implicature, even when it is the main point of an utterance, is less available for conversational uptake than the embedded clause content of a cl-complement sentence, when that embedded content constitutes the main point.

These results provide preliminary support for the claim of a difference in status between the embedded content of cl-comp sentences, and an implicature (or invited inference). And there is in fact a very plausible explanation of that difference, namely, that the embedded content of cl-comp sentences is explicitly expressed. This, then, leads to a further question: is explicit expression enough

to allow for conversational follow-up? We turn to this in the next section.

3.2 Embedded content in EMPU vs. MMPU

Recall that if the availability of embedded content is understood to be due to an implicature, this predicts that when no implicature is warranted – that is, when the matrix clause is directly relevant to the prior discourse, as in MMPUs – the embedded content should not be available for conversational continuation. In contrast, on the view that the embedded clause by default constitutes an independent discourse unit, no contrast is predicted between MMPUs and EMPUs in this respect.

This issue was tested in a further experiment. In this study, participants were shown a four-segment dialog with an anaphoric final segment.¹⁰ In this case, the target sentence (response to the question) was always a cl-comp sentence. In contrast to Experiment 1, these cl-comp sentences involved an MMPU: in each case, the matrix content, not the embedded content, was a plausible direct answer to the question preceding it. (Recall that in the embedded condition in Experiment 1, it was always the *embedded* content of the cl-comp sentence which addressed the prior question.) Dialogs 1 and 2 are shown below:

9. Dialog 1

A: Alan never ceases to amaze me.

B: Why, what did he do now?

A: He announced to everyone that he got ticketed for DUI. / that he's going into bankruptcy.

C: That's weird.

10. Dialog 2

A: I am so mad at Mike.

B: Uh oh. What happened?

A: He's going around saying that Helen's going to get fired / that there was a big security breach last night.

C: That's weird.

There were 4 different dialogs; as illustrated here, each dialog had two versions differing in the content of the clausal complement of A's second

⁸ The Chi-squared test treats each response as independent, ignoring any possible effects of subject. This variable could be explored in future work.

⁹ The outlier example (Ex.3) is as follows: A: *How was that book the kids were reading for school?* / B: *Fran thought it was really boring.*

¹⁰ The initial statement by A, preceding the question by B, was introduced to help establish the main point status of the matrix content of the target sentence (A's second utterance.) For example, in 9, the initial sentence makes clear that the main point of A's second utterance is to report on something that Alan has done that is surprising.

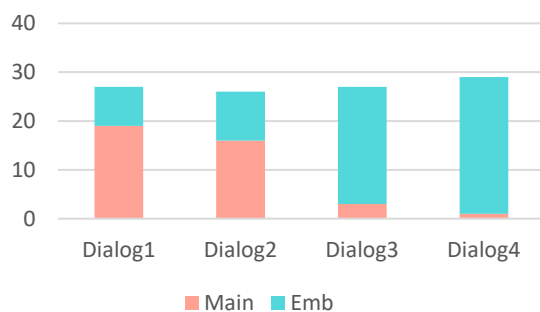


Figure 3: Exp 2, numbers of each response type (main clause reference vs. embedded clause reference), by dialog.

utterance. This was to safeguard against the possibility that one particular content might skew the effects. In fact, in no dialog was there any observable difference in distribution of responses across the two variants. As before, participants wrote in an answer to the question: “What does Cate think is weird / surprising?”; answers were coded by the author as referring to either the main clause content or the embedded clause content. 15 responses were collected for each dialog with each content (30/dialog). The results, arranged by dialog and excluding uncodable items, are shown in Figure 3.

The question under investigation here is whether the embedded content is *available* as the target of conversational continuations, when that content is not the main point in the utterance in which it is introduced. (We are not concerned here with whether that content is *preferred* as the target of conversational continuations.) The results indicate that it is so available. In dialog 1, 8 out of the 27 codable responses identified the embedded content as antecedent of the C utterance; in dialog 2, 10 out of 26 responses did so. In dialogs 3 and 4, the embedded content was the preferred understood target. These results support the view that the embedded content of cl-comp sentences constitutes an independent discourse segment even in MMPU cases, as in Hunter’s analysis. The results are also consistent with the claims of Snider 2018 that propositional contents may be available as antecedents of propositional anaphors even when not at-issue.

An important caveat is in order here, however. It is true that in the dialogs in the study, the conversational situation does not support a true Gricean implicature that the speaker intends to communicate the content of the embedded clause;

there is no conversational violation to support such an implicature. But real world knowledge may well support an ordinary inference that the embedded contents may be true or of interest. Consider Dialog 1 from the study, shown in 9. above. The answer to B’s question (*What did Alan do?*) is provided by the matrix content of A’s reply: it’s his announcing that he got ticketed that is amazing. But it is also simplest to assume that what Alan announced is true; moreover, his getting ticketed for DUI is a discussion-worthy topic in its own right. One might posit, then, that although the availability of the embedded content for a conversational continuation does not require a true Gricean implicature, it nonetheless requires inferences about how likely the embedded content is to be true or to be of conversational interest. These inferences help determine whether content will be available for a conversational uptake.

Recall, though, the results from the previous experiment, which clearly show that explicitly expressed content is significantly more likely to be the target of conversational uptake than implied content, even when the implied content is highly relevant and has more real-world significance than the explicit content. At the very least, we can conclude from the combined results of the two experiments that explicit expression of content makes that content more easily accessible for conversational uptake; and this result is better modeled by the Hunter analysis than by an analysis relying on conversational implicature or other types of conversational reasoning. Nonetheless, we should not overlook the role of pragmatic reasoning in the identification of the anaphoric antecedents in these experiments: In order for content to be the target of uptake, it must also be content that the interlocutors are likely to want to talk about.

3.3 Interim conclusions for Hunter 2016

These experimental results, although limited, provide support for the view espoused by Hunter (and others working within a rhetorical relations framework) according to which cl-comp sentences are segmented into two discourse units regardless of their discourse use, with the embedded content available as an anchor for a rhetorical relation even if it is not initially independently attached to the preceding discourse by any relation, as in the analysis below:

11. A: What did Mike say about Helen? _{α}
 B: [He said] _{β} [that she's in her office] _{γ}
 Att(β, γ), QAP(α, β)
 C: Yes that's right. _{δ}
 Affirm(γ, δ)

On the other hand, there is no response-based evidence for the presence of the “attribution” discourse-unit proposed by Hunter, with the content (roughly) “Mike said something.” (Hunter 2016, pp.17-18). There is, though, clear evidence for a discourse unit consisting of the matrix content in its entirety. This suggests a natural revision to Hunter’s analysis, according to which cl-comp sentences should be segmented into discourse units as shown in 12.:

12. [He said [that she's in her office] _{γ}] _{β}

As well as being supported by the evidence cited, this modification has the benefit of maintaining fit between the discourse units posited and the syntactic and semantic units of the sentence. This modification still allows us to posit the holding of the Attribution relation between the matrix discourse unit and the embedded discourse unit; the matrix content indeed expresses the attribution in question. We will not pursue here any further questions for Hunter’s analysis raised by this modification.

3.4 Aside on the role of the embedding predicate

While not directly relevant to the main issue, the differences between the dialogs in Experiment 2 are worth a brief comment. The embedding predicate in dialog 1 is *announced*; in dialog 2 is *going around saying*; and in dialogs 3 and 4 is *said*. Dialogs 3 and 4 are shown here:

13. Dialog 3

- A: What's going on with Jen?
 B: Nothing that I know. Why?
 A: She said she turned down that great job offer / that she isn't coming to dinner with us.
 C: I'm surprised.

14. Dialog 4

- A: I'm getting a little worried about Chris
 B: Why, what's going on?
 A: He said that Bill is avoiding him / Bill is being mean to him.
 C: I'm surprised.

While respondents overall preferred the matrix clause as antecedent in Dialogs 1 & 2 (while allowing the embedded clause as a possible

antecedent), responses to Dialogs 3 & 4 almost unanimously selected the embedded clause as antecedent. The simple reportative *say* seems to carry almost no semantic weight, and respondents seem to straightforwardly take its complement to be presented as true. The comparison with Dialog 2 (with predicate *going around saying*) is instructive: responses to “What does Cate think is surprising?” included “Cate is surprised that Mike is spreading a rumor” and “That Mike is gossiping about Helen,” both suggesting that participants did not necessarily take the content of what Mike was saying to be true.

Previous work on inferences about veridicality of events presented in texts (Sauri 2008, de Marneffe et al. 2012, de Marneffe 2012) has identified a variety of factors that contribute to these inferences. The current experiment suggests that quite fine features of the embedding predicate can have a significant effect; and also that there is a possible relation between veridicality judgments and judgments of “uptake worthiness”.

4 The role of conversational inference in interpretation of cl-comp sentences: coherence vs. commitment

The crucial distinction between EMP uses and MMP uses of cl-comp sentences lies in how the cl-comp sentence coheres with prior discourse. This is articulated both in my 2007 description of the two uses, and in Hunter’s model. Crucially, in EMP uses, the cl-comp sentence coheres with the prior discourse primarily by virtue of a rhetorical relation holding between the prior discourse and the embedded content. In MMP uses, the crucial relation is between the matrix content and the prior discourse.

But even in the EMP case, where the embedded content is one of the relata of a crucial coherence relation, the speaker need not be understood to be fully committed to that content. As we’ve noted, the speaker of a cl-comp sentence in an embedded main point use is often understood to have reduced commitment to the embedded content: this is what explains their choice to embed that content, rather than simply asserting it. In other cases, by providing a strong evidential source for the content, a speaker with full commitment to the embedded content can bolster their case for its truth e.g. *My doctor told me that it's actually ok to eat a lot of eggs.*)

The kind of reasoning described by Simons 2007 derives conclusions about both the intended main point of a cl-comp utterance, and about the speaker's degree of commitment (see section 2.2. above). Hunter concurs with Simons 2007 that the determination of speaker commitment requires "world-knowledge based reasoning." (p.11). Hunter goes on to say that "this kind of world-knowledge based reasoning...is generally independent of the reasoning used to determine rhetorical structure." Moreover, as Hunter further notes, "it can take many discourse turns to determine a speaker's commitment to the embedded content of a report."

These observations suggest a crucial distinction between two types of information: on the one hand, the rhetorical relations between elementary discourse units, modelled in rhetorical structure theories; and on the other, a higher level model of the intentional structure of a discourse, a structure which must reflect each speaker's conversational commitments. And reasoning about a speaker's likely intentions is essential to the determination of this structure.

A central case for distinguishing rhetorical structure from speaker commitment is the case of "no-commitment" uses of cl-comp sentences, as in 15. These were first discussed by Simons 2007, and taken up by Hunter 2016.

15. A: What course did Jane fail?

B: Henry falsely believes that she failed calculus. In fact, she failed swimming.

More subtle cases are possible. Consider:

16. A: So, is Trump guilty of collusion?

B: Well, Giuliani says he's completely innocent.

In these cases, the hearer is expected to infer that the speaker has no commitment to the embedded content. Because Hunter's modal rhetorical relations entail that the relata are epistemic possibilities for the speaker, her model cannot treat the embedded contents of such cases as rhetorically related to the prior discourse. She argues that in such cases, it is the attribution predicate alone which attaches to the prior discourse. In making this move, though, Hunter seems to conflate rhetorical structure with the determination of speaker commitment (a conflation which, earlier in the paper, she deems problematic; see her section 3.1.). Resolving this issue would require a significant overhaul of the semantic commitments of Hunter's analysis, which I will not undertake.

5 Concluding remarks: main point status vs. explicitness

The observations from Exp. 1, comparing cl-comp content to Relevance implicature, show an important difference between "main point" status of content, and availability for conversational continuation. Simons 2007 already noted that in both uses of cl-comp sentences, conversational contributions can target either the matrix content or the embedded content (Simons 2017, ex.16); and this is confirmed by the results of the study.

The data suggest that embedded content is available for conversational continuation even when not the speaker's intended main point. This observation is unsurprising if one considers normal, real-life talk, where interlocutors may pursue tangents or compete for topic control.

The data further point to the crucial importance of explicit expression of propositional content. That act of expression, in and of itself, makes the propositional content expressed available for conversational uptake.

These results have important consequences not only for our understanding of the discourse functions of cl-comp sentences, but for theorizing about the semantics and pragmatics of discourse in general. For example, Simons 2013 has argued that certain cases of local pragmatic effects arise through the application of conversational reasoning to the contents of non-asserted clauses (e.g. disjuncts, or the antecedent or consequent of a conditional); and that this is possible precisely because these clauses function as independent discourse units. The results reported here support that view, demonstrating that our thinking about discourse pragmatics must be attentive not only to what is implicit in discourse, but also to the function of explicit expression.

Acknowledgements

Versions of this work have been presented to the New York Philosophy of Language Workshop and to the Linguistics Department at the University of Pennsylvania. I am grateful to Judith Tonhauser for early discussion of the material and for help in setting up the experiments. Thanks to Christina Bjorndahl for statistical analysis and plots, and to three anonymous SemDial reviewers for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 0952497.

References

- Hooper, J.B. 1975. On assertive predicates. In: Kimball, J.P. (Ed.), *Syntax and Semantics Vol. 4*. Academic Press, NY: 91-124.
- Hunter, Julie. 2016. Reports in Discourse. *Dialogue & Discourse* 7(4): 1-35.
- de Marneffe, Catherine. 2012. *What's that supposed to mean?* Ph.D. dissertation, Stanford University.
- De Marneffe, Catherine, Christopher Manning and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38(2): 301-333.
- Sauri, Roser. 2008. *A Factuality Profiler for Evantualities in Text*. Ph.D. thesis, Computer Science Department, Brandeis University.
- Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117: 1034-1056.
- Simons, Mandy. 2013. Local Pragmatics and Structured Content. *Philosophical Studies* 168(1): 21-33.
- Simons, Mandy. 2016. Local Pragmatics in a Gricean Framework. *Inquiry* 60(5): 493-508.
- Snider, Todd. 2018. Distinguishing At-issueness from Anaphoric Potential: A Case Study of Appositives. In William G. Bennett, Lindsay Hracs, and Dennis Ryan Storoshenko (eds.), *Proceedings of the 35th West Coast Conference on Formal Linguistics (WCCFL)*: 374-381. Cascadilla Proceedings Project.
- Urmson, J.O. 1952. Parenthetical Verbs. *Mind, New Series* 61: 480-496.

How to Put an Elephant in the Title: Modelling humorous incongruity with enthymematic reasoning

Ellen Breitholtz and Vladislav Maraev

Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
{ellen.breitholtz,vladislav.maraev}@gu.se

Abstract

In humour theories incongruity is a crucial feature in characterising a humorous event, and giving it a formal definition is important for accounting for why something is considered amusing. In this work we address the problem of formalising incongruity within a computational framework, thereby exploring the ability of a machine to detect the source of humorous incongruity and being able to adapt its behaviour accordingly. In our formalisation we draw attention to dialogical and incremental perspectives on humour. To capture mismatches between the information states of dialogue participants, we adopt the notions of enthymemes and topoi as instances and types of defeasible argumentation items.

1 Introduction

Consider the following dialogue:

- (i)
- 1 A How do you put an elephant into a fridge?
 - 2 B Hmm, I don't know?
 - 3 A Open the door, put the elephant inside, close the door.
 - 4 B Haha okay
 - 5 A How do you put a giraffe into the fridge?
 - 6 B Open the door, put the giraffe inside, close the door?
 - 7 A Wrong! Open the door, get the elephant out, put the giraffe inside, close the door.

Jokes such as the one above rely on dialogicity and the expectations of interlocutors to reason in a certain way based on certain assumptions about acceptable reasoning. In this work we suggest an account of humorous events that calls attention

to the dialogical nature of humour, and the underlying reasoning warranting interpretations giving rise to humour.

Studies of humour often underline the importance of comprehending a temporal sequence of events for understanding a joke as it unfolds (Suls, 1972; Ritchie, 2018). However, such sequence might be interpreted differently by different interlocutors. Moreover, when telling a joke, a speaker often takes advantage of the potential to interpret a move in different ways to create a humorous exchange. Thus, to account for how a speech event is perceived as humorous, we must consider this incrementality from an interactive point of view. In our account we use techniques from dialogue semantics where game boards are used to represent the information states of interlocutors, which are updated during the course of an interaction (Cooper and Ginzburg, 2015; Ginzburg, 2012; Larsson, 2002). To capture mismatches between the information states of dialogue participants, we adopt the notions of *enthymemes* and *topoi* as instances and types of defeasible argumentation items. This approach has been used in formal analysis of dialogue to account for inferences based on background assumptions, and to account for incremental interpretation of argumentation in dialogue (Breitholtz et al., 2017). Very similar approach based on topoi and enthymemes was utilised to account for laughter-related incongruity in dialogue (Ginzburg et al., 2015; Mazzocchi et al., 2018).

In the rest of the paper we will first provide some theoretical background to humour and dialogical reasoning (section 2). We will then look at the joke above in more detail (section 3) providing an informal description that we discern in the process of its comprehension. In section 4 we will move on to describe the relevant updates of this joke using TTR, a type theory with records

(Cooper, 2012). In the final section, we will discuss the limitation of our approach in connection to dialogue systems.

2 Related work

2.1 Humour research

In the past decades competing visions on humour have been developed, introducing such notions as ‘incongruity’, ‘incongruity resolution’, ‘semantic script’, ‘superiority’, ‘relief’, ‘pseudo-logic’ and many others as key components of humour. Ritchie (2004) emphasises the importance of explicating these so-called ‘theory-internal’ concepts in ‘theory-external’ terms which will arise from more general explanations relying on underlying cognitive processes, such as text comprehension (Ritchie, 2018) and, in our case, incremental reasoning in dialogue.

Notable linguistic theories of humour, such as Semantic-Script Theory of Verbal Humour (SSTH, Raskin, 1985) and General Theory of Verbal Humour (GTVH, Attardo and Raskin, 1991; Hempelmann and Attardo, 2011) are mainly about humour competence. They abstract away from the actual process of joke comprehension and do not include processing as a crucial condition for humour (Ritchie, 2018). Acknowledging Ritchie’s claim about a deficiency of actual explanations regarding how jokes are processed as text, we view the dialogicity of joke processing as a crucial condition for getting humorous effect that may result in amusement or laughter.

One important consequence of the dialogicity of jokes is the presence of the possibility that interlocutors might interpret the same piece of discourse in distinct ways. This is often taken advantage of in humour, and one way to account for this is using a theory of enthymematic arguments warranted by topoi.

2.2 Computational humour

A considerable amount of literature has been published on computational humour, highlighting the importance of understanding humour for dialogue systems (e.g., Raskin and Attardo, 1994; Hempelmann, 2008; Binsted et al., 1995).

A number of authors have investigated *humour generation*, mainly using template-based approaches inspired by humour theories. Examples of generated humorous texts are puns (Ritchie, 2005), lightbulb jokes (Raskin and Attardo, 1994),

humorous names (Ozbal and Strapparava, 2012) and acronyms (Stock and Strapparava, 2005).

Much of the current literature on *humour recognition* pays particular attention to either detecting salient linguistic features, such as stylistic features (Mihalcea and Strapparava, 2005), hand-crafted humour-specific features (Zhang and Liu, 2014) and N-gram patterns (Taylor and Mazlack, 2004), or latent semantic structures, (Taylor, 2009; Yang et al., 2015). Yang et al. (2015), in addition, focus on humour anchors, i.e. words or phrases that enable humour in a sentence.

So far, however, there has been little discussion about detecting humour in an interactive setting. For example, recent studies were mostly concerned with scripted dialogues, such as TV series like ‘Friends’ and ‘The Big Bang Theory’. Purandare and Litman (2006) used both prosodic and linguistic features and Bertero and Fung (2016) used a text-based deep learning approach. Both of these studies marked utterances followed by laughs as humorous, and the rest as non-humorous. The main weakness of this approach is that in real dialogues laughter is not necessarily associated with humorous content: it is not always triggered by humour and can express wide range of emotions, such as amusement, aggression, social anxiety, fear, joy and self-directed comment (Poyatos, 1993; Provine, 2004) and may also be used to convey propositional content (Ginzburg et al., 2015). In addition to this, not all events that are perceived as humorous provoke laughter. Even though laughter in conversations can be predicted with a fairly high accuracy (Maraev et al., 2019), it is still not indicative of whether the preceding content was humorous as opposed to, for example, the laughter having been used to soften a bold opinion expressed by one of the interlocutors.

Therefore, in the current paper we employ a dialogue-driven rather than a humour-driven framework. In Section 2.3 we will give a brief account of enthymematic reasoning in dialogue and relate it to jokes and humour.

2.3 Rhetorical reasoning and humour

The *enthymeme* is originally a key device in the Aristotelian theory of persuasion. However, as we shall see, the concept has broader use. An enthymeme is an argument where the conclusion does not follow by necessity, usually because one

or more premises are not explicit in the discourse. Presenting an argument based on implicit information is possible since the members of an audience or participants in a conversation have knowledge and beliefs regarding the world around them, which they automatically supply to arguments where they fit. The implicit information can be of different kinds – general knowledge, contextually relevant information, socio-cultural assumptions, etc. In rhetorical theory, the rule of thumb underpinning an enthymeme is referred to as a *topos*. As noted by Jackson and Jacobs (1980), enthymemes do not only belong in rhetorical discourse, but are frequently occurring in conversation. This idea is in line with the semantic theory of topoi in Ducrot (1988, 1980); Anscombe (1995), where topoi are seen as essential for meaning that exceeds the semantic content conveyed by an utterance. So, what does enthymematic reasoning in dialogue in fact mean? In (ii) (Wilson and Sperber, 2004) we find an example of a reply to a question requiring enrichment with implicit assumptions in order to be seen as a relevant answer to the question.

- (ii) Peter Would you drive a SAAB?
 Mary I wouldn't drive any Swedish car.

The implied answer to the question in (ii) is that Mary would not drive a SAAB. This conclusion is based on the fact that a SAAB is a Swedish car. In approaches to implicit meaning like Gricean (or Neo-Gricean) pragmatics and Relevance theory (Wilson and Sperber, 2004), this conclusion is based on an assumption of relevance – why would Mary answer the way she does unless a SAAB is indeed a Swedish car? However, this view ignores the fact that Peter might not interpret the answer correctly if it is unsupported by assumptions in his information state. In Aristotelian dialectic and rhetoric, (ii) would be warranted by a *topos* – for example that if something is true for a particular *genus*, then it is also true of a *species* (subtype) of that genus – and a premise, in this case that a SAAB is a species of the genus car. If an interlocutor is not aware of either the *topos* or the premise, the answer given by Mary bears no relevance to the question. In our analysis we will not distinguish between topoi and premises. Following Ducrot (1988), we will refer to all rules or principles used to underpin reasoning as topoi.

In (iii) we see an example of where enthymematic reasoning underpinned by topoi creates a humorous effect.

- 1 A Are the bagels fresh?
 2 B No.
 (iii) 3 A What about the muffins?
 4 B Better get the bagels.

The context of the joke is that A goes into a bakery, presumably to buy bread or cakes. A first asks about the freshness of the bagels. The shop assistant, B, responds that they are not fresh. A, thinking about getting muffins instead, asks whether *those* are fresh, and B responds that A better get the bagels. This short dialogue is underpinned by two topoi – one saying that if some food is not fresh, you should not buy it, and one saying that if you have to choose between two food items, and one is fresher than the other, you should choose the fresher one:

$$\frac{\text{not_fresh}(x)}{\text{not_buy}(x)} \quad (1)$$

$$\frac{\text{fresher_than}(x, y)}{\text{buy}(x)} \quad (2)$$

Let us think of the updates of the dialogue above: After the first utterance the inquirer/customer, A, has communicated that they are considering buying some bagels, and that the freshness of the bagels will have impact on their willingness to buy them. When B has replied “no”, we know that the bagels are not fresh, and indeed, A starts inquiring about the freshness of other types of bread. We can assume that a *topos* along the lines of ‘don’t buy non_fresh food’ is accommodated in the dialogue. If B had not agreed with this, they would have said something like ‘they are not fresh, but they are actually best when they are a few days old, or similar’. The second exchange evokes the *topos* that if one food item is fresher than another, you should buy the fresher one. Both of these topoi seem acceptable, and most people would agree with them. However, in this case, two topoi are accommodated which, when instantiated in this particular context, lead to inconsistent conclusions. That is, one of the topoi says that A should buy the bagels and one that they should not, and this is of course, a type of incongruity. So the fact that a *topos* is accommodated which clashes with a previously accommodated *topos*, regarding the same question under discussion, seems to create the humorous effect in this case.

In the next section we will look at another example where humorous incongruity is achieved through clashes between reasoning items.

3 The elephant-in-a-fridge riddle: An analysis

Let's consider the example in (i), as it could be told in a dialogue situation. We use a made up example¹ because it allows us to abstract away from complex cultural and social assumptions as well as situational context, and treat the discourse on a level of very basic assumptions.

This joke is a good illustration of how interlocutors build a common ground incrementally, agreeing on and refuting topoi drawn on to underpin the dialogue.

3.1 An elephant

In the first part of the joke, in (iv), the question evokes a topos about how to put things in fridges, which is in some way restricted to the kitchen domain. In this context, the idea of how to put something into a fridge is obvious, and also restricted to things that are (usually) food, and of the right size. This leads the interlocutor, B, to say that he does not know how to put an elephant into a fridge.

- (iv)
- 1 A How do you put an elephant into a fridge?
 - 2 B Hmm, I don't know?
 - 3 A Open the door, put the elephant inside, close the door.
 - 4 B Haha okay

The joke-telling genre indicates in this instance that A's question ('How do you put x in a fridge') is not really a request for information but has an answer which is known to A and which is to be revealed to B. On the other hand, the question is odd, which leads B (or the audience) to expect a non-trivial answer. It is important to draw attention to this because it is this oddity that provokes a light chuckle from the listener when the triviality is revealed.

One way of characterising "oddity" is in terms of congruity (or incongruity) with regard to salient topoi. The activity of putting something in a fridge is associated with a particular sequence of events. However, this sequence of events or actions will

work more or less well to create the state of x being in the fridge. We can think of a scale of oddity for these kinds of questions (Table 1):

Degree	Example
Trivial	'How do you put a cheese in a fridge?'
Tricky	'How do you put a big cake in a fridge?'
Odd	'How do you put an elephant in a fridge?'

Table 1: Degrees of oddity

We can think of *trivial* and *odd* as eliciting incongruity. The trivial question addresses something that is considered to be known, and the odd one addresses something ridiculously impossible. A nice example of a trivial question is 'Why did the chicken cross the road'? Questions are usually not supposed to address knowledge that can be easily inferred from the question (crossing the road entails getting to the other side of it).² This can be also be explained by violation of Grice's Maxim of Quantity: The answer 'to get to the other side' does not provide any additional information, and is thus superfluous.

A *tricky* question requires some non-trivial resolution, for example:

- (v)
- A: How do you put a wedding cake in a fridge?
 - B: You will need to remove one of the shelves.

3.2 A giraffe

Given the answer (3 A), B relaxes the implausibility of the elephant being put inside a fridge with no additional non-trivial actions. B accepts the required sequence of actions and acknowledges that (4 B). But is this enough to answer the question about a giraffe?

- (vi)
- 5 A How do you put a giraffe into the fridge?
 - 6 B Open the door, put the giraffe inside, close the door?
 - 7 A Wrong! Open the door, get the elephant out, put the giraffe inside, close the door.

B gives an answer based on his newly acquired storyworld, where elephants fit into fridges. But,

¹This joke appears at: <http://jeremy.zawodny.com/blog/archives/009023.html>

²The authors are aware of another, suicidal, interpretation of the chick riddle.

apparently what B has acquired is not enough: putting a giraffe into the fridge requires several other assumptions to be accommodated.

1. Even given that the fridge is ‘magical’, and big enough to fit an elephant, it is still not big enough to fit two big animals.
2. The joke-teller is talking about the very same fridge (this is especially important for languages in which there is no definite article)
3. Even if B understands that A is talking about the same fridge, it is not obvious that it already has an elephant inside, since it has never been explicitly said that an elephant has been put into a fridge.

3.3 Summary

An important quality of this example is that it illustrates how common ground is built gradually and following contributions exploiting the previous updates, the joke relies on A’s priming tricks and on not specifying what exact assumptions B should accept. If, in the earlier stage, the assumptions were characterised more precisely (e.g., ‘A: It is just a really huge fridge’), then the riddle would not work, or at least would be less funny.

The joke relies on an ambiguous and uncertain setting which creates the possibility of resolutions which generate humorous effects.

4 Formal account

The formal framework we will use is *Type Theory with Records* (TTR), a rich type theory successfully employed to account for a range of linguistic phenomena, including ones particular to dialogue (Cooper and Ginzburg, 2015).

In TTR agents perceive an individual object that exists in the world in terms of being *of a particular type*. Such basic judgements performed by agents can be denoted as “ $a : \text{Ind}$ ”, meaning that a is an individual, in other words a is a *witness* of (the type) Ind (ividual). This is an example of a *basic* type in TTR, namely types that are not constructed from other types. An example of a more complex type in TTR is a *p*type which is constructed from predicates, e.g. *fresher_than*(a, b), “ a is fresher than b ”. A witness of such a type can be a situation, a state or an event. To represent a more general event, such as “one individual item is fresher than another individual item”

record types are used. Record types consist of a set of fields, which are pairs of unique labels and types. The record type which will correspond to the aforementioned sentence is the following:

$$\left[\begin{array}{ll} x & : \text{Ind} \\ y & : \text{Ind} \\ c_{\text{fresher}} & : \text{fresher_than}(x, y) \end{array} \right] \quad (3)$$

The witnesses of record types are *records*, consisting of a set of fields which are pairs of unique labels and values. In order to be of a certain record type, a record must contain at least the same set of labels as the record type, and the values must be of a type mentioned in the corresponding field of the record type. The record may contain additional fields with labels not mentioned in the record type. For example, the record (4) is of a type in (3) iff $a : \text{Ind}$, $b : \text{Ind}$, $s : \text{fresher_than}(a, b)$ and q is of an arbitrary type.

$$\left[\begin{array}{ll} x & = a \\ y & = b \\ c_{\text{fresher}} & = s \\ c_{\text{price}} & = q \end{array} \right] \quad (4)$$

TTR also defines a number of type construction operations. Here we mention only the ones that are used in the current paper:

1. *List types*: if T is a type, then $[T]$ is also a type – the type of lists each of whose members is of type T . The list $[a_1, \dots, a_n] : [T]$ iff for all i , $a_i : T$. Additionally, we use a type of non-empty lists, written as $_{ne}[T]$, which is a subtype of $[T]$ where $1 \leq i \leq n$. We assume the following operations on lists: constructing a new list from an element and a list (cons), taking the first element of list (head), taking the rest of the list (tail).
2. *Function types*: if T_1 and T_2 are types, then so is $(\lambda r : T_1. T_2)$, the type of functions from records of type T_1 to record type T_2 . Additionally, T_2 may *depend* on the parameter (the witness of type T_1 passed to the function).
3. *Singleton types*: if T is a type and $x : T$, then T_x is a type. $a : T_x$ iff $a = x$. In record types we use manifest field notation to represent singleton type. Notations $[a : T_x]$ and $[a = x : T]$ represent the same object.

4.1 Dialogue Gameboards in TTR

Following Ginzburg (2012) and Larsson (2002) we will model the progress of dialogues in terms of the *information states* of the dialogue participants. In our analysis we will focus on the part of a dialogue participant's information state that is shared. That is, what has in some way been referred to in the dialogue, or what is necessary to integrate in the information state for a dialogue contribution to be interpreted in a relevant way. We will refer to this shared part of an interlocutor's information state as the *Dialogue Game Board* (DGB) of that participant. We are particularly interested in how individual agents draw on individual (and sometimes distinct) resources. We will therefore use separate DGBs for each agent, rather than letting the DGB represent a God's eye notion of context. For example, although a topos may be of central relevance in the dialogue, it does not appear on the DGB until it has been made explicit, or until something has been said which has caused it to be accommodated. We model the DGB as a record type where labels are associated with types, as in 5.

$$\left[\begin{array}{l} \text{rhet_resources} : [\text{topoi} : [\text{Topos}]] \\ \text{dgb} : [\text{eud} : [\text{Enthymeme}]] \\ \quad : [\text{topoi} : [\text{Topoi}]] \end{array} \right] \quad (5)$$

The record type in 5 represents the type of the information state of a dialogue participant with regard to enthymematic reasoning. In the DGB we find the enthymemes under discussion and the topoi that have been evoked in the conversation. For a topos to be added to the dgb of a dialogue participant, it must have been accommodated by the participant. The field *rhet_resources* (rhetorical resources) represents the topoi that are available to a speaker for inventing and interpreting arguments.

4.2 Enthymematic Reasoning in the Elephant Joke

We model enthymematic inferences and the topoi that underpin them as functions from situations of particular types to other types of situation. For example, one topos relating to the situation described in the elephant dialogue in (iv), could be represented as a function from a situation of a type where someone opens the door of the fridge, puts an object inside, and shuts the door, to a type of

situation where the same object is in the fridge. We see this topos, τ_1 , in (6):

$$\tau_1 = \lambda r : \left[\begin{array}{l} x : \text{Ind} \\ y : \text{Ind} \\ z : \text{Ind} \\ \text{Cfridge} : \text{fridge}(x) \\ \text{Cagent} : \text{agent}(y) \\ \text{Copen} : \text{open}(y, x) \\ \text{Cput} : \text{put_in}(y, z, x) \\ \text{Csmall} : \text{small}(z) \\ [s : \text{in}(r.z, r.x)] \end{array} \right]. \quad (6)$$

A topos is to be seen as a non-monotonic principle of reasoning (Breitholtz, 2014), and as such the conclusion does not follow necessarily and in all cases. Just like the principle that if x is a bird, x flies, does not apply to situations where the bird in question is a penguin, there might be a number of situations where a topos about how food gets into a fridge does not apply. Relevant to the situation at hand is an exception regarding the size of the object. Thus, we include the constraint "small" to restrict the use of the topos to things that are small enough to fit into a fridge. τ_1 is part of B's rhetorical resources, that is, a collection of topoi that are available for B to use as warrants in reasoning. The situation suggested by A's question conveys an enthymeme ϵ_1 like that in (14).

$$\epsilon_1 = \lambda r : \left[\begin{array}{l} x : \text{Ind} \\ y : \text{Ind} \\ z : \text{Ind} \\ \text{Celephant} : \text{elephant}(z) \\ \text{Cfridge} : \text{fridge}(x) \\ \text{Cagent} : \text{agent}(y) \\ \text{Copen} : \text{open}(y, x) \\ \text{Cput} : \text{put_in}(y, z, x) \\ [s : \text{in}(r.z, r.x)] \end{array} \right]. \quad (7)$$

In order to integrate a topos based on an enthymeme under discussion, the topos accessed in the rhetorical resources of the dialogue participant must be relevant with regard to the enthymeme conveyed in the discourse. We define this as the enthymeme being a *specification* of the topos. An enthymeme ϵ is a specification of a topos τ iff the antecedent type of ϵ is a subtype of the antecedent type of τ , and, for any situation r , the result of applying ϵ to r , is a subtype of the result of applying

τ to r , as shown in (8).

$$\begin{aligned} \tau &= T_1 \rightarrow T_2 \\ \epsilon &= T_3 \rightarrow T_4 \\ T_3 &\sqsubseteq T_1 \\ \text{for any } r, \epsilon(r) &\sqsubseteq \tau(r) \end{aligned} \quad (8)$$

However, since the antecedent type of τ_1 involves a constraint “small”, which is not present in the antecedent type of ϵ_1 , ϵ_1 is not a specification of τ_1 . Interlocutor B does not have access to other relevant topoi regarding how do you put an elephant into a fridge, and replies that he does not know the answer to the question.

A’s next utterance evokes another topos — τ_2 — where the size constraint is removed, and the enthymeme under discussion is thus a specification of τ_2 , which is integrated in A’s DGB according to the update rule in (10) below.

$$\tau_2 = \lambda r : \left[\begin{array}{ll} x & : \text{Ind} \\ y & : \text{Ind} \\ z & : \text{Ind} \\ c_{\text{fridge}} & : \text{fridge}(x) \\ c_{\text{agent}} & : \text{agent}(y) \\ c_{\text{open}} & : \text{open}(y, x) \\ c_{\text{put}} & : \text{put_in}(y, z, x) \\ [s & : \text{in}(r.z, r.x)] \end{array} \right]. \quad (9)$$

$$\begin{aligned} \mathcal{F}_{\text{integrate_shared_topos}} = \\ \lambda r : \left[\begin{array}{ll} \text{rhet_resources} : \left[\begin{array}{l} \text{topoi} : [\text{Topos}] \end{array} \right] \\ \text{dgb} : \left[\begin{array}{l} \text{eud} : [\text{Enthymeme}] \\ \text{topoi} : [\text{Topos}] \end{array} \right] \end{array} \right] \cdot \\ \lambda e : \left[\begin{array}{l} t : \text{Topos} \\ c_1 : r.\text{rhet_resources}.\text{topoi}(t) \\ c_2 : \text{specification}(\text{fst}(r.\text{dgb}.\text{eud}), t) \end{array} \right] \cdot \\ \left[\text{dgb} : \left[\text{topoi} = \text{cons}(e.t, r.\text{dgb}.\text{topoi}) : [\text{Topos}] \right] \right] \end{aligned} \quad (10)$$

B then moves on to the second punchline of the joke, asking how to fit a giraffe into the fridge. Which enthymeme that is under discussion at this point is not obvious – B could interpret the situation in (at least) two ways. Either, the question is how to fit a giraffe into any fridge, or into the fridge that is already occupied by the elephant. On any of these interpretations, the enthymeme under discussion ϵ_2 (in 11) is similar to ϵ_1 , with the exception that the individual z is associated with the constraint “giraffe” rather than “elephant”, or that an individual is added which is associated with the

constraint “giraffe” without any other individual or constraint being removed.

$$\epsilon_2 = \lambda r : \left[\begin{array}{ll} x & : \text{Ind} \\ y & : \text{Ind} \\ z & : \text{Ind} \\ c_{\text{giraffe}} & : \text{giraffe}(z) \\ c_{\text{fridge}} & : \text{fridge}(x) \\ c_{\text{agent}} & : \text{agent}(y) \\ c_{\text{open}} & : \text{open}(y, x) \\ c_{\text{put}} & : \text{put_in}(y, z, x) \\ [s & : \text{in}(r.z, r.x)] \end{array} \right]. \quad (11)$$

However, since the size constraint is now gone, it should not matter. B’s DGB now looks like this:

$$\left[\text{dgb} : \left[\begin{array}{ll} \text{eud} = [\epsilon_2, \epsilon_1] : [\text{Enthymeme}] \\ \text{topoi} = [\tau_2] : [\text{Topoi}] \end{array} \right] \right] \quad (12)$$

B evaluates whether the enthymeme ϵ_2 is underpinned by the topos already integrated in the DGB, and since the the addition of a giraffe, including or excluding the elephant, does not matter since the size restriction from τ_1 is dropped in τ_2 , which means that τ_2 can be used to warrant ϵ_2 . B thus replies, in accordance with this reasoning, that you behave in the same way to put a giraffe into a fridge as you do with an elephant, that is, you open the door, put the giraffe in, and close the door.

$$\begin{aligned} \mathcal{F}_{\text{evaluate_enthymeme}} = \\ \lambda r : \left[\begin{array}{ll} \text{dgb} : \left[\begin{array}{l} \text{eud} : [\text{Enthymeme}] \\ \text{topoi} : [\text{Topos}] \end{array} \right] \\ \left[\begin{array}{l} t : \text{Topos} \\ c_1 : r.\text{dgb}.\text{topoi}(t) \\ c_2 : \text{specification}(\text{fst}(r.\text{dgb}.\text{eud}), t) \end{array} \right] \end{array} \right] \cdot r \end{aligned} \quad (13)$$

A takes advantage of the fact that B draws on the topos on his DGB, τ_2 . However, A’s final punchline evokes a third topos, τ_3 , which introduces a new constraint regarding the ability of an elephant and a giraffe to be in the fridge at the same time, possibly some kind of size restriction. Which is of course incongruous in relation to B’s previous information state. Thus, taking advantage of the set up of B’s DGB at each exchange in the dialogue,

A is able to create mismatches in B’s DGB, making use of at least one of *her* available topoi, τ_3 , (see the end of section 3 for other possible topoi which challenge B’s τ_2). In the case that A would be asked to justify this punchline, the answer could be along the following: ‘The fridge is huge but not enormous enough to fit two big animals’.

$$\tau_3 = \lambda r : \left[\begin{array}{l} x : Ind \\ y : Ind \\ z : Ind \\ c_{size} : huge_not_enormous(x) \\ c_{fridge} : fridge(x) \\ c_{agent} : agent(y) \\ c_{open} : open(y, x) \\ c_{put} : put_in(y, z, x) \\ [s : in(r.z, r.x)] \end{array} \right]. \quad (14)$$

5 Discussion

The aim of the present research was to examine how reasoning required for joke processing in dialogue situations can be explained by means of enthymemes and topoi.

The scope of this study was limited in terms of using constructed examples and abstracting away from real dialogue data. A further study with more focus on data from spoken language corpora is therefore suggested. Nevertheless, in modelling reasoning patterns, one needs to abstract away from certain local processing issues, such as speech processing and clearing out misunderstandings that do not rely on argumentation requiring common sense reasoning.

The current study indicates the importance of having resources such as topoi, that enable an agent to reason using non-logical arguments, for building future dialogue systems with a capability to recognise and understand humour. An issue that was not addressed in this study was whether topoi can be bootstrapped from any available sources, such as WordNet or massive amounts of textual data. Considerably more work needs to be done to describe how to choose the most salient topos from the available resources. A reasonable approach to tackle these issues could be to employ Bayesian networks, following Maguire (2019) who combines them with topoi to represent world knowledge in order to model conditionals.

Notwithstanding these limitations, this study offers some insight into formalising how humour

can be processed in a dialogue setting.

6 Acknowledgements

This research was supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. In addition, we would like to thank Robin Cooper, Christine Howes and our anonymous reviewers for their useful comments.

References

- Jean-Claude Anscombe. 1995. La théorie des topoi: Sémantique ou rhétorique? *Hermès*, 15.
- Salvatore Attardo and Victor Raskin. 1991. Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3-4):293–348.
- Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135.
- Kim Binsted et al. 1995. Using humour to make natural language interfaces more friendly. In *Proceedings of the AI, ALife and Entertainment Workshop, Intern. Joint Conf. on Artificial Intelligence*.
- Ellen Breitholtz. 2014. Reasoning with topoi - towards a rhetorical approach to non-monotonicity. volume *Proceedings of the 50:th anniversary convention of the AISB*, pages 190–198. AISB.
- Ellen Breitholtz, Christine Howes, and Robin Cooper. 2017. Incrementality all the way up. In *Proceedings of the Computing Natural Language Inference Workshop*.
- Robin Cooper. 2012. [Type theory and semantics in flux](#). In *Handbook of the Philosophy of Science*.
- Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. *Handbook of Contemporary Semantic Theory, The*, pages 375–407.
- Oswald Ducrot. 1980. *Les échelles argumentatives*.
- Oswald Ducrot. 1988. Topoi et formes topique. *Bulletin d’études de la linguistique française*, 22:1–14.
- Jonathan Ginzburg. 2012. [The Interactive Stance: Meaning for Conversation](#). Oxford University Press.
- Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*.

- Christian F Hempelmann. 2008. Computational humor: Beyond the pun? *The Primer of Humor Research. Humor Research*, 8:333–360.
- Christian F Hempelmann and Salvatore Attardo. 2011. Resolutions and their incongruities: Further thoughts on logical mechanisms. *Humor-International Journal of Humor Research*, 24(2):125–149.
- Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*, 66(3):251–265.
- Staffan Larsson. 2002. *Issue Based Dialogue Management*. Ph.D. thesis, University of Gothenburg.
- Eimear Maguire. 2019. Enthymematic conditionals: Topoi as a guide for acceptability. In *Proceedings of the IWCS 2019 Workshop on Computing Semantics with Types, Frames and Related Structures*, pages 65–74.
- Vladislav Maraev, Christine Howes, and Jean-Philippe Bernardy. 2019. Predicting laughter relevance spaces. In *Proceedings of the International Workshop on Spoken Dialog System Technology*.
- Chiara Mazzocconi, Vladislav Maraev, and Jonathan Ginzburg. 2018. Laughter repair. In *Proceedings of SemDial 2018 (AixDial)*, pages 16–25.
- Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics.
- Gözde Ozbal and Carlo Strapparava. 2012. Computational humour for creative naming. *Computational Humor 2012*, page 15.
- Fernando Poyatos. 1993. *Paralanguage: A linguistic and interdisciplinary approach to interactive speech and sounds*, volume 92. John Benjamins Publishing.
- Robert R Provine. 2004. Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science*, 13(6):215–218.
- Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215. Association for Computational Linguistics.
- Jonathan D Raskin and Salvatore Attardo. 1994. Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. *Pragmatics & Cognition*, 2(1):31–69.
- Victor Raskin. 1985. *Semantic mechanisms of humor*. Synthese language library, 24. Reidel, Dordrecht.
- Graeme Ritchie. 2004. *The linguistic analysis of jokes*. Routledge.
- Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.
- Graeme Ritchie. 2018. *The Comprehension of Jokes: A Cognitive Science Framework*. Routledge.
- Oliviero Stock and Carlo Strapparava. 2005. Hahacronym: A computational humor system. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 113–116. Association for Computational Linguistics.
- Jerry M Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The psychology of humor: Theoretical perspectives and empirical issues*, 1:81–100.
- Julia M Taylor. 2009. Computational detection of humor: A dream or a nightmare? the ontological semantics approach. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 429–432. IEEE.
- Julia M Taylor and Lawrence J Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.
- Deirdre Wilson and Dan Sperber. 2004. Relevance theory. In Laurence Horn and Gregory Ward, editors, *Handbook of Pragmatics*, pages 607–632. Blackwell.
- Diya Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376.
- Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898. ACM.

Co-ordination of Head Nods: Asymmetries between Speakers and Listeners

Leshao Zhang

Human Interaction Lab
Cognitive Science Research Group
Queen Mary University of London
leshao.zhang@qmul.ac.uk

Patrick G.T. Healey

Human Interaction Lab
Cognitive Science Research Group
Queen Mary University of London
p.healey@qmul.ac.uk

Abstract

Previous research suggests that if people unconsciously mimic their interaction partner's movement, they gain social influence. We compare the effectiveness of speakers that mimic listeners' head nods with speakers that use natural nods in a special customised virtual environment. The results suggest that listeners agreed more with mimicking speakers than natural speakers. However, there are also asymmetries in speaker-listener nodding in the high and low-frequency domain. Listeners nod significantly more than speakers in the high-frequency domain. This asymmetry may be an important factor in coordination. We conclude that speaker and listener nods have both different form and different functions.

1 Introduction

There is significant interest in the coordination of speaker and listener behaviour in conversation, especially mimicry of form and/or temporal synchronisation of behaviour. Previous research has suggested that people automatically mimic each others' movements and behaviours unconsciously during interaction ([Chartrand and Bargh, 1999](#)), usually within a short window of time of between three to five seconds. It is claimed that this can prompt changes in individuals' cognitive processing style, altering performance on tests of ability and creativity and shifting preferences for consumer products as well as improving liking, empathy, affiliation, increasing help behaviour and reducing prejudice between interactants ([Chartrand and Lakin, 2013](#)). Based on this idea, Bailenson and Yee conducted the "Digital Chameleons" (2005) study. They created a virtual speaker automatically mimic the listener's head nods and suggested that the mimicking agent was more persuasive than the nonmimicker. However, these effects have not been consistently replicated ([Riek](#)

[et al., 2010](#); [Hale and Hamilton, 2016](#); [Zhang and Healey, 2018](#)).

In this paper, we investigate mimicry effects in more detail by comparing natural, mimicked, acted or 'canned' (i.e. non-interactive) playback of nodding behaviour in dialogue. These experimental manipulations are achieved through the use of a special customised Immersive Virtual Environment (IVE) which supported multiple people real-time interaction. For each of these manipulations, we explore the dynamics of the joint head movements both inside the virtual environment i.e. what the participants see and respond to and compare this with the coordination of their actual nodding behaviours.

2 Background

2.1 Nonverbal Studies with Immersive Virtual Environments

Immersive virtual environments (IVEs) have provided new ways to experiment with nonverbal interaction ([Blascovich et al., 2002](#); [Healey et al., 2009](#); [Bailenson et al., 2001](#)). In face-to-face interaction studies, it is difficult to introduce experimentally controlled manipulations of nonverbal behaviours. In principle, IVEs enable control of all aspects of participant's non-verbal behaviour ([Bailenson et al., 2001](#)). They also provide researchers with access to all participant's motion data, including all visible movements, gaze, and gestures ([Blascovich et al., 2002](#)). This 'panoptic' capability allows for subsequent analysis of all behaviours from any arbitrary viewpoint, something that is impossible with video.

2.2 Digital Chameleons

The "Digital Chameleons" study ([Bailenson and Yee, 2005](#)) illustrates the potential of IVEs. Bailenson and Yee compared the persuasiveness

of a virtual agent which automatically mimics a listener's head nods at a 4 seconds delay with an agent which reproduced a previous listener's head nods (so playback of naturalistic head nods but random with respect to what is being said in the interaction). They found evidence that a mimicking agent is more persuasive than the playback condition when delivering a message to students that they should always carry their ID card.

Similar studies were repeated over recent years. Researchers either found the effects of the "Digital Chameleons" (Bailenson and Yee, 2007; Verberne et al., 2013; Stevens et al., 2016) or could not consistently replicate the result (Riek et al., 2010; Hale and Hamilton, 2016; Zhang and Healey, 2018). This suggested that we might not have enough understanding of the speaker-listener head-nodding coordination.

2.3 Head Nods

Head nods are an important conversational signal. They are the most frequent head movement behaviour among shakes and changes of angle/orientation, etc (Włodarczak et al., 2012; Ishi et al., 2014). One possible reason for the mixed evidence on head-nodding coordination is the potential for different kinds of nod with different frequencies.

Hader et al. (1983) distinguishes three different head nods by frequency: 1) slow head nods between 0.2-1.8 Hz 2) ordinary head nods between 1.8-3.7 Hz and 3) rapid head nods above 3.7 Hz. They also suggest that listeners mainly use ordinary head nods to signal 'YES', rapid head nods for synchrony and slow/ordinary nods for other tasks. Other definitions of head nods by speed have been used. For example, Hale et al. (2018) define slow head nods as between 0.2-1.1 Hz, fast head nods between 2.6-6.5 Hz and found that listeners produce more fast head nods than speakers.

Head nods also serve different functions for listeners and speakers, e.g., listeners use "back channel" nods to signal their agreement, interests or impatience or to synchronise with a speaker's head nods (Hadar et al., 1985); while speakers may nod to seek or check agreement, to signal continued speaking, to express emphasis or as 'beat' gestures that accompany the rhythmic aspects of speech (Heylen, 2005). Listener head nods are also primarily concurrent with the speaker's turn. Healey et al. (2013) showed that speakers nod more

than primary addressees and that this relationship varies depending on how fluent the speaker's performance is.

2.4 Cross Recurrence Quantification Analysis

Analysis of the coordination of speaker and listener head nods requires methods that can find coordinated patterns in time-series over a variety of temporal intervals.

Recurrence Quantification Analysis (RQA) (Webber Jr and Zbilut, 2005) is a nonlinear time-series analysis method for the analysis of chaotic systems. Cross Recurrence Quantification Analysis (xRQA) is RQA applied to two independent time-series, e.g., two participants and finds the degree of match between the two time-series at different temporal offsets. So, for example, it can detect if one person's nods are systematically repeated by another person. xRQA has been widely used in the analysis the coordination of the interactants in a conversation (Richardson and Dale, 2005; Dale and Spivey, 2006; Richardson et al., 2008).

xRQA reconstructs two one-dimensional time-series data to pairs of points in a higher Embedding Dimension phase space (Takens, 1981) using Time-Lagged copies. It calculates the distances between the reconstructed pairs of points. The points pair that fall within a specified distance (Radius) are considered to be recurrent. The recurrent points are visualised with Recurrence Plots (RPs) that show the overall amount of repetition of (%REC), the longest sequence of repeated behaviours (LMAX) and the predictability or determinism (%DET) of one sequence from another. More specifically, %REC is the percentage of recurrent points in the RP. It indexes how much the two time-series are repeated. LMAX is the length of the longest diagonal line segment in the RP. It indexes the coupling strength of the two time series. %DET is the percentage of recurrent points falls on diagonal lines. It shows how much one time-series is predictable from another.

3 Current Study

To investigate the coordination of speaker-listener head nods, we used a customised IVE (Figure 1) that supports multiple participants' real-time interaction. Participants interact through virtual avatars. An optical motion capture system (Vicon)



Figure 1: The IVE in the Listener's View

captures participant's body movements in real-time, and this drives the movement of the avatars inside the IVE. Eye and lip movements are not captured so an algorithm is used to generate naturalistic eye movements and vocal amplitude is used to drive lip movements (previous research suggests participant's find the animation broadly realistic (Zhang and Healey, 2018)). In this study we used an asymmetrical setting for the speaker-listener interaction: the listener is immersed into the IVE and sees the speaker as a virtual character while the speaker is not immersed but is in the same physical room as the listener (see Figure 2).

3.1 Procedure

One participant acts as a listener and, in the appropriate conditions, a second participant acts as the speaker. In each experiment trial, participants wear the marker suits for motion tracking after an introduction. The listener also wears the Oculus Rift HMD and interacts with the virtual representation of the speaker (avatar) in the IVE. Following Bailenson et al. (2005), the speaker is asked to deliver a short pre-written speech about student ID card regulations to the listener. The speaker faces the listener and can see their body movements but cannot make eye-contact (Figure 2). The monologue is about 2-3 minutes long. After the monologue, the listener is asked to fill an online questionnaire on a computer.

In the experiment, the virtual speaker's head nods are manipulated according to the assigned condition:

1. Mimic – the virtual speaker's head nods are exact mimics of the listener's head nods but at a 4s delay.
2. Playback – the virtual speaker's head nods are an exact replay of the nods of the previous listener's head nods.



Figure 2: Two Participants Were Doing the Experiment

3. Natural – the virtual speaker's head nods are an exact mapping of the real speaker's head nods.
4. Recording – the virtual speaker's full body movements are an exact replay of a pre-recorded animation of a speaker/actor.

The Mimic, Playback and Natural conditions were assigned in rotation while the Recording condition was applied whenever we had only one participant in an experimental trial.

3.2 Measures

The analysis is organised into two sections. First, subjective assessments of the effectiveness of the speaker. Second, the patterns of head-nodding behaviour for the virtual and real speaker-listener pairs as determined from the motion capture data.

3.2.1 Effectiveness of the Speaker

We did exactly the same measurement for the speaker's effectiveness as Bailenson and Yee did in the "Digital Chameleons" study. The effectiveness of the speaker was measured by listener ratings on a self-report questionnaire. Speaker effectiveness is assessed by 4 items about agreement (agreement, valuable, workable, needed of the student ID card regulation delivered by the speaker), 13 items (friendly, likeable, honest, competent, warm, informed, credible, modest, approachable, interesting, trustworthy, sincere and overall) on impressions of the speaker, and 8 items (enjoy, acceptable, isolating, attractive, comfortable, cooperative, self-conscious or embarrassed and overall) on the virtual speaker's social presence; with Likert scale range from 1 strongly disagree to 7 strongly agree. Based on our previous research, we made our null hypothesis:

H0 The effectiveness of the speaker does not differ across conditions.

3.2.2 Amount of Head Nods

The body movement data was recorded as the joint orientation time-series in degrees at 60 Hz. With the recorded head movement time-series data, we tested the difference of the number of head nods between the speaker and listener with the paired t-test in the frequency range 0-8 Hz. Peaks in the head-nodding time-series were treated as the point that the participant changed the direction of head movement and counted as a nod. The total amount of head nods was counted as the number of peaks in the head-nodding time-series data. A low pass filter was used on the time-series data with the cut-off frequency set to increase slowly from 0 to 8 Hz in the resolution of 0.1 Hz. Building on previous work our initial hypothesis was that:

H1 Speakers nod more than listeners in all the conditions.

3.2.3 Head-Nodding Coordination

The coordination of speaker-listener head-nodding was tested using the xRQA method. We calculated a baseline chance coordination of the speaker-listener nods by doing xRQA with randomly paired speaker's and listener's from the Natural condition. We compared the head-nodding coordination in each condition as well as the chance level coordination for both the virtual and real speaker-listener pair. Given the assumption that non-verbal communication is coordinated in actual interactions, our second hypothesis is:

H2 Coordination of the speaker-listener head nods in all conditions is higher than chance.

3.3 Pairing Participants

Instead of running separate pairs of participants in each trial, we applied a shifted overlay participant arrangement. Each participant took part in two conditions. As shown in Figure 3, participants were asked to act first as a listener, then as a speaker. In each experimental trial, we had a previous participant as the speaker and a current participant as the listener. This setting ensured that before every experiment trial, the speaker has already been in the virtual environment and heard the message delivered by the previous speaker. Thus, the speaker would understand what the listener would see in the virtual world and be familiar with the message they would need to deliver to the listener. In the case of only one participant presented in the experiment, e.g. the very first experiment trial or one participant was not showing up,

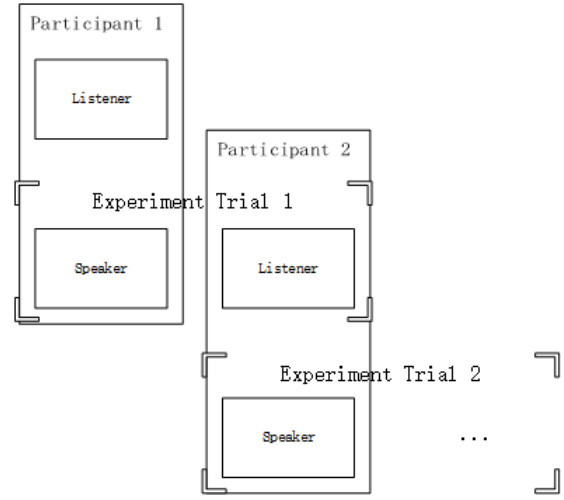


Figure 3: The Procedure Flow of Experiment II

we replaced the speaker with an animated virtual agent to deliver the message which corresponded to the recording condition.

3.4 Participant

54 participants were recruited by email, posters and through a participant panel. Each participant received 10 pounds for their participation. The final sample consisted of 29 female and 25 male students between 18 to 33 (Mean=21.89, SD=3.45). None of the participants reported severe motor, auditory or visual disabilities/disorders.

3.5 Result

3.5.1 Effectiveness of the Speaker

We tested the agreement, impression, social presence and the overall effectiveness of the speaker with the Generalized Linear Mixed Model (GLMM) analysis with the fixed factors of experiment condition (Mimic, Playback, Natural, Recording). Subject, speaker/listener's gender and the rating of their relationship were included as random effects. The result suggested that the listener's agreement with the speaker is slightly higher in the Mimic condition than in the Natural condition, $t_{50}=2.218$, $p=0.031$; the listener's impression of the speaker is higher in the Mimicry condition than in Recorded condition, $t_{50}=2.655$, $p=0.011$; the social presence of the speaker is higher in the Playback condition than in the Recorded condition, $t_{50}=2.870$, $p=0.006$; the overall effectiveness of the speaker is higher in the Mimic condition than in the Recorded condition, $t_{50}=2.491$, $p=0.016$. No other significant effect was found.

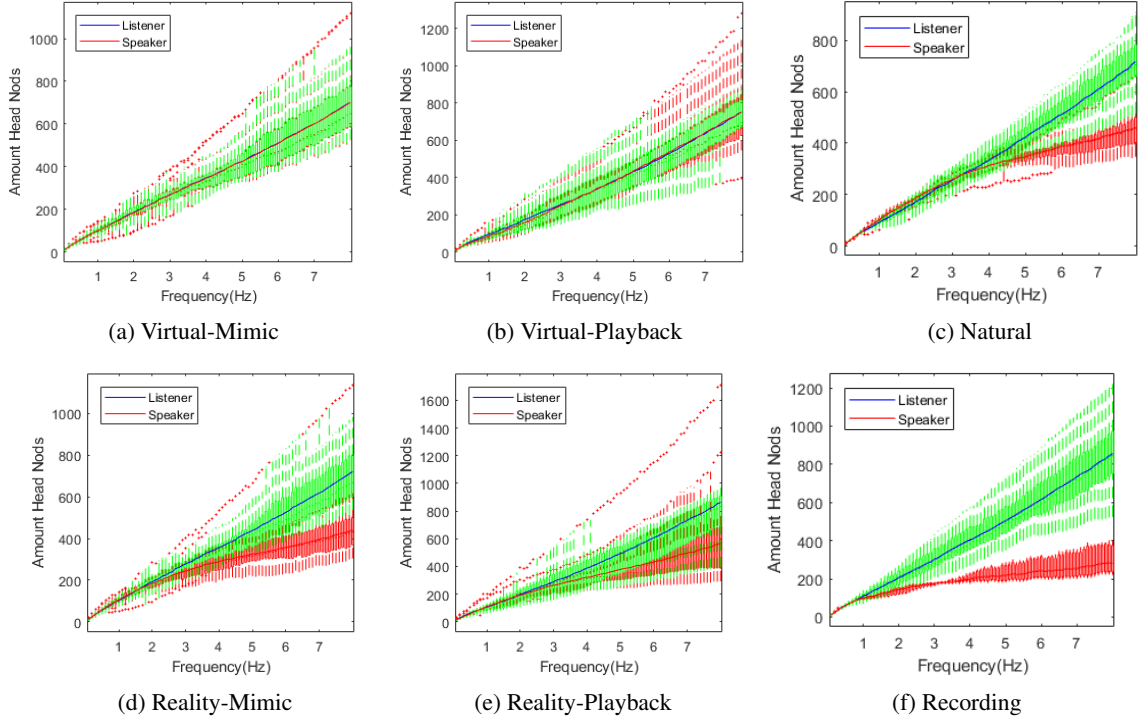


Figure 4: Boxplots of the Cumulative Amount of Head Nods for the Virtual and Real Speaker-Listener Pair.

3.5.2 Amount of Head Nods

We counted the number of head nods for every pair of participants. Figure 4 shows the distribution of the number of head nods for the virtual and real speaker-listener pair with a series of boxes. The X-axis is the cutoff frequency of the low pass filter. The Y-axis is the number of head nods for the participants through a certain low pass filter. The boxes were taken in the resolution of 0.1 Hz.

We compared the mean difference of the number of head nods between the listener and speaker below the certain frequency with the paired t-test. The result suggested that for the virtual pair of speaker and listener, there was no significant difference of the number of head nods under the condition of mimic and playback with the exception that the listener has a significantly higher amount of head nods than the real speaker in the frequency range from 4-8 Hz. Moreover, in the natural condition, the listener nodded less in the frequency range between 0.7-1.5 Hz whereas nodded more in the frequency between 3-8 Hz than the speaker. In the recording condition, the listener nodded significantly more than the speaker beyond 1 Hz. Figure 5 and 6 shows the mean difference of the number of head nods between the speaker and listener (listener to speaker) for the virtual and real pair respectively. The red dots in the graph indicate the

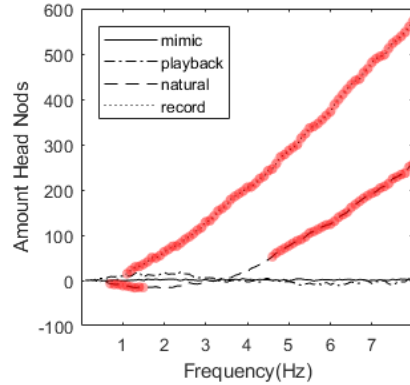


Figure 5: Cumulative Mean Difference of the Amount of Head Nods for the Virtual Listener-Speaker Pair.

points are under the significant level of 0.05.

3.5.3 Head-Nodding Coordination

xRQA was run for all the virtual and real interactional pairs with fixed parameters: Embedding Dimension=6, Time Lag=1, Radius=50, Non-normalised. The fixed parameters ensured that the parameters were kept as the controlled variables; the value of the parameters was picked to ensure no floor or ceiling effect for the xRQA outputs; not normalise the data to reduce the effect of non-movement. Figure 7 is the RP examples for the speaker-listener pair in each condition. The RPs in

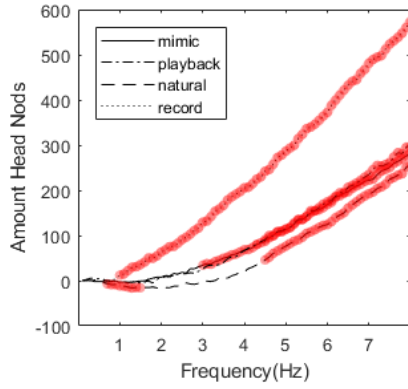


Figure 6: Cumulative Mean Difference of the Amount of Head Nods for the Real Listener-Speaker Pair.

mimic and playback condition were divided into virtual and real pair, whereas the RPs in natural and recording condition were not as they move the same in the virtual or real world. As we can see with the RPs, the virtual pairs with the mimic (Figure 7a) and playback (Figure 7b) condition were more coordinated (more dots in the RP) than the other conditions. They showed different coordinating patterns, e.g., there was a long diagonal line in the RPs of the mimic condition which was not seen in the RPs of the playback condition. The diagonal line has a tiny offset in Y-axis which indicated the 4s delay mimicry manipulation of the virtual speaker's head-nodding. The RPs of the recording (Figure 7f) condition showed the least coordination (least dots) of the speaker-listener pair. However, we cannot easily tell the difference between the RPs of the mimic (Figure 7d), playback (Figure 7e) and natural (Figure 7c) conditions with the real pairs.

The quantification outputs of the xRQA calculated the %REC, LMAX and %DET for all the virtual and real speaker-listener pairs. Figure 8 is the boxplots for those xRQA outputs by condition. The horizontal red lines are the chance level of these measures with the 95% confidence interval. We tested the %REC, LMAX and %DET for virtual and real speaker-listener pairs between conditions. The result suggested there was a significant ($p < 0.001$) difference between conditions on these items for the virtual and real speaker-listener pairs.

Games-Howell posthoc pairwise test suggested that: for the virtual speaker-listener pair, %REC was not significantly different from the chance level in the mimic, playback and natural condition, while it was significantly below the chance

level in the recording condition, Mean Difference (MD)=2.72, $p < 0.001$; with LMAX mimic was great than playback (MD=4588, $p < 0.001$), playback was great than natural (MD=99.4, $p < 0.005$), natural was at about chance level and great than recording (MD=32.5, $p < 0.001$); %DET was above the chance level in the mimic (MD=2.75, $p < 0.001$) and playback (MD=3.0, $p < 0.001$) conditions, and below the chance level in the recording condition (MD=4.37, $p < 0.001$), while not different from the chance level in the natural condition. For the real speaker-listener pair, %REC was below the chance level in the recording condition (MD=2.72, $p < 0.001$) while no significant difference from the chance level in the mimic, playback and natural condition; LMAX was not reliably different from chance in the mimic and natural conditions, whereas it was above the chance level in the playback condition and below the chance level in the recording condition; %DET was not significantly different from the chance level in the mimic, playback and natural condition, while it was significantly below the chance level in the recording condition (MD=4.37, $p < 0.001$).

4 Discussion

The results suggest that listeners may agree more with the speaker in the Mimic condition than in the Natural condition. Although this would indicate rejection of the null hypothesis H_0 , the evidence here is weak given the number of statistical comparisons made. There was also a difference in the effectiveness of the speaker when we manipulated its head movement behaviour. This was a surprise to us as we expected that there would be no difference in the speaker's effectiveness across all the conditions. Overall, the present study does not provide clear evidence for an effect of mimicry on agreement and persuasion but does indicate this might be worth pursuing in further work.

A much more salient and surprising finding is the distribution of head-nodding behaviour by the speaker and listener during the monologue. In terms of the number of head nods, the results show that listeners nodded significantly more in the high-frequency domain (above 3 Hz), and less in the low-frequency domain (between 0.7-1.5 Hz) in the Natural condition while no difference was observed in the other conditions. This suggests that we partly reject the hypothesis H_1 . In natural communication, speaker and listener nod dif-

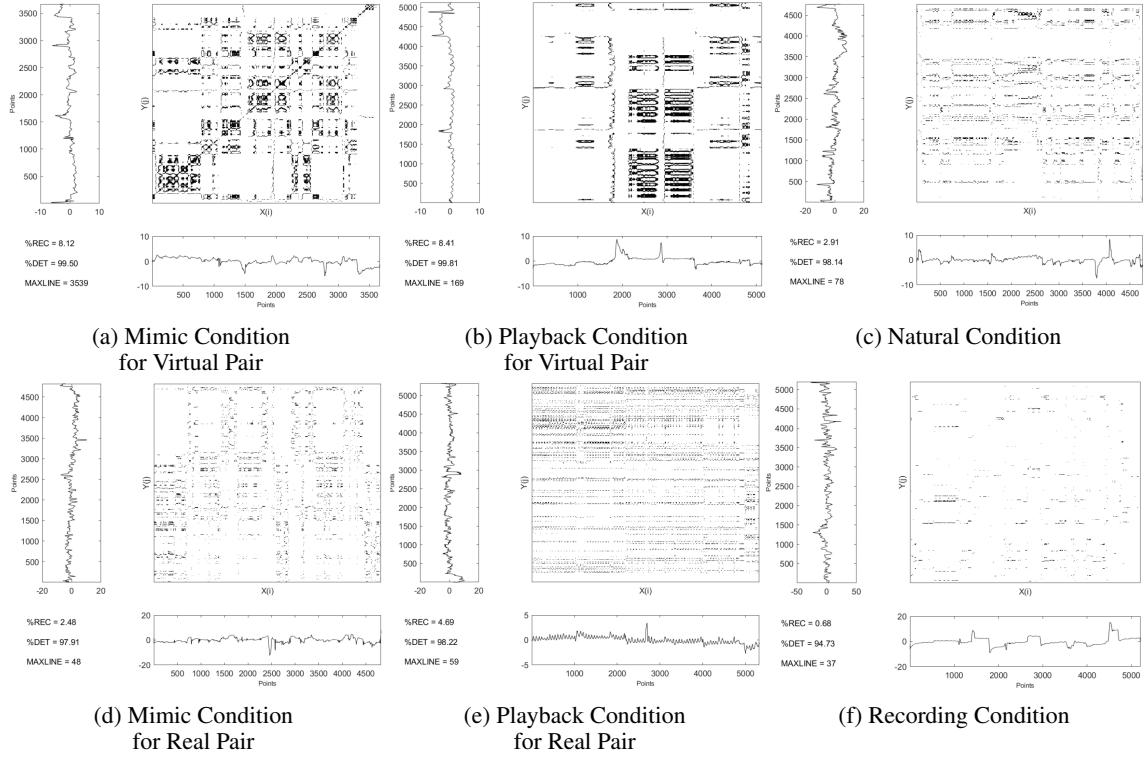


Figure 7: The Recurrence Plot for Speaker-Listener Pair in Each Condition

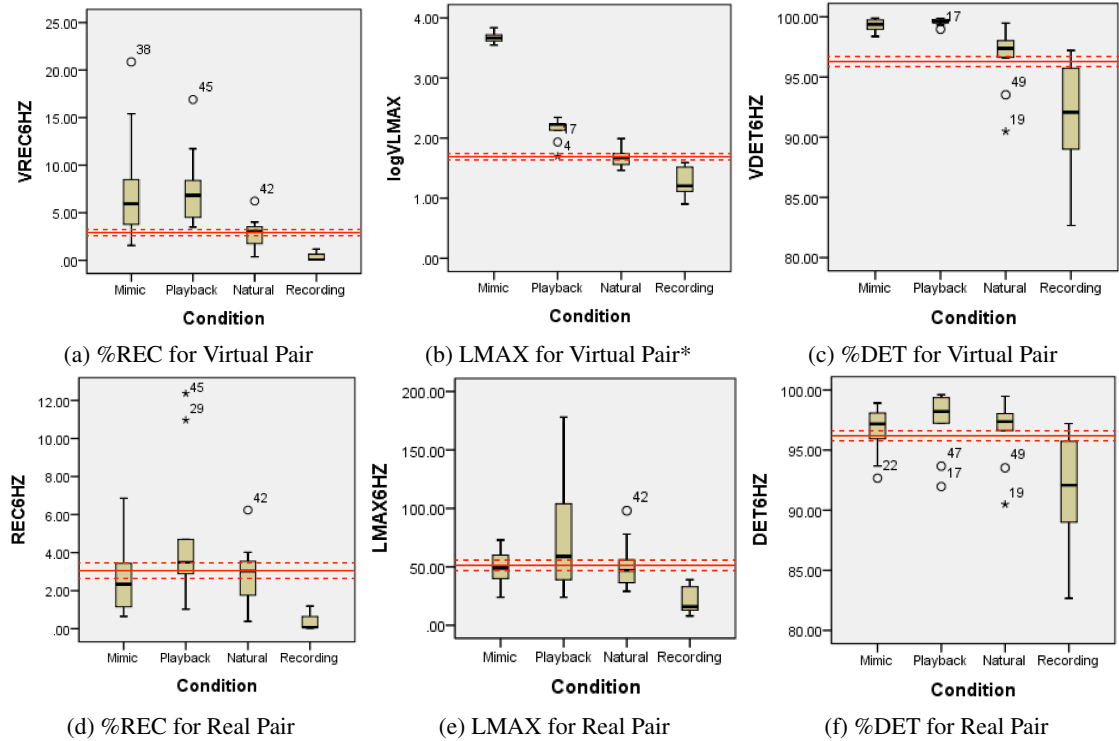


Figure 8: Boxplots of xRQA Outputs for the Virtual and Real Speaker-Listener Pair. Red line is the mean value of random pair with 95% CI. *A logarithm to base 10 was applied to the LMAX for Virtual Pair to compress the scale as the value in mimic condition is extremely high due to the experimental manipulation.

ferently in the high and low frequency domain (cf. (Hale et al., 2018)). Moreover, Figure 4f indicated that the speaker in the recording condition nodded much less in the high-frequency domain than the speaker in the other conditions. This is despite the fact that people performing the monologue in the Recorded condition moved much more overall than any of the other speakers. This might be because in the absence of a real listener, speakers perform significantly fewer fast nods. If fast nods are listener specific behaviours they might be a key contribution to the reciprocal dynamics between speakers and listeners. In other words, using an actor to perform a communication with the absence of the real listener leads to a non-verbal performance that is very different from the natural behaviour of a speaker in a live interaction - even when it is a monologue.

We also tested the speaker-listener's head-nodding coordination by applying the one-way ANOVA to the xRQA outputs. The most obvious point about the results illustrated in Figure 8 is that coordination with the Recorded speaker is consistently well below our measure of chance. The primary reason for this is that the people who recorded their monologues moved much more than those who delivered or listened to them live. These movements rarely matched those of their listeners who were relatively still.

Interestingly, the results also show that speaker-listener head-nodding coordination is not different chance in the Natural condition. In these data head-nodding coordination only exceeds chance in the Mimic and Playback conditions in the virtual speaker-listener pairs and is not different from chance with the real speaker-listener pairs. This is unsurprising in the virtual mimicry case since the experimental manipulation guarantees that nods are mimicked. The above chance coordination in the virtual Playback case is more puzzling. One possible explanation is that it occurs because we are pairing the head movements of listeners with listeners. Since the results indicate that listener head movements have a different characteristic frequency, this makes chance similarity higher than it is for speaker-listener combinations. This suggests accepting the null hypothesis for H2 as well. Natural speaker-listener head-nodding is no more coordinated than we could expect by chance. Recorded virtual speaker's head-nodding is significantly decoupled.

It is interesting to note that overall coordination of speaker-listener head-nodding is higher in the virtual world than in the real world with the mimic and playback conditions. The only difference between the two worlds is the speaker's head nods. In the virtual world, the speaker's head nods are taken from a listener, either from the listener themselves (Mimic condition) or from another listener (Playback condition), whereas in the real world, they are their actual head nods. Since listeners nod more than the speaker in the high-frequency domain, this could account for the elevated levels of virtual coordination. This is consistent with previous works (Hadar et al., 1985; Hale et al., 2018).

A potential limitation of the experimental approach used here is that the relation of the timing of head nods and vocal stress in the speech is not controlled. For example, Giorgolo and Verstraten (2008) suggest that temporally shifting the timing of hand gestures in the video away from its audio component create an anomalous feeling. Although only one participant (out of 54) reported a detachment of the head nods from the speech in debriefing, the effect of the correlation between the timing of speaker's head nods and the vocal stress in the speech is not clear in this work and needs further study.

5 Conclusion

The results suggest that in some circumstances speakers get more agreement by mimicking listener nodding behaviour. However, they also show that speaker and listener head nods are different in character. In the Natural interaction condition people do not coordinate their nodding behaviour more than would be expected by chance. The analysis of head-nodding behaviour suggests that this is because speakers nod more in the low-frequency domain and less in the high-frequency domain than the listener. The speaker-listener head-nodding coordination is above chance for the mimicking speaker, at chance for the natural speaker and below chance for an animated (recorded) virtual speaker. We also found that the fast nods are critical in the speaker-listener's coordination.

Acknowledgments

The work is supported by EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

References

- Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2001. [Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments](#). *Presence: Teleoperators and Virtual Environments*, 10(6):583–598.
- Jeremy N. Bailenson and Nick Yee. 2005. [Digital chameleons](#). *Society*, 16(10):814–819.
- Jeremy N. Bailenson and Nick Yee. 2007. [Virtual interpersonal touch and digital chameleons](#). *Journal of Nonverbal Behavior*, 31(4):225–242.
- Jim Blascovich, Jack Loomis, Andrew C Beall, Kimberly R Swinth, Crystal L Hoyt, and Jeremy N Bailenson. 2002. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2):103–124.
- Tanya L. Chartrand and John A. Bargh. 1999. [The chameleon effect: The perception-behavior link and social interaction](#). *Journal of Personality and Social Psychology*, 76(6):893–910.
- Tanya L Chartrand and Jessica L Lakin. 2013. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308.
- Rick Dale and Michael 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.
- Gianluca Giorgolo and Frans A. J. Verstraten. 2008. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, pages 31–36.
- U. Hadar, T.J. Steiner, E.C. Grant, and F. Clifford Rose. 1983. [Head movement correlates of juncture and stress at sentence level](#). *Language and Speech*, 26(2):117–129. PMID: 6664179.
- Uri Hadar, Timothy J Steiner, and F Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.
- Joanna Hale and Antonia F. De C. Hamilton. 2016. [Testing the relationship between mimicry, trust and rapport in virtual reality conversations](#). *Scientific Reports*, 6.
- Joanna Hale, Jamie A Ward, Francesco Bucccheri, Dominic Oliver, and Antonia Hamilton. 2018. Are you on my wavelength? interpersonal coordination in naturalistic conversations.
- Patrick G T Healey, Mary Lavelle, Christine Howes, Stuart Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker’s troubles. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.
- Patrick G.T. Healey, Chris Frauenberger, Marco Gillies, and Stuart Battersby. 2009. Experimenting with non-verbal interaction. In *8th international gesture workshop*.
- Dirk Heylen. 2005. Challenges ahead: Head movements and other social acts in conversations. In *Proceedings of Joint Symposium on Virtual Social Agents*, pages 45–52.
- Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2014. Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, 57:233–243.
- Daniel C Richardson and Rick Dale. 2005. Looking to understand: The coupling between speakers’ and listeners’ eye movements and its relationship to discourse comprehension. *Cognitive science*, 29(6):1045–1060.
- Michael J Richardson, Stacy Lopresti-Goodman, Marisa Mancini, Bruce Kay, and RC Schmidt. 2008. Comparing the attractor strength of intra- and interpersonal interlimb coordination using cross-recurrence analysis. *Neuroscience Letters*, 438(3):340–345.
- Laurel D Riek, Philip C Paul, and Peter Robinson. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces*, 3(1-2):99–108.
- Catherine J Stevens, Bronwyn Pinchbeck, Trent Lewis, Martin Luerksen, Darius Pfitzner, David MW Powers, Arman Abrahamyan, Yvonne Leung, and Guillaume Gibert. 2016. Mimicry and expressiveness of an eca in human-agent interaction: familiarity breeds content! *Computational cognitive science*, 2(1):1.
- Floris Takens. 1981. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.
- Frank M. F. Verberne, Jaap Ham, Aditya Ponnada, and Cees J. H. Midden. 2013. Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In *Persuasive Technology*, pages 234–245, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Charles L Webber Jr and Joseph P Zbilut. 2005. Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences*, pages 26–94.
- Marcin Włodarczyk, Hendrik Buschmeier, Zofia Malisz, Stefan Kopp, and Petra Wagner. 2012. Listener head gestures and verbal feedback expressions in a distraction task. In *Feedback Behaviors in Dialog*.
- Leshao Zhang and Patrick G.T. Healey. 2018. Human, chameleon or nodding dog? In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 428–436. ACM.

Character Initiative in Dialogue Increases User Engagement and Rapport

Usman Sohail Carla Gordon Ron Artstein David Traum

USC Institute for Creative Technologies

12015 Waterfront Drive, Los Angeles CA 90094-2536, USA

usman.f.sohail@gmail.com {cgordon|artstein|traum}@ict.usc.edu

Abstract

Two dialogue policies to support character initiative were added to the Digital Survivor of Sexual Assault, a conversational agent designed to answer questions about sexual harassment and assault in the U.S. Army: (1) asking questions of the user, and (2) suggesting conversation topics after a period of inactivity. Participants who interacted with a system that had these initiative policies reported that they felt higher engagement and rapport with the character, compared to participants who interacted with the baseline system. There was also a positive correlation between the number of instances of character initiative in a dialogue and the level of engagement and rapport reported by participants.

1 Introduction

There is a large body of work discussing the efficacy and benefits of using conversational agents as educational and assistive tools (Rickel, 2001; Kerly et al., 2009; Bickmore et al., 2013; Graesser et al., 2014; Gardiner et al., 2017). Some of these systems are designed more for formal educational learning, while others are designed to educate with the intent of changing the user’s behavior (Bickmore et al., 2013; Gardiner et al., 2017).

The Digital Survivor of Sexual Assault (DS2A: Artstein et al., 2019) was created to educate U.S. Army soldiers on the topic of sexual assault, in an effort to change attitudes and behavior and help prevent future harassment and assault. Inspired by the New Dimensions in Testimony project of conversation with Holocaust survivors (Traum et al., 2015), the DS2A system allows users to engage in a natural conversation with audio-visual recordings of Specialist Jarett Wright, a U.S. Army soldier who was sexually assaulted by senior members of his company while stationed in Iraq in 2010. Through conversation, users learn about



Figure 1: The Digital Survivor of Sexual Assault system at the SHARP Academy in Fort Leavenworth, Kansas. Photo Credit: Stephen P. Kretsinger Sr.

Jarett’s experiences of assault, retaliation, litigation, and other aspects of his life that were shaped by the assault. The interactive conversation is intended to forge a personal bond between the users and Jarett. The system is presently deployed by the U.S. Army in Fort Leavenworth, Kansas, and is used in training career professionals who deal with educating, preventing, and reporting of sexual harassment and assault in the Army (Figure 1).

An example conversation with Jarett is shown in Figure 2. This excerpt is typical of conversations with the DS2A system, and it shows the reactive nature of the system’s operation: the system answers the user’s questions, but does not take the initiative to ask questions of the user. However, there is evidence that asking questions helps the connection between two parties in a conversation. Burbules and Rice (1991) state that communicative virtues such as showing interest in a person’s thoughts and opinions promote serious and open discussion. Nystrand and Gamoran (1991) found that student engagement has a direct correlation with academic achievement, with students who were more engaged being better able to absorb and retain the information presented. Addi-

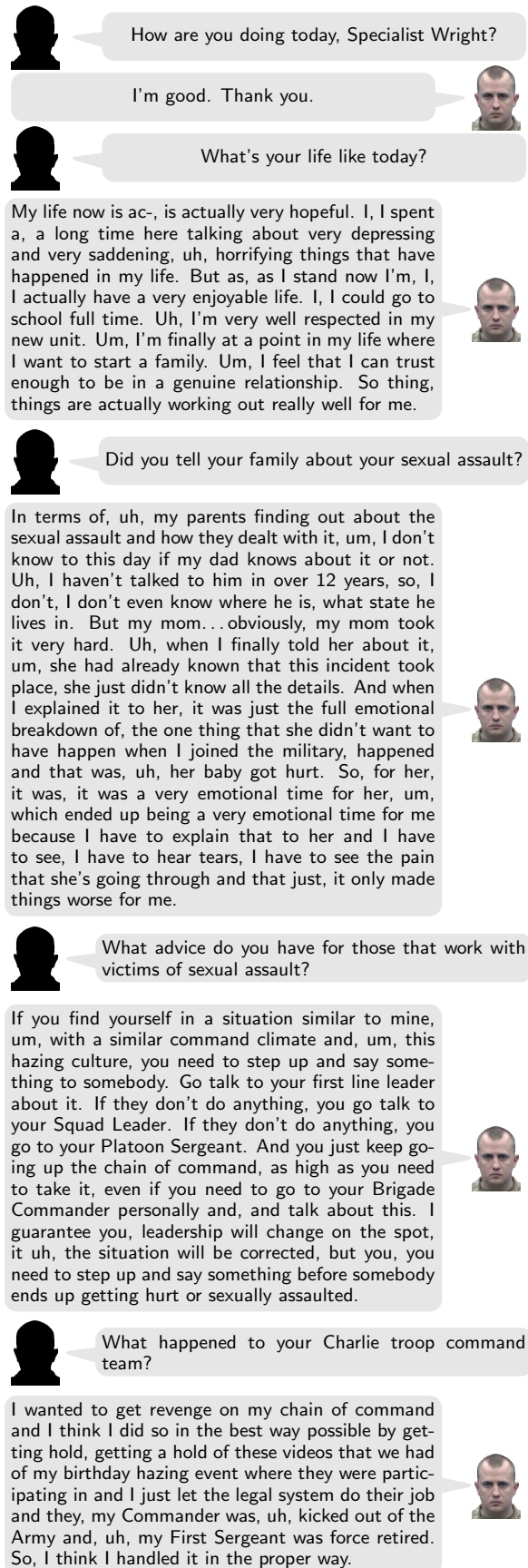


Figure 2: Sample conversation with Jarett Wright

tionally, Nystrand et al. (2003) underline the importance of maintaining student engagement with “authentic questions” which encourage them to share their own ideas and opinions. It is therefore reasonable to hypothesize that user engagement and rapport could be increased by giving the DS2A system an ability not only to respond to user questions, but also to engage the user by asking questions of its own.

Establishing rapport between users and conversational agents is an important component of creating engagement; in human-robot interaction it has been shown to increase customer comfort (Kanda et al., 2009) and social influence (Artstein et al., 2017). There has been a fair amount of work on using online measures to track user engagement in real time, using visual cues such as eye-gaze and head movement (see Sidner et al., 2005; Nakano and Ishii, 2010; Bohus and Horvitz, 2014); similar on-line measures have also been developed for assessing rapport (Zhao et al., 2016). However, such online measures were not available to us due to time and budget constraints, so we estimate user engagement and rapport using a post-interaction questionnaire.

This paper presents two main contributions: a set of policies for a reactive, question-answering character to take initiative and ask the user questions (section 2), and an experiment that shows that these policies increase user engagement and rapport, compared to interaction with a baseline system (section 3). We conclude by discussing some limitations of the experiment (section 4).

2 System Description

2.1 Baseline system

The baseline DS2A system is designed as an integrated application in the Unity game engine (<https://unity3d.com>); it incorporates several components from the USC ICT Virtual Human Toolkit (Hartholt et al., 2013), which is publicly available (<http://vhtoolkit.ict.usc.edu>). Input to the system is user speech, and the output is video clips of Jarett, recorded in our Light Stage studio. The overall system architecture is shown in Figure 3; a more detailed description is given in Artstein et al. (2019).

The baseline system uses a fairly standard pipeline for processing user utterances: audio from the user is sent to a speech recognizer, and the text output of the recognizer is sent to

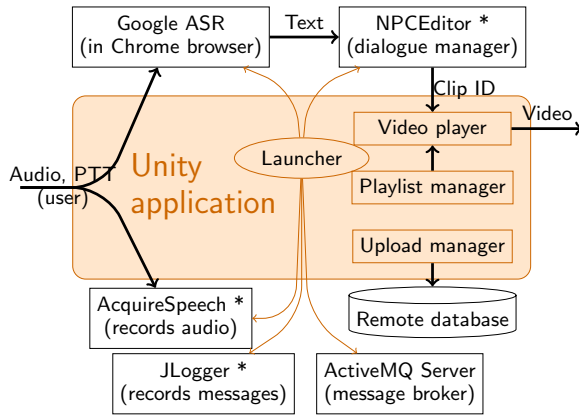


Figure 3: System architecture (* = Toolkit component)

the dialogue manager component, which selects a video response to be played back to the user. DS2A uses the NPCEditor component from the Virtual Human Toolkit (Leuski and Traum, 2011), which combines statistical Natural Language Understanding (NLU) with rule-based dialogue management. The NLU functionality is trained on question-answer pairs, and for each user question it returns a ranked list with zero or more appropriate responses (an empty list means that the classifier wasn't able to find an appropriate response). The dialogue manager functionality uses this ranked list to choose a response. The default action is to pick the top-ranked response; additional policies for avoiding repetition and handling non-understanding are described in section 2.3.

The baseline system is completely reactive: it acts in response to each user utterance, and it acts only in response to user utterances.

2.2 Mixed-Initiative Dialogue

We define character initiative as any character utterance which is not a reaction to a user utterance. To create a mixed-initiative dialogue experience, we implemented two types of character initiative by adding rules to NPCEditor's rule-based dialogue manager: *follow-up questions* and *timeout suggestions*.

Follow-up questions are designed to build rapport and encourage the user to engage more deeply with a topic; they are directly tied to the character utterances selected by the default dialogue manager. For example, if in response to a user question Jarett talks about his affinity for video games, he may follow up by asking the user about their video game preferences (Figure 4). If the user responds to the question, then Jarett will reply with

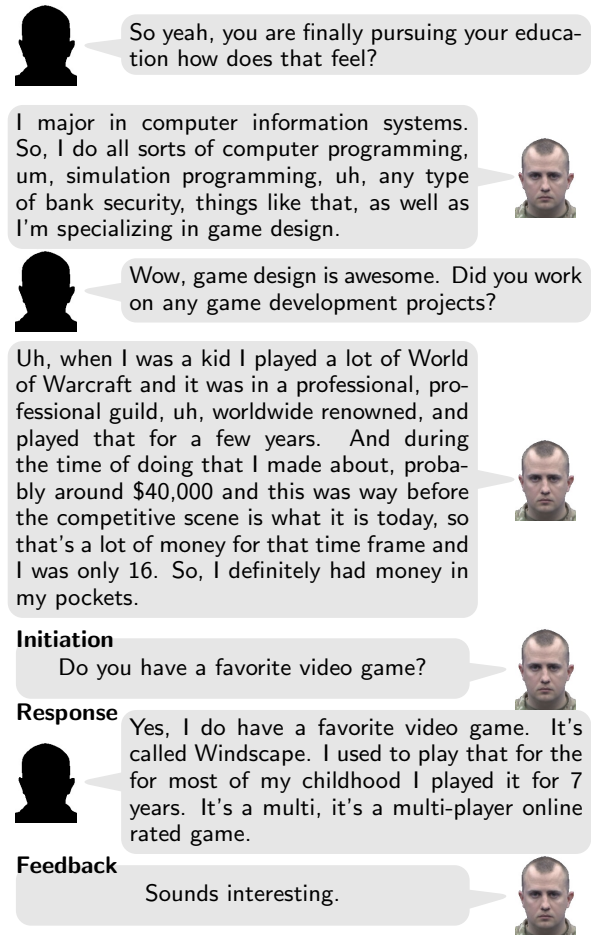


Figure 4: Follow-up question

a feedback utterance, completing an Initiation-Response-Feedback (IRF) exchange unit (Sinclair and Coulthard, 1975).

We annotated some of Jarett's responses with potential follow-up questions (see section 2.4 below); whenever Jarett issues a response, the dialogue manager checks to see if the response is associated with a potential follow-up question. In order to avoid asking too many questions, Jarett will ask a follow-up question only if an initiative utterance has not been delivered in any of the last three system utterances. Also, to avoid repeating himself, Jarett will not ask a follow-up question that he has already used in the same session, except for a few generic follow-ups which may be repeated, such as "How about you" or "If you were in my situation, what do you think you would have done?"

After Jarett asks a follow-up question, the system needs to determine whether the following user utterance is a response to the question. In principle this could be achieved by the NLU functional-

ity; however, at this point we do not have enough data to reliably train the NLU to identify user responses to follow-up questions. We therefore use a simple time-based heuristic, and assume that a user's utterance is a response to Jarett's question if a substantial portion of the question was actually played; in this case, Jarett will react to the user's utterance with a feedback utterance. However, if the user interrupts the initiative question more than two seconds before it is scheduled to end, it is assumed the user did not hear the initiative question, and the system will process the user's utterance using its default policy.

Timeout questions are designed to re-engage a participant after a period of inactivity. The time interval required to trigger a timeout varies between installations. The instructional system in Fort Leavenworth is typically used in front of a class, and we found that a threshold of 40 seconds provided a good balance between prompting the instructor and not being too disruptive. However, in piloting for the experiment reported below we found this threshold to be too long for one-on-one conversations, so we reduced it to 15 seconds for the experiment.

The timeout question can be a follow-up question to Jarett's previous utterance, if such a follow-up is available but wasn't asked previously due to the restriction of not asking too many follow-ups in succession. If a follow-up question is not available, then the timeout question utilizes the *topic suggestion* concept from NPCEditor's default dialogue manager (Leuski et al., 2006; Leuski and Traum, 2011). Originally designed to bring the user back into the domain of conversation after several consecutive instances of non-understanding (see section 2.3.2), we added topic suggestions as timeout questions, which serve not only to re-engage participants, but also to inform them of what the system can talk about in the event they cannot think of anything to ask. The system has 49 topic suggestion utterances covering 20 varied topics such as sexual assault prevention, reporting, retaliation, and bystander intervention. At the beginning of each session the system generates a list of all the topic suggestions, and then goes through the list throughout the session.

2.3 Additional dialogue policies

In addition to the policies above, the initiative system retains the reactive dialogue management

policies of the baseline system. The policies below are all default policies that come with NPCEditor and are described in Leuski and Traum (2011). These policies are triggered by an incoming user utterance and its NLU interpretation, which takes the form of a (possibly empty) ranked list of appropriate responses by Jarett; they handle responding to user utterances, avoiding repetition, dealing with non-understanding, and special cases.

2.3.1 Responses and repetition avoidance

If the NLU functionality returns a non-empty list of responses, the dialogue manager will choose one of those responses to play to the user. The choice balances rank (for best match) and recency (to avoid repetition): if the list contains utterances that were not used by Jarett in the last ten turns, it will use the highest ranked of these; otherwise, it will use the least recently used utterance.

2.3.2 Handling of non-understanding

If the NLU returns an empty list, the dialogue manager uses a strategy designed to gradually move the speaker to saying something the system can understand (Artstein et al., 2009). With each successive non-understanding, Jarett will go further down the following list.

1. Say again: Jarett asks the user to repeat their question.
2. Don't understand: Jarett tells the user he doesn't understand their question.
3. Can't answer: Jarett informs the user he can't answer their question.
4. Topic suggestion: Jarett suggests a new topic.
5. Topic story: Jarett tells a story based on the topic suggested in step 4.

If at any point the user says something that is understandable (that is, the NLU returns a non-empty list), then the policy goes back to that in section 2.3.1 and the non-understanding counter resets to zero.

2.3.3 Special utterances

While choosing a response typically means playing the video clip associated with the response, NPCEditor also allows for special response tokens that do some other action. One such token is used in the DS2A system: replaying the previous response. This token is selected by the NLU as the



Figure 5: Follow-up to positive sentiment

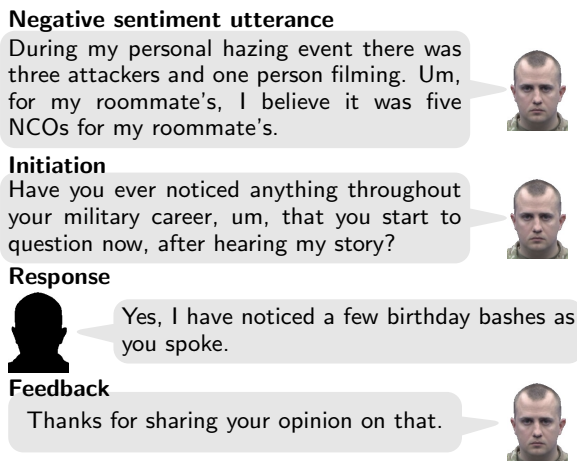


Figure 6: Follow-up to negative sentiment

interpretation of user utterances like “Could you repeat that?” If this token is selected by the dialogue manager (according to the policy in section 2.3.1), then Jarett will repeat his most recent utterance.

2.4 Annotations

The initiative policies require annotations that are not part of the default set-up of NPCEditor; these annotations were achieved by defining several new fields, detailed below (NPCEditor allows an unlimited number of user-defined fields).

Follow-up Questions. Each of Jarett’s utterances is annotated with a field that lists (zero or more) potential good follow-up questions.

Follow-up Sentiment. Character utterances with follow-up questions are annotated with a sentiment label (positive, negative, or neutral). Jarett’s feedback to a user’s response matches the sentiment of the utterance that triggered the follow-up question (Figures 5 and 6).

Utterance Length. Each of Jarett’s follow-up questions is annotated with its length, so that in case the question is interrupted by the user, the dialogue manager will know whether to issue a feedback utterance.

3 Experiment

In order to determine whether the mixed-initiative dialogue strategy has an effect on user engagement and rapport, we conducted an experiment comparing interactions with the baseline system and the initiative system. Each participant interacted with one version of the system, and we measured participant engagement and rapport using a post-interaction questionnaire. This section describes the experimental design and results, demonstrating that mixed initiative interaction does lead to increased engagement and rapport.

3.1 Method

Materials. We compared two versions of the DS2A system: a baseline system with the default dialogue management policies, and an initiative system with the default and initiative policies. The content was the same as in the system used by the U.S. Army, except that we removed some of Jarett’s utterances which included “colorful” language. Participants interacted with the system on a MacBook Pro laptop running the Windows 10 operating system, using the laptop’s built in display, speakers, microphone and trackpad.

Participants. A total of 58 participants were recruited through an ad on Craigslist (<https://www.craigslist.org>), paper flyers posted on a university campus, and a mass email sent to the Computer Science department. Participants were divided equally between the two experimental conditions. One participant was excluded from analysis because they chose to end the interaction early, resulting in a total of 29 participants in the baseline condition, and 28 in the initiative condition. The participants were 25 female, 32 male; most were between the ages of 18–27; the most common ethnic background was Asian; and the majority of participants had no affiliation or relation with the military (Table 1). Participants were given \$10 as compensation for their participation in the study.

Procedure. Each participant interacted with either the baseline system or the initiative version of

Age		Military Affiliation	
18–27	50	None	53
28–37	4	Close friend or family	4
38–47	1		
48–57	2		
Gender		Race	
Male	32	Asian	45
Female	25	Black/African American	2
		White	3
		Other	7

Table 1: Demographics of study participants

[Engagement questions]	
	I was interested in hearing what Jarett had to say.
	Jarett seemed interested in what I had to say.
	During the interaction, I lost track of time.
R	I found my mind wandering during the interaction.
[Rapport questions]	
	I felt Jarett created a sense of closeness or camaraderie between us.
R	Jarett created a sense of distance between us.
	I think Jarett and I understood each other.
R	Jarett communicated coldness rather than warmth.
	Jarett was warm and caring.
R	I wanted to maintain a sense of distance between us.
	I felt I had a connection with Jarett.
	Jarett was respectful to me.
R	I felt I had no connection with Jarett.
	I tried to create a sense of closeness or camaraderie between us.
R	I tried to communicate coldness rather than warmth.

Figure 7: Post-interaction questionnaire. Each question is rated on a 5-point scale. The label **R** indicates reverse-coded items.

the system for 20 minutes. Interaction took place in a quiet room on a university campus, and no experimenters or other personnel were present in the room during the interaction. This was done to ensure participants did not have any distractions in the room which might affect their overall engagement with the system, or their ability to build rapport. Participants were seated at a table in front of a laptop which displayed the video responses, and interacted with the system by pressing on the trackpad when asking their questions and releasing when their question was finished. At the end of the 20-minute interaction the experimenter re-entered the room and administered two questionnaires: one with questions about demographic information, and one designed to quantify the level of engagement and rapport felt by the user.

	Baseline	Initiative	Hi Init
Participants	29	27/28	21/22
Engagement	3.60	3.84	3.95
Rapport	3.66	3.83	3.94

Table 2: Means of questionnaire responses

Measures. The engagement and rapport questionnaire was given on two sheets of paper, the first with the engagement questions and the second with the rapport questions (Figure 7). The questions about engagement were devised specifically for this study, while the questions about rapport were adapted from [von der Pütten et al. \(2010\)](#) and [Artstein et al. \(2017\)](#). Each question was rated on a 5-point scale: 1 Strongly Disagree, 2 Disagree, 3 Neither Agree nor Disagree, 4 Agree, 5 Strongly Agree. We devised two measures, one for engagement and one for rapport, by summing the responses to the positive questions, subtracting the reverse-coded questions, and normalizing the result to the interval 1–5.

We compared the baseline and initiative groups using t-tests, and used ANOVAs to test for interactions with gender. Since initiative behavior is dependent on the course of the dialogue, some participants in the initiative condition experienced very little initiative behavior by the system. It is not immediately clear how to treat these low-initiative dialogues in the initiative group, because the user experience in these dialogues is similar to that of the baseline group. We therefore tested comparisons both between the baseline group and the full initiative group, and also between the baseline group and a “high initiative” subset, defined (somewhat arbitrarily) as those members of the initiative group who experienced at least two initiative utterances in their conversation with Jarett.

For the initiative group, we also measured the correlation between the number of initiative utterances and the level of engagement or rapport.

3.2 Results

The mean values of engagement and rapport for the various groups are shown in Table 2. One participant in the initiative group and high initiative subset did not answer the questions on engagements, so they were excluded from the analysis of engagement. The engagement and rapport scores are highly correlated, both for the baseline group ($r = 0.54$, $df = 27$, $p < 0.005$) and for the initia-

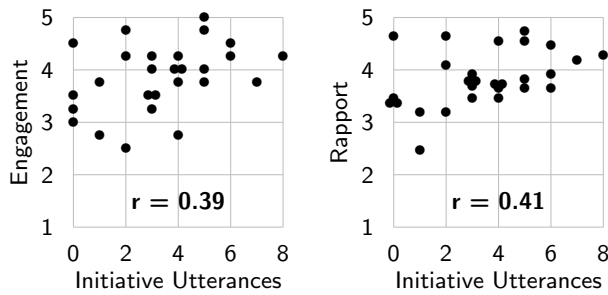


Figure 8: Positive correlation between number of initiative utterances and engagement/rappor

tive group ($r = 0.67$, $df = 25$, $p < 0.001$). This correlation could be interpreted as showing that the notions of engagement and rapport go hand in hand, or that the two instruments are actually tapping into a common notion, for example general satisfaction with the interaction, with the system, or with Jarett.

The difference in means between the baseline and initiative groups is not significant but shows a trend for engagement ($t(54) = 1.36$, $p < 0.1$) and a weak trend for rapport ($t(53) = 1.24$, $p = 0.11$). Between the baseline group to the high-initiative subset of the initiative group, the difference is significant for both engagement ($t(45) = 1.89$, $p < 0.05$) and rapport ($t(47) = 2.22$, $p < 0.05$). The above tests are one-tailed Welch’s two-sample t-tests. ANOVAs found no effect of gender nor any interactions with gender.

For the initiative group, we also calculated the correlation between the number of initiative utterances in the dialogue and the participant’s perceived engagement and rapport (Figure 8). Pearson’s correlation is positive and significant for both engagement ($r = 0.39$, $df = 25$, $p < 0.05$) and rapport ($r = 0.41$, $df = 26$, $p < 0.05$).

3.3 Discussion

Our results suggest that the mixed-initiative dialogue strategy we employed increases the user’s perception of their engagement and rapport with the DS2A system. The correlation of engagement and rapport scores with the number of system initiative utterances suggests that it is the character initiative that is responsible for the improved perception of the system.

We should make two notes about the observed correlations with system initiative utterances. First, since the length of the dialogues was fixed by the experiment at 20 minutes, we cannot tell whether the effect is due to the *number*

of initiative utterances or the *rate* of initiative utterances. Second, recall that the rate of initiative utterances was explicitly limited by the dialogue policies (section 2.2): by design, the system is not capable of asking initiative utterances all the time, reflecting our belief that such behavior is undesirable. Therefore, the observed correlation should not be extrapolated into making conclusions about very high rates of initiative utterances.

4 Conclusion

Our experiment shows that the mixed-initiative dialogue strategy increases the levels of perceived user engagement and rapport. In conclusion, we wish to discuss several limitations of the experiment.

Measuring engagement. Much previous research has noted the difficulty of measuring engagement (see e.g., Nystrand and Gamoran, 1991; Cerrato and Campbell, 2017). While many studies have had success in measuring engagement online through the analysis of eye-gaze behaviors, affective postures, and other auditory and visual cues (e.g., Nakano and Ishii, 2010; Sanghvi et al., 2011; Yu et al., 2004; Huang et al., 2016), this was not a strategy available to us for this study. As noted above, our off-line engagement measure was highly correlated with our rapport measure; a better, more direct way of measuring participant engagement would be helpful.

Repetitive feedback. The system contains a total of 16 feedback utterances, and many of these are fairly similar (for example, all 8 instances of feedback to negative sentiment are variations on “thank you for sharing”). In informal discussion, some participants mentioned that the feedback they were receiving seemed repetitive; this may have negatively impacted engagement or rapport. We suspect that a system with more varied feedback utterances could have a more pronounced effect on engagement and rapport.

Removed content. Since this experiment was conducted on a non-military population, the U.S. Army requested that we remove some of Jarett’s utterances that the Army felt were unsuitable for a civilian population (primarily utterances involving “colorful” language). The effect of removing these utterances was a slight reduction in overall coherence, since some user questions which could have

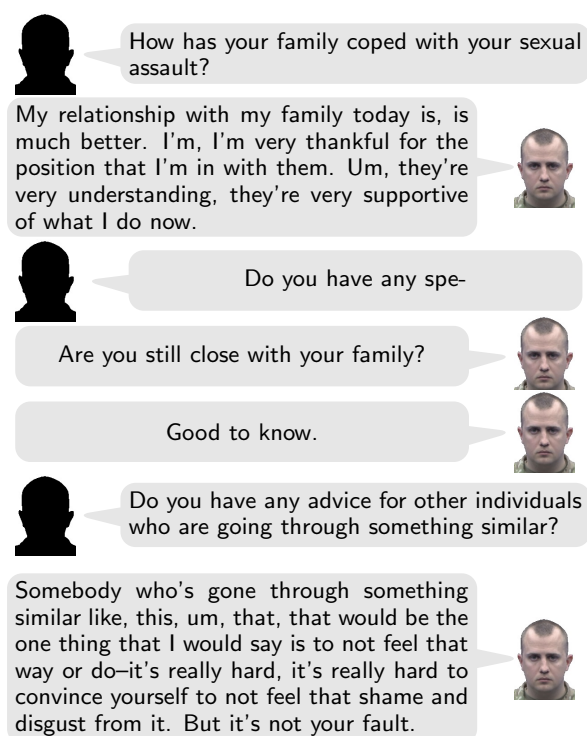


Figure 9: Timing issues

received a “colorful” response in the military system had to be treated as non-understanding in the experiment system. The lower coherence of civilian Jarett could have the effect of reducing participant engagement and rapport (though it would have a similar effect on both baseline and initiative conditions).

Timing of follow-up questions. As discussed in section 2.2, follow-up questions are triggered by a preceding utterance by Jarett, and are implemented by playing the question video clip directly after the conclusion of the video clip of the previous utterance. However, since each of Jarett’s video clips is a self-contained conversational turn, the appearance is as if Jarett is yielding the turn to the participant, and then immediately claims the turn and asks a question. This somewhat unnatural sequence of cues often led participants to believe that Jarett was done talking at the end of his original response, so the participant would ask another question immediately; in some cases, participants interrupted early enough that they never heard Jarett’s follow-up question, while in other cases they heard part of the follow-up question, which was cut short by their interruption.

Figure 9 shows a case in which Jarett’s follow-up question cut off the participant’s next question.

Jarett’s follow-up question started while the participant was mid-utterance, and an examination of the audio files reveals that the question was played in its entirety. Consequently, the system treated the participant’s cut-off utterance as a response to the follow-up question, and Jarett immediately replied with the feedback “Good to know” even though this was not conversationally appropriate. This is an example of the unnatural conversational cues causing a communicative breakdown, and incidents like this may have had a negative effect on the participants’ engagement and rapport. A better implementation of follow-up questions would be for the dialogue manager to somehow modify the ending of the trigger utterance clip, so that it does not give the impression of yielding the turn to the participant; however, this is not possible using the current tools.

Despite the above limitations, our study shows that a mixed-initiative dialogue strategy can lead to higher levels of perceived user rapport and engagement compared to a fully reactive strategy, when talking to agents designed to educate and inform users. As mentioned in the introduction, higher levels of engagement can lead to better retention of information, and higher levels of rapport can lead to increased social influence. This suggests that agents designed to educate for the purposes of effecting behavioral change would benefit greatly from implementing a mixed-initiative dialogue strategy.

Acknowledgments

The work depicted here 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, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum. 2009. *Semi-formal evaluation of conversational characters*. In Orna Grumberg, Michael Kaminski, Shmuel Katz, and Shuly Wintner, editors, *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *Lecture Notes in Computer Science*, pages 22–35. Springer, Heidelberg.

- 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.
- Ron Artstein, David Traum, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2017. [Listen to my body: Does making friends help influence people?](#) In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, pages 430–435, Marco Island, Florida. AAAI Press.
- Timothy W. Bickmore, Daniel Schulman, and Candace Sidner. 2013. [Automated interventions for multiple health behaviors using conversational agents](#). *Patient Education and Counseling*, 94:142–148.
- Dan Bohus and Eric Horvitz. 2014. [Managing human-robot engagement with forecasts and... um... hesitations](#). In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*, pages 2–9, Istanbul, Turkey. ACM.
- Nicholas Burbules and Suzanne Rice. 1991. [Dialogue across differences: Continuing the conversation](#). *Harvard Educational Review*, 61(4):393–417.
- Loredana Cerrato and Nick Campbell. 2017. [Engagement in dialogue with social robots](#). In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, volume 427 of *Lecture Notes in Electrical Engineering*, pages 313–319. Springer, Singapore.
- Paula M. Gardiner, Kelly D. McCue, Lily M. Negasha, Teresa Cheng, Laura F. White, Leanne Yinusa-Nyahkoon, Brian W. Jack, and Timothy W. Bickmore. 2017. [Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial](#). *Patient Education and Counseling*, 100:1720–1729.
- Arthur C. Graesser, Haiying Li, and Carol Forsyth. 2014. [Learning by communicating in natural language with conversational agents](#). *Current Directions in Psychological Science*, 23:374–380.
- Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. [All together now: Introducing the virtual human toolkit](#). In Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira, editors, *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29–31, 2013 Proceedings*, volume 8108 of *Lecture Notes in Computer Science*, pages 368–381. Springer, Heidelberg.
- Yuyun Huang, Emer Gilmartin, and Nick Campbell. 2016. [Conversational engagement recognition using auditory and visual cues](#). In *Proceedings of Interspeech 2016*, pages 590–594, San Francisco. ISCA.
- Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. [An affective guide robot in a shopping mall](#). In *HRI '09: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, pages 173–180, La Jolla, California, USA.
- Alice Kerly, Richard Ellis, and Susan Bull. 2009. [Conversational agents in e-learning](#). In *Applications and Innovations in Intelligent Systems XVI: Proceedings of AI-2008, the Twenty-eighth SGA International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 169–182. Springer.
- Anton Leuski, Jarrell Pair, David Traum, Peter J. Mc-Nerney, Panayiotis Georgiou, and Ronakkumar Patel. 2006. [How to talk to a hologram](#). In *IUI '06: Proceedings of the 11th International Conference on Intelligent User Interfaces*, pages 360–362, Sydney, Australia. ACM.
- Anton Leuski and David Traum. 2011. [NPCEditor: Creating virtual human dialogue using information retrieval techniques](#). *AI Magazine*, 32(2):42–56.
- Yukiko I. Nakano and Ryo Ishii. 2010. [Estimating user's engagement from eye-gaze behaviors in human-agent conversations](#). In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI)*, pages 139–148, Hong Kong. ACM.
- Martin Nystrand and Adam Gamoran. 1991. [Instructional discourse, student engagement, and literature achievement](#). *Research in the Teaching of English*, 25:261–290.
- Martin Nystrand, Lawrence L. Wu, Adam Gamoran, Susie Zeiser, and Daniel A. Long. 2003. [Questions in time: Investigating the structure and dynamics of unfolding classroom discourse](#). *Discourse Processes*, 35:135–198.
- Astrid M. von der Pütten, Niole C. Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. [“It doesn't matter what you are!” Explaining social effects of agents and avatars](#). *Computers in Human Behavior*, 26(6):1641–1650.
- Jeff Rickel. 2001. [Intelligent virtual agents for education and training: Opportunities and challenges](#). In Angélica de Antonio, Ruth Aylett, and Daniel Ballin, editors, *Intelligent Virtual Agents: Third International Workshop, IVA 2001, Madrid, Spain, September 10–11, 2001 Proceedings*, volume 2190 of *Lecture Notes in Computer Science*, pages 15–22. Springer.
- Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, Andre Pereira, Peter W. McOwan, and Ana Paiva. 2011. [Automatic analysis of affective postures](#)

and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-robot Interaction (HRI)*, pages 305–312, Lausanne, Switzerland. ACM.

Candace L. Sidner, Christopher Lee, Corry D. Kidd, Neal Lesh, and Charles Rich. 2005. [Explorations in engagement for humans and robots](#). *Artificial Intelligence*, 166:140–164.

John McHardy Sinclair and Richard Malcolm Coulthard. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press, London.

David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alecia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. 2015. [New Dimensions in Testimony: Digitally preserving a Holocaust survivor’s interactive storytelling](#). In Henrik Schoenau-Fog, Luis Emilio Bruni, Sandy Louchart, and Sarune Baceviciute, editors, *Interactive Storytelling: 8th International Conference on Interactive Digital Storytelling*, volume 9445 of *Lecture Notes in Computer Science*, pages 269–281. Springer, Heidelberg.

Chen Yu, Paul M. Aoki, and Allison Woodruff. 2004. [Detecting user engagement in everyday conversations](#). In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, pages 1329–1332, Jeju Island, Korea. ISCA.

Ran Zhao, Tanmay Sinha, Alan W. Black, and Justine Cassell. 2016. [Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior](#). In David Traum, William Swartout, Peter Khooshabeh, Stefan Kopp, Stefan Scherer, and Anton Leuski, editors, *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016 Proceedings*, volume 10011 of *Lecture Notes in Computer Science*, pages 218–233. Springer, Cham.

Modeling Intent, Dialog Policies and Response Adaptation for Goal-Oriented Interactions

Saurav Sahay, Shachi H Kumar, Eda Okur, Haroon Syed, Lama Nachman

Intel Labs, Anticipatory Computing Lab, USA

{saurav.sahay, shachi.h.kumar, eda.okur, haroon.m.syed, lama.nachman}
@intel.com

Abstract

Building a machine learning driven spoken dialog system for goal-oriented interactions involves careful design of intents and data collection along with development of intent recognition models and dialog policy learning algorithms. The models should be robust enough to handle various user distractions during the interaction flow and should steer the user back into an engaging interaction for successful completion of the interaction. In this work, we have designed a goal-oriented interaction system where children can engage with agents for a series of interactions involving ‘Meet & Greet’ and ‘Simon Says’ game play. We have explored various feature extractors and models for improved intent recognition and looked at leveraging previous user and system interactions in novel ways with attention models. We have also looked at dialog adaptation methods for entrained response selection. Our bootstrapped models from limited training data perform better than many baseline approaches we have looked at for intent recognition and dialog action prediction.

1 Introduction

Language technologies have benefited from recent progress in AI and Machine Learning. There have been major advancements in spoken-language understanding (Mikolov et al., 2013; Mesnil et al., 2015). Machine-learning approaches to dialog management have brought improved performance compared to traditional handcrafted approaches by enabling systems to learn optimal dialog strategies from data (Paek and Pieraccini, 2008; Bangalore et al., 2008). With availability of large amounts of data and advancements in deep learning research, end-to-end trained systems have also shown to produce state of the art results in both open-ended (Dodge et al., 2015) and goal-oriented

applications (Bordes et al., 2016) in the research community.

With the emergence of reliable ASR and TTS systems, we are now seeing platforms such as Alexa and Google Home and a plethora of domain and task-based dialog agents that allow users to take specific actions via spoken interface-based systems. Microsoft Research released Language Understanding Intelligent Service (LUIS) (Williams et al., 2015a,b), which helps software developers create cloud-based, machine-learning powered, language-understanding models for specific application domains. Google Dialogflow is an SDS platform for quick development of various conversational agents that can be deployed on various platforms such as Amazon Alexa, Google Home and several others. Systems like Google’s Dialogflow offer mechanisms such as explicit context management on linear and non-linear dialog flows to manage the conversations. The developer can attach input and output context states explicitly on various intents and create if-then-else flows via context variables. To go from explicit context management to implicit/automatic context management in SDS, probabilistic and neural network based systems are emerging in the research and development community (Bocklisch et al., 2017; Burtsev et al., 2018; Ultes et al., 2017).

Most dialog systems are categorized as either chatbots or task-oriented where chatbots are open-ended to allow generic conversations and the task-oriented system is designed for a particular task and set up to have short conversations (Jurafsky and Martin, 2018 (Online)). Goal-oriented Interaction Systems are somewhat midway between the two and should support longer duration interactions having various tasks to fulfill as well as support some non-task interactions.

We are developing a goal-oriented multimodal conversational system that engages 5 to 7 years old

children in concise interaction modules (Anderson et al., 2018). The overall goal of the system is to engage children in multimodal experiences for playful and learning oriented interactions.

Our application consists of interactions where children get introduced to the agent and they can play a simplified version of ‘Simon Says’ game with the agent. The dialog manager ingests verbal and non-verbal communication via the Natural Language Understanding (NLU) component (entities and intents) and other engines that process vision and audio information (faces, pose, gesture, events and actions) and generates sequential actions for utterances and non-verbal events. We describe the NLU, Dialog Manager (DM) and Dialog Adaptation modules in this work as shown in Figure 1. We build our NLU and DM based models on top of the Rasa framework (Bocklisch et al., 2017). We enrich the NLU Intent Recognition model in Rasa by adding additional features to the model. Typical goal-oriented dialog managers are modeled as sequential decision making problems where optimal interaction policies are learned from a large number of user interactions via reinforcement learning approaches (Shah et al., 2016; Liu et al., 2017; Su et al., 2017; Levin and Pieraccini, 1997; Cuayáhuítl, 2017; Dhingra et al., 2016). Building such systems to support children interactions is particularly difficult and we use a supervised learning approach using some initial training data collected via Amazon Mechanical Turk (AMT) to bootstrap the agents. Dialog Adaptation has been explored as part of parametric generation (Mairesse and Walker, 2007) as well as end-to-end NLG for generating contextually appropriate responses (Dušek and Jurčiček, 2016). We look at parametric generation methods and develop a simple ML classifier for Dialog Adaptation.

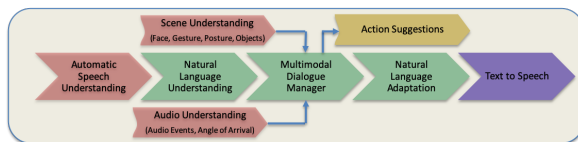


Figure 1: Multimodal Spoken Dialog System

2 Data Collection

Guided by User Experience (UX) research and Wizard of Oz user studies involving kids, we have designed sample interactions between the agent

and the kid (Figure 2) for the ‘Meet & Greet’ and ‘Simon Says’ interactions. The left half of Fig. 2 shows the task-oriented dialog flow. However, in a real system, one could imagine a lot of non-task-oriented or non-cooperative dialog involved. Especially in a dialog involving young kids, there could be a lot of chit-chat style conversations on things such as preferences or likes/dislikes of kids, a description of how the day went at school, or even the kids asking several questions about the agent itself. As shown in the right half of the Fig. 2, these can be categorized as either simple chit-chat or some unexpected utterances, as well as complaints or requests for help. To support our application, which includes a mixture of task and non-task-oriented conversations, we collect data via AMT for two types of interactions: ‘Meet & Greet’, and ‘Simon Says’, between the agent and the kid. We requested the turkers to provide us with a dialog, with agent on one side and kid on the other by providing sample interaction flows as shown in Fig. 2.

We collected 80 dialogs in total (65 stories for training, and 15 for test) for the two interactions. After the annotation process, we observed 48 user intents (including verbal intents and physical activities for the ‘Simon Says’ game), as well as 48 distinct agent actions. Note that for our NLU models, 26 distinct verbal intents are observed in the dataset, where 16 of them are categorized as goal-oriented (i.e., related to our ‘Meet & Greet’ and ‘Simon Says’ scenarios), and the remaining 10 are non-goal-oriented (e.g., chit-chat or out-of-scope). Further details of our NLU and dialog datasets can be found in Table 1 and Table 2, respectively.

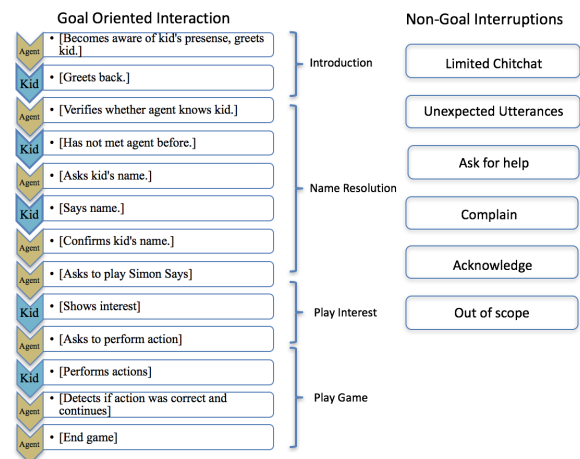


Figure 2: Domain of Application

	Training	Test
# of intents (distinct)	26	26
goal-oriented	16	16
non-goal-oriented	10	10
total # of samples (utterances)	779	288
total # of tokens	3850	1465
mean # of samples per intent	29.96	11.08
mean # of tokens per sample	4.94	5.09

Table 1: NLU Dataset Statistics

	Training			Test		
# of dialogs	65			15		
Meet & Greet	49			12		
Simon Says	16			3		
	Kid	Agent	All	Kid	Agent	All
# of intents/actions	48	48	96	28	28	56
goal-oriented	39	41	80	19	22	41
non-goal-oriented	9	7	16	9	6	15
# of turns	441	560	1001	97	112	209
goal-oriented	374	501	875	63	84	147
non-goal-oriented	67	59	126	34	28	62
# of turns per dialog	6.8	8.6	15.4	6.5	7.5	14.0

Table 2: Dialog Dataset Statistics

3 Models and Architecture

In this section, we describe the architectures we developed for the NLU, DM and Dialog Adaptation modules of the spoken dialog system pipeline.

3.1 NLU/Intent Understanding

The NLU module processes the input utterance to determine the user intents and entities of interest.

3.1.1 Intent Classifier

We use an intent classifier based on supervised embeddings provided as part of the Rasa framework (Bocklisch et al., 2017). This embedding-based intent classifier is based on ideas from the StarSpace algorithm (Wu et al., 2017) and embeds user utterances and intents into the same vector space. These embeddings are trained by maximizing the similarity between them. We also adapt sequence models and Transformer networks¹ to work with various features for developing our models.

3.1.2 Features and models for the NLU module

We utilize and extend the Rasa NLU module by adding various textual features and models to improve the performance of the intent classifier.

Textual features: We used text features such as number of words, first word, last word, bigrams, dependencies such as 1st/2nd/3rd person subject, inverted subject-verb order and imperative verbs, morphology features, hand constructed word lists,

¹<https://github.com/RasaHQ/rasa/pull/4098>

‘wh’ words, top n words, and many more. We add about 580 such textual features to our custom feature extractor.

Speech Act features: Verbal Response Modes (VRM) is a principled taxonomy of speech acts that can be used to classify literal and pragmatic meaning within utterances (Lampert et al., 2006). Utterances are classified into disjoint sets comprising Question, Disclosure, Edification, Advice, Acknowledgement, Reflection, Interpretation and Confirmation according to this model². The classifier (Sahay et al., 2011) also used the above text features for modeling the task and the top features in this classification task were domain independent features such as ‘?’, length of words, ‘you’, ‘i’, ‘okay’, ‘well’, etc.

Universal Sentence Embeddings: Universal Sentence Encoders (Cer et al., 2018) encode sentences into high dimensional vectors that has shown success in a variety of NLP tasks. We use the encoder model trained using a Transformer encoder³ to generate fixed length vectors as features for the NLU module in our pipeline. The motivation for using this model is to hope to recognize short utterances with similar meaning where the word level vectors may not provide enough information for correct intent classification.

Sequence Models: Long Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Bidirectional LSTMs (Schuster and Paliwal, 1997) can capture patterns and dependencies in sequential data using their memory gates and can robustly encode information. We use LSTM and BiLSTM layers to generate representations that are used in place of the fully connected embedding layer in the baseline model.

Transformers: (Vaswani et al., 2017) proposed a novel sequence-to-sequence network, the Transformer, entirely based on attention mechanism. The performance of Transformer model has generally surpassed RNN-based models and achieved better results in various NLP tasks. We use rich representations from the transformer model as an extension to the baseline model.

Bidirectional Encoder Representations from Transformers (BERT): Bidirectional Encoder Representations from Transformers, BERT (Devlin et al., 2018) represents one of the latest de-

²For the classifier, Disclosure and Edification classes were combined into one class

³<https://tfhub.dev/google/universal-sentence-encoder-large/3>

velopments in pre-trained language representation and has shown strong performance in several NLP tasks. We use pre-trained BERT model based features to generate representations for our dataset and use this in our NLU pipeline.

3.2 Dialog State Tracking

The task of a Dialog Manager is to take the current state of the dialog context and decide the next action for the dialog agent by using some policy of execution. Policies that are learnt from actual conversational data either use Markov Decision Processes (MDP) or Memory augmented Neural Networks. Memory augmented networks can update the context of the dialog by storing and retrieving the dialog state information. The dialog states and past system actions can be encoded as domain knowledge in the network to encode dialog manager based rules using Hybrid Code Networks (HCN) (Williams et al., 2017). One such dialog management policy that combines memory operations with domain knowledge embedding is the Recurrent Embedding Domain Policy (REDP) (Vlasov et al., 2018). This policy represents the dialog state embedding as a recurrent network with explicit management of the user memory and system memory modules. The user memory is used to attend to previous user intents and the system memory is used to attend to previous system actions. REDP uses Bahdanau Attention scoring to attend to previous user utterances (user memory) and memory address by location as developed in Neural Turing Machines (NTM) for system memory. With NTM, the previous system actions can be copied directly into the final dialog state representations to help recover from a sequence of uncooperative user utterances. REDP learns embeddings for dialog and system actions in the same space using ideas from the StarSpace algorithm (Wu et al., 2017) and uses these embeddings to rank the similarity between the recurrent embedding and the system action. Figure 3 shows the overall architecture of the dialog state tracker.

Our application handles multimodal streams of high frequency non-verbal input such as person recognition via face and speaker identification, gestures, pose and audio events. We pass on the information via separate modules to the dialog manager (bypassing the NLU) as relevant intents for goal-oriented interaction.

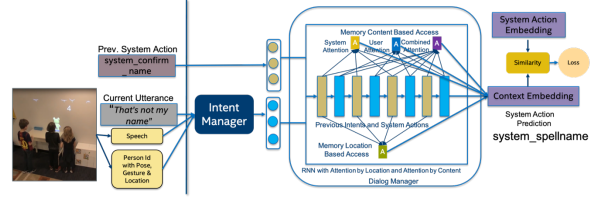


Figure 3: Dialog State Tracking Architecture

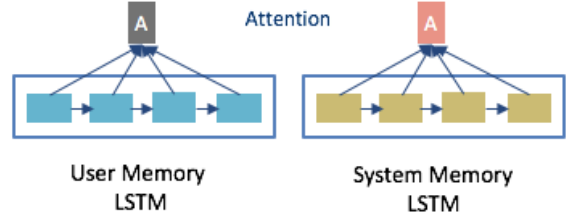


Figure 4: User Memory and System Memory

3.2.1 User and System Memories

Attention-based models (Bahdanau et al., 2014) can dynamically retrieve relevant pieces of information via selective reading through a relatively simple matching operation. In the REDP baseline architecture, separate mechanisms are used for attending to past user utterances and past system actions. While the system memory block helps the agent recover from uncooperative dialog by attending to and copying past system actions (using NTMs), the user memory mechanism uses Bahdanau Attention based matching to attend to relevant parts of past user utterances and enhance the user intent signal in the LSTM.

3.2.2 Fusion of User and System Memory

In this work, we capture the previous user-system interactions into the recurrent architecture by fusion of the signals from the user inputs and system actions. We hypothesize that attending to previous combinations of user utterances and system actions can help the bot choose the right action by directly leveraging multiple discrete views of the dialog context information. This may be useful in contexts involving multi-party interactions. Agents can also benefit from discrete attention in situations where deviations from the task-oriented dialog flow can lead to small multi-turn interaction where the context of the dialog (combinations of previous user utterances and responses) is crucial.

Figure 4 shows the memory units based on previous user intents and system actions. Figure 5 shows the fused memory unit obtained using the

dot product of user memory and system memory. It computes the product of all possible interactions between the intent features and system action features to obtain a memory of size $[user\ embedding \times system\ action\ embedding]$. This results in an enhanced representation of the dialog embedding vector (2 below). This dialog embedding vector is further enhanced using NTM based copying of the relevant previous system action as described in (Vlasov et al., 2018).

We incorporate another fusion technique as described in (Zadeh et al., 2017) where one larger memory component explicitly represents user memory, system memory and the interactions between the two (3 below).

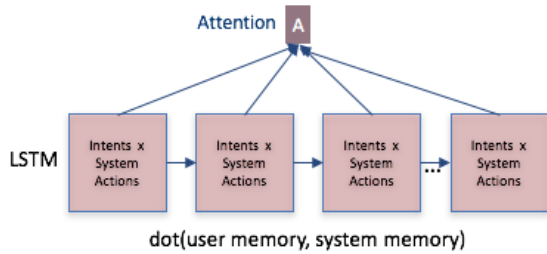


Figure 5: Capturing interactions between intents and system actions

We create the following combinations of the user and system memory blocks as part of the LSTM cell for attending to these memories for computing the relevant system action:

1. Concatenation of User and System memories (Single Memory Unit)
2. Tensor Dot of User and System memories (Single Memory Unit)
3. Tensor Fusion of User and System memories (Single Memory Unit)
4. Separate Attention on User and System memories (Two Memory Units)
5. Separate Attention on User memory and Tensor Dot (Two Memory Units)
6. Separate Attention on User memory, System memory and Tensor Dot (Three Memory Units)
7. Separate Attention on User memory, System memory and Tensor Fusion (Three Memory Units)

The configurations in the list above are a combination of the user memory, system memory and

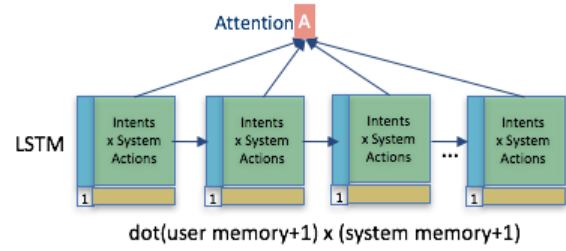


Figure 6: User Memory x System Memory Fusion

the fused user-system memory blocks. Configuration 6 above is also conceptually shown in Figure 7 with separate attention blocks to user memory, system memory and the combined user and system memory blocks.

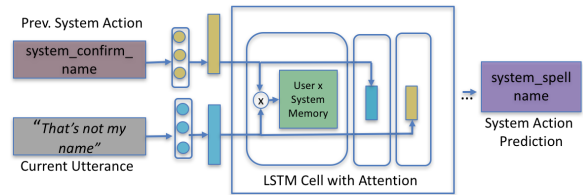


Figure 7: Simplified View of Contextual Attention

3.3 Response Adaptation

Children adapt to syntax and words used by their dialog partner more frequently than adults (Nilsenová and Nolting, 2010). This phenomenon is called Entrainment and generally applies to copying the conversational partner's attributes related to phonetics (Pardo, 2006), syntax (Reitter et al., 2006), linguistic style (Niederhoffer and Pennebaker, 2002), postures, facial expressions (L. Chartrand and A. Bargh, 1999), etc. It has also been linked to overall task success (Reitter and Moore, 2007) and social factors (Ireland et al., 2011). In this work, we explore lexical and syntactic adaptation, by using the similar referring expressions or sentence structure. SDS could pick a response at random from a list of responses for actions to match the predicted dialog state. In our work, we use the response adaptation module to score the responses and choose the best response instead of picking any random response. Figure 8 shows our architecture for response adaptation.

4 Experiments and Results

In this section we present and discuss the results for the NLU and DST experiments based on the architectural explorations explained in Section 3.

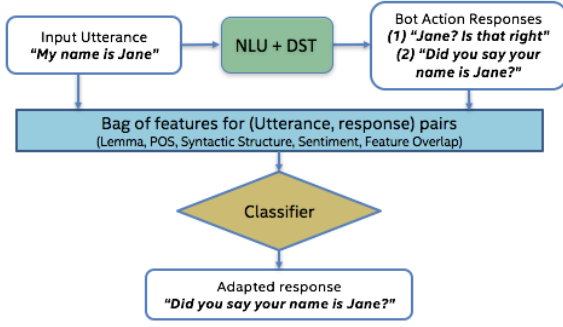


Figure 8: Response Adaptation

4.1 NLU Experiments

To evaluate the feature explorations for the NLU model, we present an ablation study on the LSTM, BiLSTM, Transformer, USE, text + speech acts features (SA) and BERT feature additions presented in Section 3.1. We use the embedding intent classifier as the baseline for the NLU experiments. For sequence models and Transformer model, we use word level features instead of the sentence-level features. Table 3 shows the precision, recall and F1-scores for the overall intent identification using different features.

The performance of the baseline model with the text + speech act features added is very similar to the baseline. Adding USE to the baseline improves the scores and this can be explained from the fact that USE, which is a transformer based model, generates better representations compared to individual text/speech act features. Consistent with the fact that BERT has shown a lot of success in several NLP tasks recently, we notice that adding BERT representations alone to the baseline improves the performance significantly. On the word-level features, we observe that while LSTM, BiLSTM and Transformer representations show slight improvement, adding BERT along with these shows a significant performance increase. In general, from Table 3, we can conclude that the speech act features have a very minor impact on the performance. The results show that adding BERT with the Transformer model helps achieving best performance on the dataset.

Table 4 presents some qualitative comparison of the baseline model with the text + speech acts (B+SA), USE (B+USE) and BERT (B+BERT) features against the ground truth labels (GT). From the table, we see that a short utterance such as ‘i am okay’ is labeled as ‘mynameis’

	Prec	Rec	F1
Baseline (StarSpace)	0.74	0.65	0.66
Baseline + LSTM	0.75	0.70	0.70
Baseline + BiLSTM	0.76	0.70	0.70
Baseline + Transformer	0.72	0.70	0.68
Baseline + BERT + BiLSTM	0.84	0.81	0.81
Baseline + BERT + Transformer	0.87	0.82	0.83
Baseline + SA	0.74	0.68	0.68
Baseline + USE	0.76	0.71	0.72
Baseline + BERT	0.81	0.77	0.77
Baseline + USE + SA	0.76	0.72	0.73
Baseline + USE + BERT	0.82	0.76	0.77
Baseline + USE + SA + BERT	0.83	0.78	0.78

Table 3: NLU Ablation Experiments on Meet & Greet and Simon Says Dataset

by the B+SA model. We believe that this can be attributed to the text features that look at words/phrases such as ‘i am’, and the ‘mynameis’ intent would usually start with ‘i am’. B+USE model predicts the correct label for this utterance. Although B+BERT model assigns the incorrect label, the prediction is semantically close to the GT. We again observe that a very short phrase such as ‘oh shit’ is classified incorrectly by both the B+SA and B+BERT models. We believe that for very short utterances, the USE model generates better representations as compared to BERT, and hence can produce more meaningful predictions.

An interesting point to observe is that, for utterances such as ‘that ain’t me Oscar’ or ‘are you Alexa?’, the BERT feature based model associates them with the ‘deny’ and ‘useraskname’ intents, respectively. Although these intents are wrongly identified, they are semantically very close to the ground truth labels ‘wrongname’ and ‘askabout-bot’. While the B+SA model tends to generate incorrect predictions for challenging examples, the B+USE model classifies ‘are you Alexa?’ as out-of-scope. Longer utterances such as ‘where is the nearest shopping mall’ are well handled by the BERT model, while the other models fail. We can conclude that the USE model could better handle very short sentences, and the BERT model performs better on the longer ones.

4.2 Dialog State Tracking Experiments

We investigate the role of single vs. multiple memories for attention as well as the impact of system memory and user memory fusion with the policy explorations. In Figure 9, we compare the results of the baseline REDP policy from (Vlasov

Kid's utterance	B+SA	B+USE	B+BERT	GT
'i am okay'	mynameis	userdoinggood	askhowdoing	userdoinggood
'oh shit'	askhowdoing	usermissedIt	affirm	usermissedIt
'are you Alexa?'	askhowdoing	outofscope	useraskname	askaboutbot
'that ain't me Oscar'	outofscope	mynameis	deny	wrongname
'where is the nearest shopping mall'	nextstep	useriamSick	outofscope	outofscope

Table 4: Qualitative analysis of baseline with USE+SA as features vs baseline with BERT

et al., 2018) with our policy changes on the dataset used by the authors that contain uncooperative and cooperative dialogs from hotel and restaurant domain. We use the same test set from the hotel domain and use a combination of cooperative and uncooperative dialogs from both hotel and restaurant domain for the training set. We divide the training set into 7 splits with 0, 5, 25, 50, 70, 90, 95 percent exclusion in the number of dialog stories in the domains. The baseline policy from (Vlasov et al., 2018) applies Bahdanau Attention scoring to the history of user utterances only. The policy does not explore attending to previous system actions or combinations of those for updating the RNN states as part of Bahdanau Attention scoring. In our experiments, we reuse the NTM based memory copying mechanism for system actions but explore additional effects of leveraging previous system actions and their combinations with the previous user intents. We see that using separate attention blocks on user memory, system memory and the combined memory using their dot product interactions help achieve slightly improved performance on this dataset. We can see some advantages in the case of very limited training data (when the agent cannot perhaps copy previous system actions) as well as in the case of full training set, where we

see a slightly higher number of correct test stories with other policies. Further investigation is needed to understand if the proposed policy changes in REDP would always benefit in certain scenarios. We also try to investigate the comparison between using a single (larger) memory attention vs. using multiple memory attentions. For example, as shown in Figure 9, 3 policy changes perform better than the baseline policy, all of which use separate memory blocks and all of these attend to the fused interaction representation of user and system memory.

Figure 10 shows the action prediction F1-scores for our 'Meet & Greet and Simon Says' dataset. Our test stories were collected using AMT and we allowed the bot and user to engage in any number of utterances and actions before passing control to the other. Since we also did not also impose any particular sequence in the game play, we didn't expect to get all action predictions correct in any one of test stories. We show the F1-scores for action predictions for the test stories varying the size of the training set. The more training data the agent has seen, more user utterance and system action interactions it has seen capturing application regularities, therefore we can hope to see improved

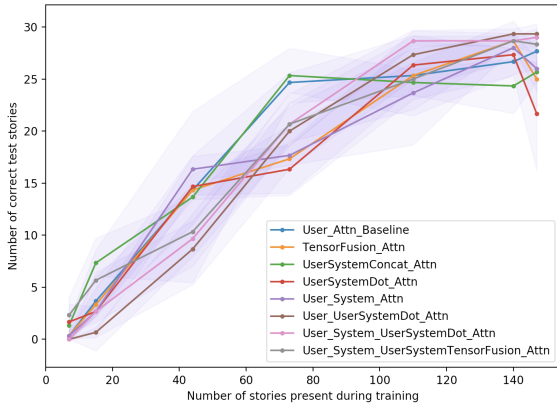


Figure 9: Performance of Models with RNN Attention over User and System Memory configurations with varying Training Sizes

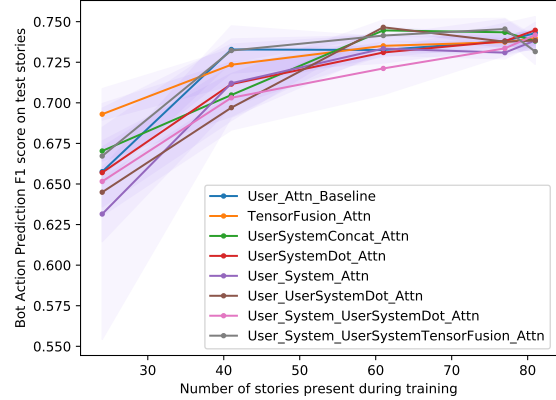


Figure 10: Action Prediction F1-score of Models with RNN Attention over User and System Memory configurations with varying Training Sizes

bot performance on unseen test sets with multiple memories and fusion configurations of attention units. From Figure 10, we can only say that there is a lot of variance in predictions with lesser training data, general trend for all these policies is get better with more training data. We see that the overall best F1-score is achieved with 75% of the training data with two separate attentions, one for user memory and another for the user-system fusion interactions.

4.3 Dialog Adaptation

For generating contextually appropriate responses, we collected multiple responses for bot utterances and trained ML models for selecting the most appropriate response from the list of responses. The goal of the classifier was to match the syntactic and linguistic style of the speaker. For this, we used 60 dialogs, with 32400 instances (2753 positive instances, 29647 negative instances) and 243 features. We created positive and negative instances automatically using feature overlap counts between the context dialog and the responses to be adapted. For feature generation, we extracted lemmas, Parts of Speech, Syntactic Structures and Sentiment Polarities using Stanford CoreNLP suite (Manning et al., 2014).

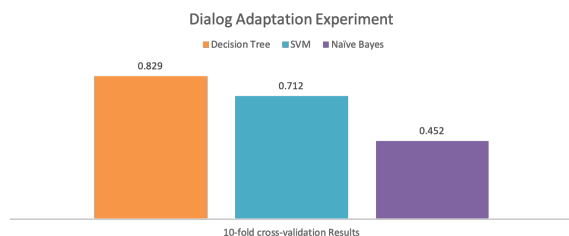


Figure 11: Dialog Adaptation

Figure 11 shows the cross-validation performance of the classifiers on the dataset. We use the Decision Tree based classifier for response adaptation in the spoken dialog system. Figure 12 shows a couple of response adaptation examples along with the positive and negative classification results for the two context utterances.

5 Conclusions & Future Work

We report preliminary explorations and results for our data driven spoken dialog system development for the multimodal ‘Meet & Greet and Simon Says’ goal-oriented application. The application involves phases of interactions for in-



Figure 12: Dialog Adaptation Examples

troducton, name resolution, game related interaction and actual game play involving children. We collect NLU and Dialog Data for our application using AMT, and manually identify non-goal-oriented intents and design interactions to include various non-goal-oriented or ‘uncooperative’ paths in the interaction. We have used and extended the Rasa NLU module and Rasa Core module for Dialog Management. Our application involves five to seven year-old children communicating with agents and we have seen from data that many children use very short utterances. In order to have a robust NLU, we have explored the use of lexical, syntactic and speech act related features (SA features), Universal Sentence Encoders as well as BERT embeddings for the embedding-based intent classifier which is a part of the Rasa NLU stack. We see the largest improvement in the NLU performance using the pre-trained BERT features and the Transformer model. For Dialog State Tracking, we extended the REDP policy by including different configurations of User and System Memory for RNN based Attention. We looked at a method for Single Memory Unit Tensor Fusion for combining User Memory, System Memory and tensor fused representation of User and System Memory. We explore other multiple memory unit configurations for RNN based Attention on history of User Intents, System Actions and their combinations. We saw improvements over the REDP baseline policy for the hotel and restaurant domain dataset as well as the ‘Meet & Greet and Simon Says’ dataset. We also explored Response Selection from the list of response templates as an Adaptation Classification problem using features such as Lemma/POS, Syntactic feature overlap and Sentiment of the response. As part of future work, we plan to extend the NLU and DM based models to include multimodal information in the pipeline.

Acknowledgments

Special thanks to Glen Anderson and the Anticipatory Computing Lab Kid Space team for conceptualization and the UX design for all the interactions. We thank Zhichao Hu (UCSC) who worked as a Summer Intern with us in 2017 and worked on the Dialog Adaptation module. We greatly acknowledge and thank the Rasa Team and community developers for the framework and contributions that enabled us to further our research and build newer models for the application.

References

- Glen J. Anderson, Selvakumar Panneer, Meng Shi, Carl S. Marshall, Ankur Agrawal, Rebecca Chierichetti, Giuseppe Raffa, John Sherry, Daria Loi, and Lenitra Megail Durham. 2018. [Kid space: Interactive learning in a smart environment](#). In *Proceedings of the Group Interaction Frontiers in Technology Workshop, GIFT@ICMI 2018, Boulder, CO, USA, October 16, 2018*, pages 8:1–8:9. ACM.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. [Neural machine translation by jointly learning to align and translate](#). *CoRR*, abs/1409.0473.
- Srinivas Bangalore, Giuseppe Di Fabbrizio, and Amanda Stent. 2008. Learning the structure of task-driven human–human dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1249–1259.
- Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. [Rasa: Open source language understanding and dialogue management](#).
- Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.
- Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. [DeepPavlov: Open-source library for dialogue systems](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder for english](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174. Association for Computational Linguistics.
- Heriberto Cuayáhuitl. 2017. Simpleds: A simple deep reinforcement learning dialogue system. In *Dialogues with Social Robots*, pages 109–118. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Bhuwan Dhingra, Lihong Li, Xiujuan Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. Towards end-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777*.
- Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.
- Ondřej Dušek and Filip Jurčiček. 2016. A context-aware natural language generator for dialogue systems. *arXiv preprint arXiv:1608.07076*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Computation*, 9(8):1735–1780.
- Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. [Language style matching predicts relationship initiation and stability](#). *Psychological Science*, 22(1):39–44. PMID: 21149854.
- Daniel Jurafsky and James H. Martin. 2018 (Online). *Ch 24: Dialog Systems and Chatbots. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd (draft) edition. Prentice Hall PTR, Upper Saddle River, NJ, USA.
- Tanya L. Chartrand and John A. Bargh. 1999. [The chameleon effect: The perception-behavior link and social interaction](#). *Journal of personality social psychology*, 76(6): 893–910. *Journal of personality and social psychology*, 76:893–910.
- Andrew Lampert, Robert Dale, and Cécile Paris. 2006. Classifying speech acts using verbal response modes. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 34–41.
- Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Fifth European Conference on Speech Communication and Technology*.
- Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2017. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. In *NIPS Workshop on Conversational AI*.

- 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.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. [The Stanford CoreNLP natural language processing toolkit](#). In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Kate G. Niederhoffer and James W. Pennebaker. 2002. [Linguistic style matching in social interaction](#). *Journal of Language and Social Psychology*, 21(4):337–360.
- Marie Nilsenová and Palesa Nolting. 2010. Linguistic adaptation in semi-natural dialogues: Age comparison. In *Text, Speech and Dialogue*, pages 531–538, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Tim Paek and Roberto Pieraccini. 2008. Automating spoken dialogue management design using machine learning: An industry perspective. *Speech communication*, 50(8-9):716–729.
- Jennifer S. Pardo. 2006. [On phonetic convergence during conversational interaction](#). *The Journal of the Acoustical Society of America*, 119(4):2382–2393.
- David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 808–815.
- David Reitter, Johanna D. Moore, and Frank Keller. 2006. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.
- Saurav Sahay, Hua Ai, and Ashwin Ram. 2011. Intentional analysis of medical conversations for community engagement. In *Twenty-Fourth International FLAIRS Conference*.
- M. Schuster and K.K. Paliwal. 1997. [Bidirectional recurrent neural networks](#). *Trans. Sig. Proc.*, 45(11):2673–2681.
- Pararth Shah, Dilek Hakkani-Tur, and Larry Heck. 2016. Interactive reinforcement learning for task-oriented dialogue management.
- Pei-Hao Su, Paweł Budzianowski, Stefan Ultes, Milica Gasic, and Steve Young. 2017. [Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 147–157, Saarbrücken, Germany. 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 Gasic, 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- Vladimir Vlasov, Akela Drissner-Schmid, and Alan Nichol. 2018. Few-shot generalization across dialogue tasks. *arXiv preprint arXiv:1811.11707*.
- Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. [Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning](#). *CoRR*, abs/1702.03274.
- Jason D Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoff Zweig. 2015a. Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luís). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 159–161.
- Jason D Williams, Nobal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015b. Rapidly scaling dialog systems with interactive learning. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 1–13. Springer.
- L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston. 2017. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*.
- Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Poster Presentations

Analysis of satisfaction and topics in repeated conversation through days

Tsunehiro Arimoto¹, Hiroaki Sugiyama¹, Masahiro Mizukami¹,
Hiromi Narimatsu¹, and Ryuichiro Higashinaka^{1,2}

¹ NTT Communication Science Laboratories, NTT Corporation

² NTT Media Intelligence Laboratories, NTT Corporation

{tsunehiro.arimoto.ub,hiroaki.sugiyama.kf,masahiro.mizukami.df,
hiromi.narimatsu.eg,ryuichiro.higashinaka.tp}@hco.ntt.co.jp

Abstract

For a dialogue system to function as a daily conversation partner, it must behave naturally not only in a single conversation but also in multiple conversations with its users. Analyzing how human satisfaction and topics change in conversations when conversations accumulate is useful for developing such systems. In this study, we analyzed multiple text-chats between two strangers for four days on a controlled schedule and revealed that their satisfaction and topic distribution depend on the length of the intervals between conversations.

1 Introduction

Chat-oriented dialogue systems are currently used for various tasks (eg., recommendation, therapy, and entertainment). While most systems assume that each user has a single conversation for a few minutes, for certain tasks, some systems assume that they have longer conversations with its users.

In order to make users have longer conversations, research for improving the naturalness or consistency of multi-turn dialogues has been actively investigated (Zhang et al., 2018). However, there are not many studies that focus on multiple conversations with its users. One of the differences between a single conversation and multiple conversations is that speakers have a short or long interval between conversations.

When an interval length between conversations is short, speakers might not care the small interval and behave as if they continue a long single conversation and might strengthen their engagement to the dialogue gradually. In contrast, when an interval between conversations is long, speakers may feel a difficulty in strengthening the engagement. If a system does not consider the effect of the length of the intervals, the system may speak to the user with wrong engagement strength that makes the user disappointed.

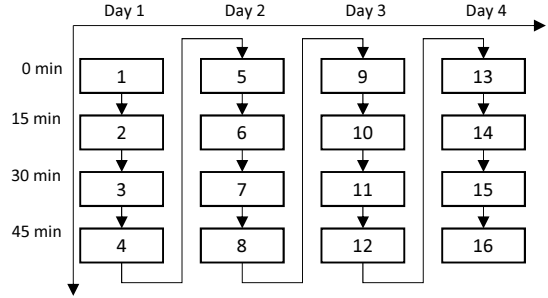


Figure 1: Scheduled collection of text chats per conversation pair (number in each cell means the order of the conversation or the cumulative number of conversations)

This study investigates how an interval length between conversations affects human satisfaction and topic selection in multiple conversations. We also analyze the trend as the number of conversation accumulates. This study focuses on dialogues in text chat in order to avoid the influence of the behavior or appearance of participants. We analyze a human-human text-chat corpus to investigate the natural behaviors of humans.

2 Repeated text-chats corpus

To investigate the effect of interval between conversations and the effect of their accumulation, we must analyze data where the time intervals between conversations are controlled. Since the level of intimacy between the speakers also affects the conversation contents, their relationships must also be controlled (Taylor and Altman, 1966).

A text-chat data collected by Higashinaka et. al., to implement an interactive system satisfies these conditions (Higashinaka et al., 2014). They collected four-days long chat data from two strangers who met on text-chat.

In their data, there is a controlled time interval between each text-chat. Figure 1 shows the

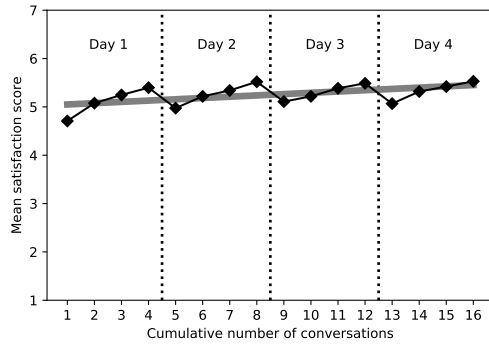


Figure 2: Satisfaction trends; Satisfaction is on a 7-Lickert scale (7 = very satisfied, 1 = completely unsatisfied). A gray line shows the regression line for all data.

recording schedule for one pair who text-chatted four times a day for four consecutive days. Because the four text-chats in each day are recorded within an hour, their corpus reveals the effect of short time intervals by comparing them. Since there is a long time interval when day changes, the comparison of two text-chats before and after the day ends reveals the effect of long time intervals. Our study analyzes part of their data: 2496 dialogues, 156 pairs, and 89 people.

3 Analysis

3.1 Human satisfaction

We analyze how human satisfaction is related to the length of interval between conversations and their accumulated amounts. We analyzed speaker satisfaction by the questionnaire results reported when they finished each text-chat.

Figure 2 shows the transition of the average scores of all the participants. Figure 2 shows that mean satisfaction score increased when the text-chats were repeated each day. Satisfaction decreased when the days changed (ex., cumulative number=4 vs number=5). These results suggest that human satisfaction increased during short intervals and decreased during long ones. The regression line for all data illustrates that human satisfaction gradually increased as the number of conversations increased.

3.2 Topic selection

Next we analyzed how topic selection changed over time. An annotator labeled the conversation topic (ex., fashion) per text-chat. The heat map in Fig. 3 shows the frequency of each topic at

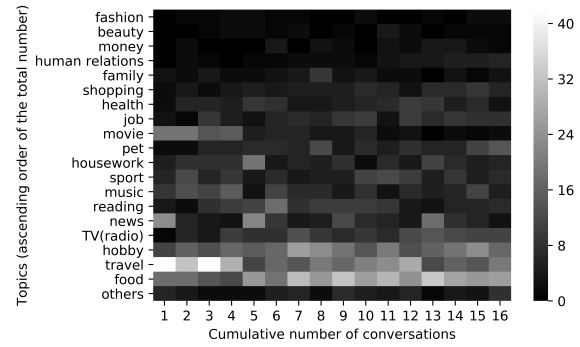


Figure 3: Number of chats on each topic with the cumulative number of conversation

each conversation timing (the cumulative number of conversation). A bright cell indicates that many conversations with its topic occur in its conversation timing. Dark cells are infrequent cells.

Distribution is uneven rather than uniform depending on each interval length and the accumulation amount of the conversations. For example, “news” often appeared after long intervals (ex., cumulative number=4 vs number=5). “Movie” often appeared on the first day (the cumulative number=1, 2, 3, 4) when the talkers are not familiar yet.

4 Conclusion

We examined the effect of multiple conversations on human satisfaction and topic selection in repeated text-chats. Our results suggest that human satisfaction and topic selection are affected by the length of the time intervals between conversations and the accumulation of the dialogue.

References

- Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939.
- Dalmas A Taylor and Irwin Altman. 1966. Intimacy-scaled stimuli for use in studies of interpersonal relations. *Psychological Reports*, 19(3):729–730.
- 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? *arXiv preprint arXiv:1801.07243*.

On Visual Coreference Chains Resolution

Simon Dobnik Sharid Loáiciga

Department of Philosophy, Linguistics, Theory of Science (FLoV)

Centre for Linguistic Theory and Studies in Probability (CLASP)

University of Gothenburg, Sweden

{simon.dobnik, sharid.loaiciga}@gu.se

Abstract

We explore to what degree an existing textual coreference resolution tools can be applied to visual dialogue data. The analysis of error of the coreference system (i) demonstrates the extent to which such data differs from the written document texts that these tools are typically applied on; (ii) informs about the relation between information expressed in language and vision; and (iii) suggests further directions in which coreference tools should be adapted for visual dialogue.

1 Introduction and Related Work

“Situating” dialogue involves language and vision. An important aspect of processing situated dialogue is to resolve the reference of linguistic expressions. The challenging aspect is that descriptions are local to the current dialogue and visual context of the conversation (Clark and Wilkes-Gibbs, 1986) and that not all information is expressed linguistically as a lot of meaning can be recovered from the joint visual and dialogue attention. Coreference resolution has been studied and modelled extensively in the textual domain where the scope of the processing coreference is within a document. Robust coreference resolution for dialogue systems is a very much needed task. The aim of this paper is to provide a preliminary investigation of to what degree an existing off-the-shelf textual coreference resolution tool can be used in the domain of the visual dialogue.

Given its popularity in contexts with scarce amounts of training data, such as dialogue systems, we use the Lee et al.’s 2011 sieve-based system here. For comparison, we also use Clark and Manning’s 2015 mention-pair system. Both are freely available through the Stanford CoreNLP distribution.

Unlike the neatly structured written text which is organised in documents, dialogue data is messy.

The text is structured in turns that are pronounced by different speakers, and sentence boundaries are not clear (cf. Byron (2003) for an overview). Work on referring expressions generation (e.g. Krahmer and van Deemter, 2011; Mitchell et al., 2012; Xu et al., 2015; Lu et al., 2017), on its part, does not typically involve dialogue or the notion of coreference chain – a central construct for coreference resolution systems. Furthermore, coreference resolution tools for dialogue are often custom built to the specific needs of companies or datasets (Rolih, 2018; Smith et al., 2011).

2 Data Processing

The dataset We take the English subsection of the Cups corpus (Dobnik et al., 2015) which consists of two dialogues between two participants with 598 turns in total. The goal of this corpus is to sample how participants would refer to things in a conversation over a visual scene. A virtual scene involving a table and cups has been designed in with a 3-d modelling software (Figure 1). Some cups have been removed from each participant’s view and the participants are instructed to discuss over a computer terminal in order to find the cups that each does not see. The task therefore resembles the Map Task (Anderson et al., 1984, 1991).

Annotation In this pilot study two annotators additionally annotated the first 100 turns of the GU-EN-P1 dialogue for coreference chains. The annotation follows the CoNLL format with the last column containing the coreference chains (Pradhan et al., 2011). Each chain is assigned a number id, where the first and the last tokens of a mention within the chain are identified with opening and closing brackets, as illustrated in Figure 2. In this example, the mentions ‘cups and containers’, ‘some white’, ‘some red’, ‘some yellow’, and ‘some blue’, all belong to the same chain.

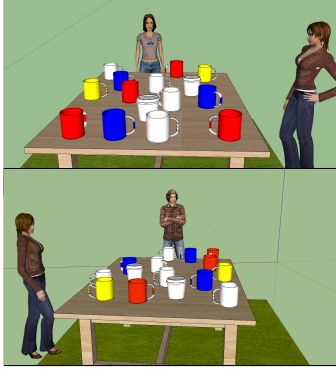


Figure 1: Scene as seen by Participants 1 and 2.

This is the standard scheme used on textual data consisting of documents, but presented two challenges for annotation: (i) in the dialogue data descriptions are made by two conversational participants in turns from their own point of view hence pronouns ‘I’ and ‘you’ as well as spatial descriptions such as ‘from my view’ will have a different referent depending on the speaker; and (ii) a description ‘the red cup’ does not have a unique referent through the dialogue but this changes depending on the previous dialogue states and the focus on the scene. Hence, the annotators also used a visual representation of the scene and descriptions were identified as belonging to the same coreference chain only if they were referring to the same physical object. We assigned fixed ids to all existing objects in the scene (the cups and the table), person A and B, ‘Katie’ and the table as well as frequently used parts of the scene such as B’s-left, Katie’s-right. Dialogue participants also dynamically create ‘objects’ throughout the conversation that they are later referred to as normal objects, e.g. ‘the empty space in front of you’, ‘my white ones (cups)’. For these, annotators introduced additional ids and their approximate location was marked in the representation of the scene.

2.1 Results

We run the annotated data through both the sieve-based and statistical systems from the CoreNLP distribution. Both yielded the exact same output, so our analysis does not distinguish between them.

The official coreference scorer commonly used in the domain searches for complete coreference links, and since the systems were unable to find any of the gold links in our data, the scorer produced appallingly negative results. A major cause behind this inability to identify the coreference

A	1	i	(2)	A	1	some	(5)
A	2	see		A	2	white	5)
A	3	lots		A	3	,	
A	4	of		A	4	some	(5
A	5	cups	(5	A	5	red	5)
A	6	and		A	6	,	
A	7	containers	5)	A	7	some	(5
A	8	on		A	8	yellow	5)
A	9	the		A	9	,	
A	10	table	(4)	A	10	some	(5
				A	11	blue	5)
B	1	me	(1)				
B	2	too					

Figure 2: Annotation of coreference chains

chains accurately lies on the deictic nature of this particular type of dialogue text and the fact that it consists of speaker turns. For instance, the systems grouped all pronouns ‘I’ and ‘me’ into the same chain (and therefore the same entity) because they have identical forms which is a strong feature for assessing coreference in these systems. This problem affects basically all mentions that refer back to some description in a changing context such as ‘my left’ and ‘your left’.

Concerning the parser, a central element to these systems, we observed that the sentences boundaries were identified often correctly (162 versus 157 in the gold), meaning that almost every turn in the dialogue was identified as a sentence. Looking at the mentions, however, from 293 manually annotated mentions distributed over 43 entities, the systems were not able to identify any of them correctly. On the contrary, the systems proposed 88 mentions and 28 entities. Further investigation reveals that a major problem was the correct identification of the mention span. For instance, in one sentence, in the gold the mentions ‘left’ and ‘red mug’ were annotated, but the system identified the maximum spans ‘her left’ and ‘a red mug’ instead. We counted only 12 mention matches due to this problem, yielding a precision of $12 / 88 = 0.14$ and a recall of $12 / 293 = 0.04$.

3 Conclusions

The results of our pilot study show that at least the two coreference resolution systems tested cannot handle visual dialogue data. We expect that our annotations will help us create a data-driven coreference resolution system able to simultaneously model both the language and visual components of this dataset, similar to Kelleher (2006).

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.

References

- Anne Anderson, Gillian Brown, Richard Shillcock, and George Yule. 1984. *Teaching talk: Strategies for production and assessment*. Cambridge University Press, United States.
- 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.
- Kevin Clark and Christopher D. Manning. 2015. [Entity-centric coreference resolution with model stacking](#). 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 1405–1415. Association for Computational Linguistics.
- 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.
- John D Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1-2):21–35.
- Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.
- Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. [Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task](#). In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. ArXiv:1612.01887 [cs.CV].
- Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. 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. Association for Computational Linguistics.
- Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems. Master’s thesis, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden.
- 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.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv*, 1502.03044v3 [cs.LG]:1–22.

Rezonator: Visualizing Resonance for Coherence in Dialogue

John W. DuBois

University of California, Santa Barbara
dubois@ucsb.edu

Abstract

To meet the challenge of understanding coherence in extended dialogue, new methods are needed for analyzing the structure of resonance and engagement in interaction. To advance this work we introduce Rezonator, a powerful new tool designed to support human-in-the-loop annotation of discourse via intuitive, informative, and quantifiable visualizations of multilevel resonance. Rezonator is designed to produce scalable gold standard data via crowdsourcing and gamification. We illustrate with examples to show how interlocutors use multilevel resonance to build a unified structure for alignment, engagement, and coherence in naturally occurring conversation.

1 Introduction

Against the background of the triumphant success of Natural Language Processing and Artificial Intelligence in simulating linguistic behaviors such as question-answering and machine translation, a shadow is cast by the recurrent failure to meet a basic challenge of everyday language use: sustaining coherence in extended dialogue. The deep learning and related techniques that seem to work so well for answering a single question in isolation collapse once the task extends to modeling a sustained, two-way collaborative exchange. Noting the failure of state-of-the-art tools at this task, some leading researchers have called for renewed attention to the problem of coherence in dialogue as a critical frontier in the work of language production and comprehension (Lai & Tetreault, 2018; Li, Monroe, Ritter, & Jurafsky, 2016). A related line of research emphasizes the need for syntax and semantics to come to terms with how conversational participants coordinate their common ground

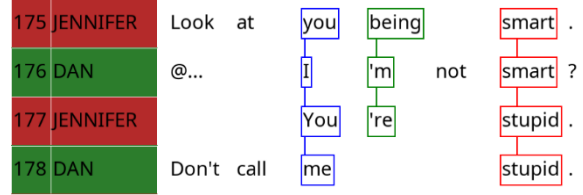


Figure 1: Diagram representation of multilevel resonance in Rezonator. Resonance reflects paradigmatic (*you : I*), inflectional (*being : 'm : 're*), semantic (*smart : stupid*), and other linguistic equivalence classes, including surface word order that often overrides differences in syntactic rules applied. Here, parallel surface order of resonating lexemes aligns a finite main clause (*I'm not smart*) with a non-finite complement clause (*you being smart*). Similarly, a second finite main clause (*you're stupid*) maps onto the reduced syntactic construction of a small clause (*me stupid*).

(Ginzburg, 2012; Gregoromichelaki & Kempson, 2013).

These issues inform the present effort, which introduces Rezonator as a tool designed to support the annotation of multi-level resonance, a key factor in sustaining an attractive mix of coherence, informativeness, and novelty in extended dialogue.

1.1 Resonance

To address these issues, the current approach highlights the critical role that resonance plays in building coherence in extended dialogue. Resonance is defined as “the catalytic activation of affinities across utterances” (Du Bois, 2014, p. 372). Resonance is analyzed within the theory of Dialogic Syntax, which “encompasses the linguistic, cognitive, and interactional processes involved when speakers selectively reproduce aspects of prior utterances, and when recipients recognize the resulting parallelisms and draw inferences from them” (Du Bois, 2014, p. 366). For a quantitative analysis of resonance, see (Moscoso del Prado Martín & Du Bois, 2015).

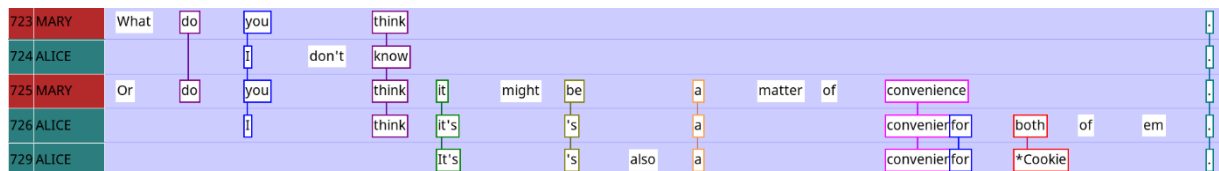


Figure 2: Rezonator representation of multilevel resonance, with structural parallelism of main clauses (*think : know*) and their clausal complements, where structurally aligned contrasts (*might be : it's*) frame the interlocutors' respective stances as relevant to collaborative epistemic problem-solving.

Figure 1 shows an annotation produced using Rezonator, which illustrates how even a brief excerpt may contain affinities at multiple levels of linguistic structure, including paradigmatic (*you : I*), inflectional (*being : 'm : 're*), antonymy (*smart : stupid*), argument structure, and clausal embedding, among others.

Resonances often come in clusters, organized via structural parallelism. This is illustrated again in Figure 2, where the structural parallelism serves to foreground subtle differences in epistemic stance. Such convergence of resonance across lexical, morphological, syntactic, semantic, and pragmatic levels is precisely what Rezonator is designed to study. Rezonator is designed to make it easy for annotators to mark their perceptions of resonance relations at all levels, yielding a rich representation of complex patterns of resonance. Inter-annotator agreement can be assessed by recruiting multiple annotators to independently evaluate the same conversations.

2 Resonance and priming

While evidence for priming seems compelling to many (Bock, 1986; Branigan & Pickering, 2016; Pickering & Ferreira, 2008), controversy remains: Is priming significantly syntactic, or is it merely reducible to lexical priming (Patrick G.T. Healey, Purver, & Howes, 2014)? More troubling is the lack of agreement on the function, if any, of structural priming: Why align? One prominent suggestion holds that priming “makes conversation easy” (Garrod & Pickering, 2004). But broad-spectrum analysis of the full range of syntactic constructions in naturally occurring conversation sometimes yields negative results (Patrick G.T. Healey, Howes, & Purver, 2010; Patrick G.T. Healey et al., 2014). The approach favored here sidesteps the lexical vs. syntax debate by combining the effects of resonance at all linguistic levels, positing a surface-oriented representation of how interlocutors build a single unified alignment structure for resonance and coherence in dialogue.

3 Future development

Because corpus annotation is very labor-intensive, some researchers have sought new ways to incentivize the work, whether through appeals to “citizen science” (Cieri, Fiumara, Liberman, Callison-Burch, & Wright, 2018) or “games with a purpose”, (Habernal et al., 2017; Jurgens & Navigli, 2014; Poesio et al., 2019). Rezonator was designed from the ground up using game design software (GameMaker Studio). This will support our development of “games of resonance” that feel like real games to the players.

For future development, Rezonator stands to benefit from incorporating relevant NLP tools such as word2vec, sense2vec, and pair2vec, several of which are integrated in a recently released toolkit for analyzing linguistic alignment in dialogue, ALIGN (Duran, Paxton, & Fusaroli, 2019).

3.1 Availability

Rezonator is free and open-source software, distributed at <https://rezonator.com> under the MIT license, with source code and documentation at <https://github.com/johnwdubois/rezonator>.

4 Conclusions

In this paper we introduce Rezonator, a tool for representing the complexity of multilevel resonance in dialogue. Rezonator leverages the node-link data structure of the directed acyclic graph to create a unified, holistic, surface-level representation of resonance between utterances. Rezonator further innovates in using gamification to provide new incentives for human-in-the-loop production of gold standard annotations, scalable to crowd-sourced levels suitable for training data, in support of the analysis of naturally occurring conversation. We argue that such explicit, quantifiable representations can help to clarify how interlocutors use multilevel resonance to build a unified structure for alignment, engagement, and coherence in extended dialogue.

Acknowledgements

The author would like to thank Terry DuBois, Georgio Klironomos, and Brady Moore for their inspired contributions to the design and programming of Rezonator and the games of resonance.

5 References

- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355-387.
- Branigan, H. P., & Pickering, M. J. (2016). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, 1-61. doi:10.1017/S0140525X16002028
- Cieri, C., Fiumara, J., Liberman, M., Callison-Burch, C., & Wright, J. (2018, May 7-12, 2018). *Introducing NIEUW: Novel incentives and workflows for eliciting linguistic data*. Paper presented at the Language Resources and Evaluation Conference (LREC 2018), 11th Edition, Miyazaki, May 7-12.
- Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive Linguistics*, 25(3), 359-410. doi:10.1515/cog-2014-0024
- Duran, N. D., Paxton, A. S., & Fusaroli, R. (2019). ALIGN: Analyzing Linguistic Interactions with Generalizable techNiques-a Python Library. *Psychological Methods*, 4, 419-438. doi:10.1037/met0000206
- Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1), 8-11.
- Ginzburg, J. (2012). *The interactive stance: Meaning for conversation*. Oxford: Oxford University Press.
- Gregoromichelaki, E., & Kempson, R. (2013). Grammars as processes for interactive language use: Incrementality and the emergence of joint intentionality. In A. Capone, F. Lo Piparo, & M. Carapezza (Eds.), *Perspectives on linguistic pragmatics* (pp. 185-216): Springer.
- Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., & Gurevych, I. (2017, 19 July 2017). *Argotario: Computational argumentation meets serious games*. Paper presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language
- Healey, P. G. T., Howes, C., & Purver, M. (2010). *Does structural priming occur in ordinary conversation?* Paper presented at the Linguistics Evidence 2010, Tuebingen. <http://www.eecs.qmul.ac.uk/~mpurver/publications.html>
- Healey, P. G. T., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PLoS One*, 9(6), e98598. doi:10.1371/journal.pone.0098598
- Jurgens, D., & Navigli, R. (2014). *It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation*. Paper presented at the Transactions of the Association for Computational Linguistics.
- Lai, A., & Tetreault, J. R. (2018, July 2018). *Discourse coherence in the wild: A dataset, evaluation and methods*. Paper presented at the Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue.
- Li, J., Monroe, W., Ritter, A., & Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *EMNLP*.
- Moscato del Prado Martín, F., & Du Bois, J. W. (2015). *Syntactic alignment is an index of affective alignment: An information-theoretical study of natural dialogue*. Paper presented at the Proceedings of the 37th Annual Conference of the Cognitive Science Society, San Jose, California.
- Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, 134(3), 427-459. doi:10.1037/0033-2909.134.3.427
- Poesio, M., Chamberlain, J., Paun, S., Yu, J., Uma, A., & Kruschwitz, U. (2019, June 2019). *A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation*. Paper presented at the 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), Minneapolis.

Within and Between Speaker Transitions in Multiparty Casual Conversation

Emer Gilmartin

ADAPT Centre, Trinity College Dublin
gilmare@tcd.ie

Carl Vogel

Trinity College Duboin
vogel@tcd.ie

1 Introduction

Casual conversation, ‘talk for the sake of talking’, has been observed to occur in two main phases or sub-genres – interactive *chat* where most or all participants contribute, and more monologic *chunk* phases, where one speaker dominates the conversation, often telling a story or giving an extended opinion (Eggins and Slade, 2004). Previous work has shown differences in the length, composition in terms of speech, silence and overlap, and in the relative frequencies of chat and chunk phases in casual conversation (Gilmartin et al., 2019). In this work we use the timing of speech and silence in chat and chunk phases to explore transitions between single party speech by a speaker and the next stretch of single party speech by the same speaker (*within speaker transition*) or another speaker (*between speaker transition*). We define *ISp* as an interval of single party speech and *ISp1* as a *ISp* of duration one second or more. We also adapt the terminology used in (Heldner and Edlund, 2010) for dyadic interaction. For speakers A and B, within speaker silence (WSS) is defined as **A_GX_A** and between speaker silence (BSS) is defined as **A_GX_B** where GX denotes global silence, while within and between speaker overlap are **A_AB_A** and **A_AB_B**. Thus, *ISp* can transition back to *ISp* with one intervening interval of silence or overlap, e.g. **1_0_1** or **1_2_1**. For multiparty interaction, more possibilities emerge. As multiparty transitions can involve a combination of overlap and silence, we define only two transition types – *within speaker transitions* (WST) beginning and ending with the same speaker, and *between speaker transitions* (BST), which start with one single speaker and transition to another single speaker.

2 Data and Annotation

The CasualTalk dataset is a collection of six 3 to 5 party casual conversations of around one hour each, drawn from the d64, DANS, and TableTalk corpora (Oertel et al., 2010; Hennig et al., 2014; Campbell, 2008).

The data were segmented and transcribed manually and a total of 213 chat and 358 chunk phases were identified and annotated, as described in (Gilmartin and Campbell, 2016). The data were also segmented into 30688 floor state intervals reflecting the participants speaking or silent at any time.

3 Transitions between Single Speakers

For each *ISp1*, we searched forward in the dataset to locate the next *ISp1* and extracted the sequence of intervals (in terms of speaker numbers) from the initial *ISp1* to the next *ISp1*. As an example, **1_2_3_2_1_0_1** contains 5 intervening intervals between the two stretches of *ISp1*.

Distributions of *ISp1*–*ISp1* transitions are shown in Figure 1, where it can be seen that the vast majority of intervening intervals are in stretches of odd numbers of intervals, with the number of cases dropping with increasing intervals. Overall, 95.53% of all *ISp1*-intervals are closed by a later *ISp1* in fewer than 16 intervening intervals. Even-number cases accounted for only 112 (2.1%) of the 5382 transitions between 1 and 15 intervals long. The most frequent class of transitions are those with one intervening interval which account for 41.13% of cases. 21.74% were WSTs while BSTs accounted for 73.78%. For the remaining 4.47% of *ISp1*-intervals, labelled 16+, at least 16 intervals occurred before a second *ISp1* interval was encountered.

In both chat and chunk, disregarding the even-number cases, the number of transitions

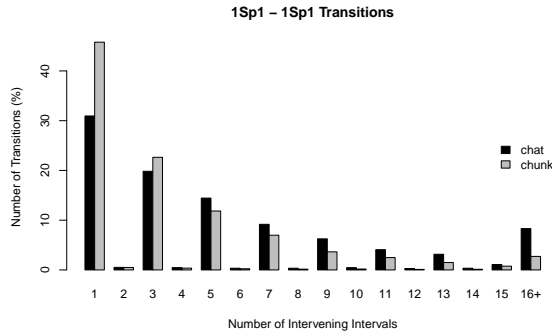


Figure 1: Number of floor state intervals between single-speaker intervals of 1 second or more in duration

declines monotonically with the number of intervening intervals between *1Sp1* intervals. The chat condition starts with a smaller percentage of 1-interval transitions and declines at a lower rate than the chunk condition. In both conditions, it is likely that numbers continue to decline with increasing intervals in a long tail. The 16+ category, a bucket category, is more than three times as large proportionally in chat (8.31%) as in chunk (2.71%).

The odd numbered cases and the 16+ interval bucket class were excluded from the *1Sp1-1Sp1* transition data, leaving 5270 transitions, comprising 77.24% WST and 22.76% BST with intervening intervals ranging from 1 to 15. Figure 2 shows these BST and WST transitions by number of participants, while Figure 3 shows interval types in chat and chunk phases, and the proportion of transitions per interval total. One-interval transitions were the largest group for BST and WST for both chat and chunk, with the proportion of 1-interval transitions particularly high for WST, and very much so in the case of chunk

4 Discussion and Conclusions

The results on transition n-grams between intervals of one speaker speaking in the clear for at least one second (*1Sp1*) show that chat and chunk differ in that between speaker transitions in chat interaction are spread over more intervening intervals than in chunk, thus increasing the frequency of more complex transitions. This could reflect more turn competition, or indeed more backchannels and acknowledgment tokens being contributed by more partic-

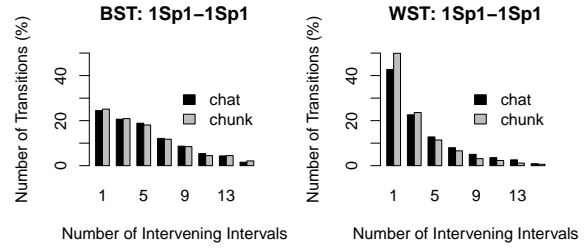


Figure 2: Number of floor state intervals between (*1Sp1*) intervals in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right) in chat and chunk phases.

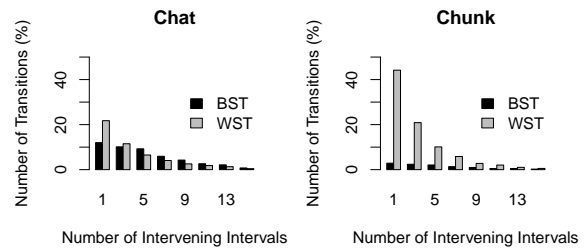


Figure 3: Percentage of Between and Within Speaker Transitions per number of floor state intervals in (*1Sp1-1Sp1*) in chat and chunk phases.

ipants. Within speaker transitions are predominantly one-interval, perhaps reflecting breathing pauses. One-interval transitions comprise the largest class, with a higher proportion of one-interval transitions in chunk than chat, and higher proportions of within speaker than between speaker one-interval transitions in both, but particularly in monologic chunk. However, one-interval transitions only account for 41.03% of transitions overall, reflecting the need to consider more complex transitions around turn change and retention. It would be very interesting to separate within speaker breathing pauses from other transitions in order to better understand transitions around silence. Future work involves further classification of transitions depending on the number of distinct speakers involved, and investigation of the duration of transitions. It is hoped that this study, and similar studies of other corpora, will allow us to inventory transition types in multiparty spoken interaction, and then analyse examples of the statistically more likely transitions in detail to better understand speaker transitions.

Acknowledgments

This work is supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

- N. Campbell. 2008. Multimodal processing of discourse information; the effect of synchrony. In *Universal Communication, 2008. ISUC'08. Second International Symposium on*, pages 12–15.
- S. Eggins and D. Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.
- Emer Gilmartin and Nick Campbell. 2016. Capturing Chat: Annotation and Tools for Multiparty Casual Conversation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- Emer Gilmartin, Benjamin R Cowan, Carl Vogel, and Nick Campbell. 2019. Chunks in multiparty conversation—building blocks for extended social talk. In *Advanced Social Interaction with Agents*, pages 37–44. Springer.
- Mattias Heldner and Jens Edlund. 2010. [Pauses, gaps and overlaps in conversations](#). *Journal of Phonetics*, 38(4):555–568.
- Shannon Hennig, Ryad Chellali, and Nick Campbell. 2014. The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. Reykjavik, Iceland.
- Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2010. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, pages 1–10.

A Wizard of Oz Data Collection Framework for Internet of Things Dialogues

Carla Gordon Volodymyr Yanov David Traum Kallirroi Georgila

Institute for Creative Technologies, University of Southern California

12015 Waterfront Drive, Los Angeles, CA 90094-2536, USA

{cgordon|yanov|traum|kgeorgila}@ict.usc.edu

Abstract

We describe a novel Wizard of Oz dialogue data collection framework in the Internet of Things domain. Our tool is designed for collecting dialogues between a human user, and 8 different system profiles, each with a different communication strategy. We then describe the data collection conducted with this tool, as well as the dialogue corpus that was generated.

1 Introduction

The Internet of Things (IoT) refers to a network of physical devices which are connected to the Internet, and can perform services and provide information to satisfy remote requests. We describe a novel Wizard of Oz (WOz) tool that can be used to investigate several questions relating to how users could communicate in natural language with a Virtual Home Assistant (VHA) that is connected to IoT devices. The tool is designed to address several issues for this kind of dialogue, including relaying aspects of the environmental context to the user and testing different communication styles.

When interacting with a VHA, a user will typically be inside a home and will know what room they are in, what devices exist, and may be able to see or hear changes in a device's state if they are in the same room. Therefore, in order to generate realistic dialogues in this domain, there needs to be some environmental context provided to the user. Our WOz tool allows us to provide this context (see Figure 1).

Additionally, previous analysis of observer ratings of IoT dialogues authored by linguists (Georgila et al., 2018; Gordon et al., 2018) suggested several features of VHA interaction that may affect user satisfaction. The tool allows us to define system profiles each with a different communication style (with arbitrary system names shown to the user).

2 System Profiles

We examine 3 binary system behavior features: Register (Direct, e.g., "Thank you.", or Conversational, e.g., "Thanks, it's my pleasure to help."), Explicitness (Explicit, e.g., "I've turned on the light in the kitchen.", or Implicit, e.g., "Your request has been taken care of."), Errors (misunderstandings exist or not). Combining these 3 features leads to 8 different system profiles. The wizard interface allows the wizard to toggle between profiles, each with a different set of utterances that conform to the system behavior features.

3 The WOz Tool

The WOz tool includes 3 different views: one for the user, and two for the wizard. The user view can be seen in Figure 1. It shows 3 rooms (Bedroom, Kitchen, Living Room) and indicates what room the participant is currently in. In each room all devices in that room are displayed, along with information about the current device state (on/off), as well as other relevant information for that device (e.g., current channel, temperature, volume).

Below the display of rooms and devices is the text chat window in which users can enter their commands to the system. A log of all system and user utterances is available to the user at all times when interacting with the interface. The system profile identities are displayed to the user in the upper right corner of the user interface and by the name in the log (the monkey in Figure 1), but no explicit description is given of the features associated with the profile. The wizard interface includes one view containing buttons to communicate with the user and change the state of the devices, and a second view with information on the state of all the devices.



Connect bluetooth for the light in the kitchen.

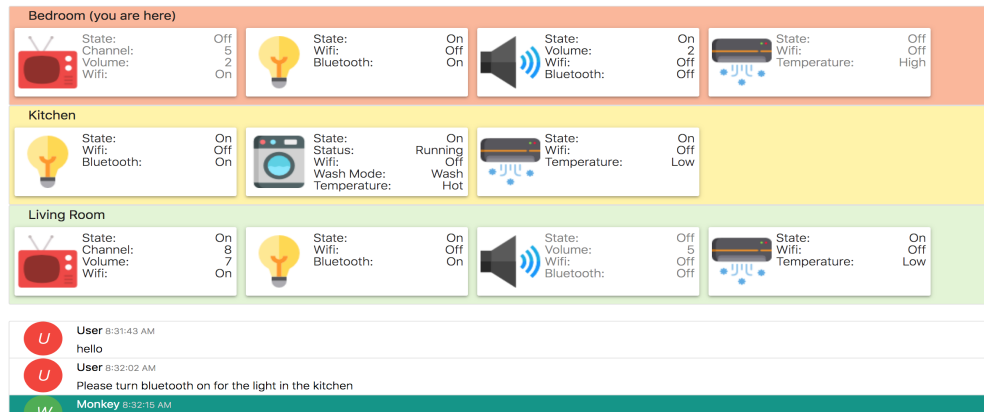


Figure 1: The user interface view

4 Data Collection

Users were recruited via Craigslist (<https://www.craigslist.org>), and interaction with our WOz tool took place on a Macbook Pro laptop, in a room without any experimenters present. After filling out a short survey, participants typed commands in text to accomplish 12 different tasks such as turning on a light. Each user interacted with only 4 of the 8 system profiles (3 dialogues per profile). Users had to accomplish their task before they could move on to another task; if a user tried to advance to the next task before accomplishing the current one, the wizard would send a message that the task was not complete.

The WOz tool provided the environmental context necessary for users to understand whether a task had been completed correctly, even without explicit confirmation from the system, as exemplified by Table 1. The system initially turns on the AC in the wrong room, but because of the context provided to the user they were able to see the AC in the kitchen was not turned on like they asked, and repeated their command to complete the task.

Users were asked to fill out a post-task Likert-scale survey designed to measure the level of user satisfaction, as well as whether the user perceived the system as friendly, natural, intelligent, or as having a personality. There was also a post-experiment survey administered once the user had completed all tasks, in which users were asked to rank the 4 systems they interacted with from best

to worst.

This data collection yielded a corpus of 216 human-system dialogues (18 users, 12 dialogues per user, 27 dialogues per system profile), each with accompanying Likert-scale survey data. This corpus will be used to further investigate the interaction between these subjective features, the system behaviors, and the user's overall satisfaction.

Actor	Text	Action
User	Please turn on the air conditioner in the kitchen.	
System	Ok, I will do that.	AC on in Bedroom
System	It has been done.	
User	Please turn on the air conditioner in the kitchen.	
System	I'm sorry, I will.	AC on in Kitchen
User	Thank you.	
System	You're welcome.	

Table 1: Environmental context example

Acknowledgments

This work was supported by Samsung Electronics Co., Ltd. and the U.S. Army. Any statements or opinions expressed in this material are those of the authors and do not necessarily reflect the policy of the U.S. Government, and no official endorsement should be inferred.

References

- Kallirroi Georgila, Carla Gordon, Hyungtak Choi, Jill Boberg, Heesik Jeon, and David Traum. 2018. Toward low-cost automated evaluation metrics for Internet of Things dialogues. In *Proceedings of the 9th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, Singapore.
- Carla Gordon, Kallirroi Georgila, Hyungtak Choi, Jill Boberg, and David Traum. 2018. Evaluating subjective feedback for Internet of Things dialogues. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (SemDial:AixDial)*, Aix-en-Provence, France.

Normativity, Meaning Plasticity, and the Significance of Vector Space Semantics

Eleni Gregoromichelaki
Heinrich-Heine University
elenigregor@gmail.com

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

Arash Eshghi
Heriot-Watt University
a.eshghi@hw.ac.uk

Ruth Kempson
King's College London
ruth.kempson@kcl.ac.uk

Mehrnoosh Sadrzadeh
University College London
m.sadrzadeh@ucl.ac.uk

Julian Hough Matthew Purver Gijs Wijnholds
Queen Mary University of London
{j.hough,m.purver,g.j.wijnholds}@qmul.ac.uk

Abstract

This paper continues the discussion started in (Lücking et al., 2019), on the suitability or otherwise of Vector Space Semantics (VSS) as a model of semantics for NL in interaction.

1 Introduction

Lücking et al. (2019) argue that the *distributional hypothesis* (DH) cannot lead to a psychologically realistic modelling of natural language (NL) due to its inability to stand as an autonomous basis for semantics. Instead, they propose a model of the conceptual mechanisms underpinning NL interaction involving direct encoding of conceptual structures and processes on individual brains (Cooper, 2019). Problems of agent coordination are then resolved as meaning negotiation and learning based on game-theoretic modelling of symbolic signalling that presupposes mental states with hardwired discourse structure (DGBs).

We find many points of agreement with Lücking et al. (2019). However, we believe that not all versions of implementing DH/VSS fall under their criticism. Although in the past most operationalisations of DH have involved only word distributions, the recent multimodal trend involves not only textual but also image and even audio contexts (e.g. Kiela and Clark, 2017; Bruni et al., 2014). Indeed, from early on, such models have envisaged their extension to distributional representations that include situational contexts (see e.g. Landauer and Dumais, 1997, a.o.) and, in our view, at least the combination of Dynamic Syntax (DS; Kempson et al., 2001, 2016) and VSS (DS-VSS, Kempson et al., 2019; Sadrzadeh et al., 2018; Wijnholds et al., 2019) operates under assumptions resolving the issues the authors raise.

Instead of employing individualistic referential mechanisms, DS proposes that semantic content emerges in interaction rather than in the correspondence of representations in the brain to entities in the world (Gregoromichelaki, 2019; Gregoromichelaki et al., 2019). Hence, the structures manipulated by DS constitute complex,

highly-structured predictive triggers (*affordances*) for further verbal/nonverbal actions. This idea has been computationally implemented in DS-TTR (Eshghi et al., 2017; Kalatzis et al., 2016; Eshghi and Lemon, 2014) where, in a Reinforcement Learning model, it is Record Types (RTs) of Type Theory with Records (TTR; Cooper, 2005; Cooper and Ginzburg, 2015) that are the triggers for further action (dialogue contexts): clusters of RTs are learned from interaction histories and, accordingly, a potential next response is chosen. But, under the same assumptions, the DS-VSS integration appears to be equally suitable for the same purpose, especially since it would appear to better capture the nondiscrete, gradient effects associated with such triggers.

On both the DS-TTR and DS-VSS views, normative semantic attributions do not concern facts about individuals but relational facts about characterisations of transactions of individuals with the sociomaterial environment. Meaning then arises in interaction, on the basis of affordances made available to agents by sociomaterial settings ('forms of life') that groups establish to direct their perceptual and action capacities. In concrete situations, agents selectively engage with multiple affordances available in such *affordance landscapes* (Rietveld et al., 2018; Bruineberg et al., 2018). These socially-established affordances constitute a general basis of normativity both for action/perception and NL meaning, in that individual agents can have partial or imperfect grasp of such potentials depending on their level of expertise. This is because individuals engage with affordances through the experience of *solicitations* (Dreyfus and Kelly, 2007): agents have abilities, dispositions, and concerns regarding their interactions which define the saliency of particular affordances in concrete situations; and individual abilities and values are acquired through histories of interactions in particular settings.

This, we suggest, is where the aptness of DH and VSS tools lies. In combination with DS, such models can be seen as implementing exemplar ac-

counts of categorisation (Nosofsky, 2011) in that the matrix representations record episodic memories of contexts of perception/action involving particular stimuli (here, *words*). Word forms in DS trigger sets of incremental actions and predictions; and past experiences with such stimulus-situation pairs is what is stored and retrieved in processing. Past co-occurrence, “similarity” relations, can then underpin associationist and probabilistic mechanisms of online selective attention (*affordance-competition*) that result in incrementally appropriate word retrieval (via activation facilitation) in production and contextualisation (narrowing-down or enrichment) in comprehension. Thus the significance of words emerges from joint (re)constructive acts during use: runtime operations over high-dimensional VS representations (e.g., context-aware analogy cf. Landauer, 2002) enable agents to engage with probabilistic distributions over fields of predictions of further opportunities for action thus grounding normativity in local exchanges. On the other hand, abstractions underpinning *explicit* normative judgements, e.g., truth-conditional judgements, reference, grammaticality etc, are phenomena definable only at a historical and group level of analysis, “bootstrapped” from more basic, domain-general psychological capacities, and do not play a fundamental grounding role in NL performance.

On this view, then, individual agents’ memories do not store transductions of perceptual input into symbolic conceptual representations (cf. Larsson, 2015). Instead, conceptual capacities are abilities to discriminate alternative responses to similar or dissimilar stimuli arrays (cues). Classical theories of learning, like reinforcement or discriminative learning (Rescorla and Wagner, 1972), can then be employed to model the constantly evolving fluid responsiveness to NL stimuli, even highly underspecified ones, like indexical pronouns, wh-elements, and names. For example, in learning the distinction between the English words *I*, *you*, and *he/she*, infants are initially expected to display inconsistencies and individual differences depending on their personal experience with input, as they will not be “attuned” sufficiently to the ambient social invariances that license the use of each form. Recorded episodes of experience with pronouns as cues for action initially will be too few and too restricted to enable development of speaker/addressee/non-participant discriminatory features to ‘solicit’ the affordances that characterise appropriate pronominal usage. But, in the face of discrepancies between their own predictions and actual experience, the infant will gradually come to discriminate salient aspects of the discourse environment to serve as cues for the choice of form. Such a shift only becomes possible, however, if there are options available, namely, suf-

ficiently “similar” competing cues (in DS, *triggers* of actions) that occur in similar contexts (language games) like the various alternative forms of pronouns. Such triggers compete with each other on the basis of their predictive value regarding subsequent events (in DS, further opportunities for rewarding interaction or avoidance of undesirable consequences). Competition means that loss of associative strength by one cue results in reinforcement of the other(s) in the same category ensuing in an emergent systematic pattern of contrasts (Rescorla and Wagner, 1972). Moreover, given that lexical triggers are necessarily fewer than discourse situations/features, the same forms can come to acquire added triggering effects by the same process, i.e., contextual co-occurrence overlap and subsequent discrimination on the basis of prediction error: for example, *you* can come to include or not multiple addressees in multiparty dialogue, or acquire an impersonal use that might include the speaker when the combination of contextual features are sufficiently discriminative. Such cases cannot be handled easily by the model presented in Lücking et al. (2019) because these uses underdetermine, disregard, or eliminate the hard-wired distinctions postulated in their DGB-based modelling with arbitrary homonymies appearing as the only available solution.

Similarly, regarding proper names, it is storage of life episodes incidentally involving particular interlocutors that resolves the problem of “referential” uncertainty by means of relying solely on domain-general memory mechanisms rather than specific assumptions about conceptual/discourse structure. In fact, Gregoromichelaki et al. (2011) argue that “mindreading” effects can be accounted for exactly because of such co-occurrence mechanisms that employ names as cues for invoking past interactions with discourse participants to ground appropriate redeployment (Horton and Gerrig, 2005), rather than assuming explicit representations of common ground or metarepresentational reasoning.

Overall then, given DS-VSS modelling of both word meaning and syntax alike as (socio)-cognitive predictive and incremental mechanisms, compositional VSS employing tensors and tensor contraction provides a fruitful implementation of exemplar-based categorisation, thus modelling the emergence of NL polysemy, as well as ‘ad hoc concept’ enrichment/narrowing effects, which otherwise remain a mystery (Partee, 2018). Without such an extension of our theoretical vocabulary (Sadrzadeh et al., 2018), we believe that progressive achievement of NL acquisition, the emergent fluency of conversational exchange not only with familiars but in arbitrary multiparty exchanges, and the inexorability of NL change all threaten to continue to elude us (Kempson et al., 2018).

References

- Bruineberg, J., E. Rietveld, T. Parr, L. van Maanen, and K. J. Friston (2018). Free-Energy Minimization in Joint Agent-Environment Systems: A Niche Construction Perspective. *Journal of Theoretical Biology* 455, 161–178.
- Bruni, E., N. K. Tran, and M. Baroni (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research* 49, 1–47.
- Cooper, R. (2005). Records and record types in semantic theory. *Journal of Logic and Computation* 15(2), 99–112.
- Cooper, R. (2019, June). Representing Types as Neural Events. *Journal of Logic, Language and Information* 28(2), 131–155.
- Cooper, R. and J. Ginzburg (2015). Type theory with records for natural language semantics. *The Handbook of Contemporary Semantic Theory*, 375–407.
- Dreyfus, H. and S. D. Kelly (2007, March). Heterophenomenology: Heavy-Handed Sleight-of-Hand. *Phenomenology and the Cognitive Sciences* 6(1), 45–55.
- Eshghi, A. and O. Lemon (2014). How domain-general can we be? Learning incremental dialogue systems without dialogue acts. In *Proceedings of Semdial 2014 (DialWatt)*.
- Eshghi, A., I. Shalymov, and O. Lemon (2017). Bootstrapping incremental dialogue systems from minimal data: the generalisation power of dialogue grammars. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Gregoromichelaki, E. (2019). Natural Languages as Distributed Action Systems. In *Third Dynamic Syntax Conference*, Valletta, Malta. May 2019.
- Gregoromichelaki, E., C. Howes, and R. Kempson (2019). Actionism in syntax and semantics. In *Dialogue and Perception*. CLASP Papers in Computational linguistics. under review.
- Gregoromichelaki, E., R. Kempson, M. Purver, G. J. Mills, R. Cann, W. Meyer-Viol, and P. G. T. Healey (2011). Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse* 2(1), 199–233.
- Horton, W. and R. Gerrig (2005). Conversational common ground and memory processes in language production. *Discourse Processes* 40(1), 1–35.
- Kalatzis, D., A. Eshghi, and O. Lemon (2016). Bootstrapping incremental dialogue systems: using linguistic knowledge to learn from minimal data. In *Proceedings of the NIPS 2016 workshop on Learning Methods for Dialogue*, Barcelona.
- Kempson, R., R. Cann, E. Gregoromichelaki, and S. Chatzikyriakidis (2016). Language as mechanisms for interaction. *Theoretical Linguistics* 42(3-4), 203–276.
- Kempson, R., E. Gregoromichelaki, and C. Howes (2018). Language as mechanisms for interaction: Towards an evolutionary tale. In A. Silva, S. Staton, P. Sutton, and C. Umbach (Eds.), *Proceedings of the 12th TbiLLC Conference*, pp. 209–227.
- Kempson, R., J. Hough, C. Howes, M. Purver, P. G. T. Healey, A. Eshghi, and E. Gregoromichelaki (2019). Why natural language models must be partial and shifting: a Dynamic Syntax with Vector Space Semantics perspective. In *Vector Semantics for Dialogue and Discourse workshop at IWCS 2019*.
- Kempson, R., W. Meyer-Viol, and D. Gabbay (2001). *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.
- Kiela, D. and S. Clark (2017, December). Learning Neural Audio Embeddings for Grounding Semantics in Auditory Perception. *Journal of Artificial Intelligence Research* 60, 1003–1030.
- Landauer, T. K. (2002, January). On the computational basis of learning and cognition: Arguments from LSA. In *Psychology of Learning and Motivation*, Volume 41, pp. 43–84. Academic Press.
- Landauer, T. K. and S. T. Dumais (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2), 211–240.
- Larsson, S. (2015). Formal Semantics for Perceptual Classification. *Journal of Logic and Computation* 25(2), 335–369.
- Lücking, A., R. Cooper, S. Larsson, and J. Ginzburg (2019). Distribution is not enough: going Further. In *NLCS - Natural Language and Computer Science 6 workshop at IWCS 2019*.
- Nosofsky, R. M. (2011). The Generalized Context Model: An Exemplar Model of Classification. In *Formal Approaches in Categorization*. Cambridge University Press.
- Partee, B. H. (2018). *Changing notions of linguistic competence in the history of formal semantics*, pp. 172–196. Oxford University Press.
- Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* 2, pp. 64–99. Appleton-Century-Crofts.
- Rietveld, E., D. Denys, and M. Van Westen (2018). Ecological-Enactive Cognition as engaging with a field of relevant affordances. In *The Oxford Handbook of 4E Cognition*, pp. 41. Oxford University Press.
- Sadrzadeh, M., M. Purver, J. Hough, and R. Kempson (2018, November). Exploring Semantic Incrementality with Dynamic Syntax and Vector Space Semantics. In *Proceedings of the 22nd SemDial Workshop on the Semantics and Pragmatics of Dialogue (AixDial)*, Aix-en-Provence, pp. 122–131.
- Wijnholds, G., M. Purver, M. Sadrzadeh, and R. Kempson (2019). Incremental Semantic Judgements. In *3rd Dynamic Syntax Conference*, Valletta, Malta. May, 2019.

Comparing Cross Language Relevance vs Deep Neural Network Approaches to Corpus-based End-to-end Dialogue Systems*

Seyed Hossein Alavi and Anton Leuski and David Traum

Institute for Creative Technologies

University of Southern California

{seyedhoa@ict.usc.edu and leuski@ict.usc.edu and traum@ict.usc.edu}

Abstract

We compare two models for corpus-based selection of dialogue responses: one based on cross-language relevance and a cross-language LSTM model. Each model is tested on multiple corpora, collected from two different types of dialogue source material. Results show that while the LSTM model performs adequately on a very large corpus (millions of utterances), its performance is dominated by the cross-language relevance model for a more moderate-sized corpus (ten thousands of utterances).

1 Introduction

End-to-end neural network models of conversational dialogue have become increasingly popular for conversational tasks (e.g., (Ritter et al., 2011; Serban et al., 2015; Zhao et al., 2017)). These models eschew traditional modeling approaches that include internal hand-crafted domain models and representations of dialogue context and multimodal input signals, and separate components for understanding natural language (converting to the internal representation language), updating dialogue state, state-based response generation, and natural language generation (e.g., (Traum and Larsson, 2003; Raux et al., 2005; Nasihati Gilani et al., 2018)). Instead, these models learn to respond directly from a corpus, either by generating new responses or selecting a response from the corpus training data, using dual encoding and hidden layers to learn appropriate dialogue continuations. However, there are still a number of questions remaining about how well such models really work for real applications, and how much data is needed to achieve acceptable performance. Other

machine learning approaches have been shown to be useful, with much smaller datasets.

In this paper, we compare two different kinds of end-to-end system, a neural network model based on (Lowe et al., 2015) and an older kind of end-to-end dialogue model, based on cross-language retrieval (Leuski et al., 2006), implemented in the publicly available NPCEditor (Leuski and Traum, 2011), and previously used for systems that have been displayed in museums (Traum et al., 2012, 2015). We compare these models on two different datasets: the Ubuntu Corpus (Lowe et al., 2015), and one derived from one of the museum system datasets (Traum et al., 2015).

2 Datasets and models

We utilized a number of datasets in our experiments to compare NPCEditor with a deep neural network model. The *Ubuntu Dialogue corpus* (Lowe et al., 2015) was constructed from Linux support message boards, where people posted problems and solutions. It contains 1 million multi-turn dialogues, with a total of over 7 million utterances and 100 million words. The training set has 50% relevant and 50% irrelevant pairs of < context, response >. In the development set, for a given context it has 1 relevant response and 9 distractors (irrelevant responses).

We constructed three other datasets out of the data made available from the system described in (Traum et al., 2015). Pinchas_10 consists of 33350 samples for the training set, 50% of which are negative samples and the rest are positive. In the development and test sets, for each question, there is a relevant response and 9 randomly selected non-relevant responses. (Similar to the dev and test sets in the Ubuntu corpus)

Pinchas_1444 is constructed to investigate how the models would perform on a task inspired by a

This work was supported by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

real problem (Traum et al., 2015) in which we may have more than one thousand possible responses gathered from interviews. The training set is created similar to Pinchas_10. Nonetheless, for the development and test sets, instead of 10 distractors, we used the whole set of possible responses. Another important difference between Pinchas_1444 and Pinchas_10 is that in this new set there might be more than one relevant response for a given question. Given that very few of the 1444 responses are appropriate for any given question, showing an even number of positive and negative examples might inappropriately prefer recall over precision. In a second version, Pinchas_1444_v2, we increased the negative samples in the training set from 50% to 90%.

The first model we test is NPCEditor (Leuski and Traum, 2011), which was used for the system in (Traum et al., 2015). At the core of NPCEditor is a statistical regression approach based on cross-lingual language model proposed by Lavrenko for cross-lingual information retrieval (Lavrenko, 2004). Leuski and Traum successfully adopted his approach to question answering and applied it in many different applications (Leuski and Traum, 2008, 2011).

From the pool of previous deep neural net models, such as (Hochreiter and Schmidhuber, 1997), (Olabiya et al., 2018), (Shao et al., 2017), (Zhou et al., 2018), (Zhang et al., 2018), (Devlin et al., 2018), (Mehri and Carenini, 2017), we chose the Dual encoder model first introduced by (Lowe et al., 2015). We trained the model with the same parameters that (Lowe et al., 2015) did.

3 Experiments and Evaluation

We conduct a series of experiments to compare the NPCEditor and the Dual-Encoder model. Following (Lowe et al., 2015), we use R@k as the evaluation metric, which is the percentage of times that the expected response is retrieved in the top-k responses. R@1 is equivalent to accuracy. We first test the Dual-Encoder model on both the Ubuntu corpus (to compare with the model in (Lowe et al., 2015), as a sanity check on the implementation), and on the Pinchas_10 dataset, which has a test-set parallel in structure to Ubuntu. Next we compare the NPCEditor and the Dual-Encoder model on the Pinchas_10 dataset. Then we compare the performance of the NPCEditor and Dual-Encoder model on Pinchas_1444_v1 and Pinchas_1444_v2

datasets.

Dataset	Pinchas_10		Ubuntu
Model	NPCEditor	DE	DE
1 in 10 R@1	0.78	0.64	0.60
1 in 10 R@2	0.84	0.83	0.74
1 in 10 R@5	0.92	0.97	0.92

Table 1: Results from the experiment 1 and 2 using various R@k measures.

Pinchas_1444	v2		v1
Model	NPCEditor	DE	DE
1 in 1444 R@1	0.7663	0.1238	0.0625
1 in 1444 R@2	0.8175	0.1939	0.1305
1 in 1444 R@5	0.8758	0.3089	0.2392
1 in 1444 R@10	0.9106	0.4217	0.3441

Table 2: Results from experiment 3 and 4.

4 Results

Experiment 1 showed that the Pinchas data appears easier than the Ubuntu data - with a much smaller training set size, the Dual-Encoder model was able to improve on R@k in the Pinchas_10 dataset compared to the Ubuntu dataset. Experiment 2 showed that given the amount of available training data (10s of thousands of examples), the NPCEditor significantly out-performs the Dual-Encoder model in R@1 on this data set. Experiment 3 showed that the results are even more striking for a more real-world example, where the system’s task is to pick the best response out of a set of over 1000 available. Here, the Dual-Encoder model does not perform well enough to engage in a meaningful dialogue, while the NPCEditor performs similarly to results reported in (Traum et al., 2015), which led to much-reported user engagement. The improved performance of the Pinchas_1444_v2 training set, with a much higher proportion of negative examples, does perhaps point to a direction for improvement. Future work should perhaps look at the even higher distribution of negative to positive examples.

These results do show that despite the recent popularity of deep learning models, there is still a place for more traditional machine learning algorithms, that can operate well on more moderate-sized data sets for problems of interest. It may also be the case that different types of dialogue have different optimal models. For example, (Gandhe and Traum, 2010) show very different upper bounds for retrieval approaches to dialogue in different domains/datasets.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Sudeep Gandhe and David Traum. 2010. I’ve said it before, and i’ll say it again: an empirical investigation of the upper bound of the selection approach to dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 245–248. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Victor Lavrenko. 2004. *A Generative Theory of Relevance*. Ph.D. thesis, University of Massachusetts at Amherst.
- Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27.
- Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In *Proceedings of the 26th Army Science Conference*, Orlando, Florida, USA.
- Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. [The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems](#). *CoRR*, abs/1506.08909.
- Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *IJCNLP*.
- Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal dialogue management for multiparty interaction with infants. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 5–13. ACM.
- Oluwatobi Olabiyi, Alan Salimov, Anish Khazane, and Erik T. Mueller. 2018. [Multi-turn dialogue response generation in an adversarial learning framework](#). *CoRR*, abs/1805.11752.
- Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let’s go public! taking a spoken dialog system to the real world. *Proceeding of the International Speech Communication Association*.
- Alan Ritter, Colin Cherry, and William B. Dolan. 2011. [Data-driven response generation in social media](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11*, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. [Hierarchical neural network generative models for movie dialogues](#). *CoRR*, abs/1507.04808.
- Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. [Generating long and diverse responses with neural conversation models](#). *CoRR*, abs/1701.03185.
- David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and grace: Direct interaction with museum visitors. In *International conference on intelligent virtual agents*, pages 245–251. Springer.
- David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 2015. [Evaluating spoken dialogue processing for time-offset interaction](#). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 199–208, Prague, Czech Republic. Association for Computational Linguistics.
- David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer.
- Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. [Modeling multi-turn conversation with deep utterance aggregation](#). *CoRR*, abs/1806.09102.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. [Learning discourse-level diversity for neural dialog models using conditional variational autoencoders](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.
- Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. [Multi-turn response selection for chatbots with deep attention matching network](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

Collection and Analysis of Meaningful Dialogue by Constructing a Movie Recommendation Dialogue System

Takashi Kodama
Kyoto University

Ribeka Tanaka
Kyoto University

Sadao Kurohashi
Kyoto University
NII CRIS

{kodama, tanaka, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 Introduction

Intelligent dialogue systems must be able to produce the utterance that match to the previous utterance of the conversational partner, as well as the *dialogue context* more broadly. This context can be influenced by former utterances, dialogue participants intentions, focus of interests, and so many other factors.

Especially, in human-human dialogues, dialogue participants have some purposes with the dialogue even in chat dialogues, e.g., telling some information and influencing the partner. Based on the purposes, we humans maintain a coherent dialogue for multiple turns, take care of what the partner thinks, and behave appropriately based on the former conversation. Handling such factors that govern the multiple dialogue turns is one of the biggest issues towards a better dialogue system.

In order to address this issue, it is essential to analyze multi-turn dialogue data. However, because the recent studies on chat dialogues have paid most attention to dealing with the variety of user utterances in open-domain chit-chat, the resultant dialogues are locally appropriate but not suitable for context-level analysis.

In this study, we propose a new large human-machine dialogue dataset for Japanese. We constructed a scenario-based movie recommendation dialogue system and collected 1060 dialogues by crowdsourcing. Figure 1 shows an example of collected dialogues¹. The system starts the dialogue with a movie-related topic such as news topic, introduces a related movie, and presents some recommendation points. The system produced mostly natural dialogues, while it also raised some context-level errors that are worth considering toward a better dialogue system.

¹Examples of dialogues presented in this paper are originally in Japanese and were translated by the authors.

S1: It is a hot topic that it was revealed by the 11th that the second child was born to actor Shota Someya and actress Rinko Kikuchi.
U1: That's a good news.
S2: Shota Someya appeared in the movie called "The Lion in March: Part 1."
U2: Right.
S3: There is a way of attacking in *Shogi* for each players, and their face express the conflict and distress in the game. You will be moved and nailed to the picture.
U3: I want to see it once.
S4: The performance of the casts is the best, so the video and music are at a satisfactory level.
U4: I'm interested in it.
S5: I think you can enjoy it.

Figure 1: Example of collected dialogue (S is the system and U is the user).

2 System Configuration

Figure 2 shows an overview of our proposed system. There are two main components in our system: a movie database and a dialogue manager. A movie database stores information of 213 movies acquired by web crawling. For each movie, we prepare the system's utterances beforehand as a *scenario* using the information of the movie database.

To produce natural dialogue, we created scenarios based on the three strategies: (1) the system starts the dialogue by introducing the surrounding topic and later presents the title of the recommended movie; (2) sentences that describe recommended points of the movie are uttered multiple times; (3) in principle, the system utters according to the scenario, as if it takes the dialogue initiative. Since the last point can be a cause of disagreement, we designed the dialogue manager to answer easy questions and to react to some typical user utterances.

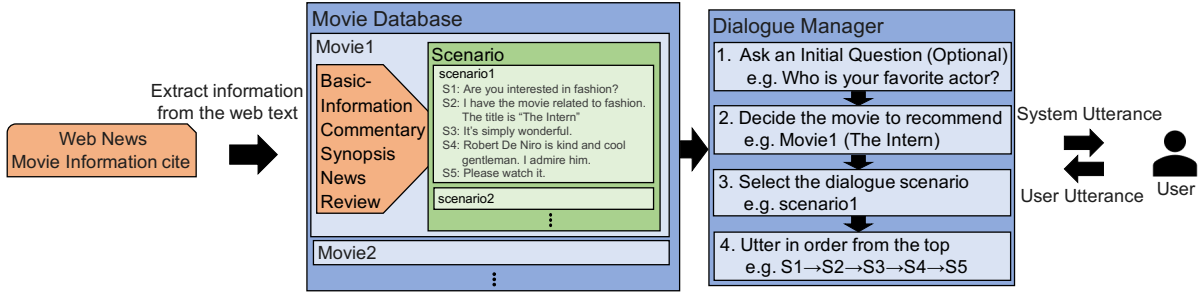


Figure 2: The overview of movie recommendation dialogue system

Based on the three strategies, we created one or more scenarios for each movie. In the first part of the scenario, the system started with movie-related topic and presented recommended movies. In the remaining part, the system made two utterances that describe the recommendation points and finally prompted the user to watch the movie. We extracted recommendation points from the user reviews of a film review website and converted them to appropriate syntactic form by some rules.

Which movie to recommend as well as which scenario to use are decided by the dialog manager. It drives the dialogue based on each scenario by referring to the movie information if necessary.

3 Dialogue Collection and Analysis

We collected dialogue by our movie recommendation system on crowdsourcing. After each dialogue, we asked the worker to answer some questionnaire. We selected the following four referring to the evaluation metrics adopted by Hiraoka et al. (2013); Li et al. (2018); Yoshino et al. (2018).

- (1) **Potential Interest:** Do you like movies?
- (2) **Watching Experience:** Have you seen this recommended movie?
- (3) **Persuasiveness:** Do you want to see this recommended movie?
- (4) **Naturalness of Flow:** Was the flow of dialogue natural?

The result we got suggested that our scenario-based system was able to produce natural utterances that keep the dialogue purpose.

We labeled each of the system utterances in the collected dialogues whether it is natural in the concerned dialogue context. We conducted this annotation task also on crowdsourcing. For each system utterance except the first utterance, we asked three workers to annotate it with one of the three labels, *Natural*, *Possibly Unnatural*, and *Unnatural*. When an utterance is judged as unnatural by

Main category	H+	Ours
Utterance-level	12.7%	4.1%
Response-level	51.1%	41.6%
Context-level	29.9%	54.3%
Environment-level	6.3%	0.0%

Table 1: Distribution of error categories (H+ is numbers obtained by Higashinaka et al. (2015a)).

multiple annotators, we further classified its error type. We adopted the error taxonomy provided by Higashinaka et al. (2015a,b) and we further classified the error types for some subcategories based on the observed cases. As they proposed, we also distinguished the four hierarchical levels, namely, *utterance-level*, *response-level*, *context-level*, and *environment-level* (see Higashinaka et al. (2015a) for details). Table 1 shows the distribution of error categories in dialogue data proposed by Higashinaka et al. (2015a) as well as in dialogue data we collected. Although the direct comparison is difficult, our data contains less utterance-level errors and more context-level errors compared to dialogue collected by Higashinaka et al. (2015a). This suggests dialogues we collected contain more errors concerned with dialogue context, which are worth analyzing in more depth towards a better dialogue system.

4 Conclusion

We proposed a human-machine dialogue collection in Japanese. The dialogue data will be made available for research use on a request basis. Future research will have to investigate more on the cause of contextual errors and the way to avoid the unnatural utterances.

Acknowledgments

This research was supported by NII CRIS Contract Research 2019.

References

- Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015a. [Towards taxonomy of errors in chat-oriented dialogue systems](#). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 87–95, Prague, Czech Republic. Association for Computational Linguistics.
- Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015b. [Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Lisbon, Portugal. Association for Computational Linguistics.
- Takuya Hiraoka, Yuki Yamauchi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Dialogue management for leading the conversation in persuasive dialogue systems. In *Proceedings of IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 114–119. IEEE.
- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. [Towards deep conversational recommendations](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9725–9735. Curran Associates, Inc.
- Koichiro Yoshino, Yoko Ishikawa, Masahiro Mizukami, Yu Suzuki, Sakriani Sakti, and Satoshi Nakamura. 2018. [Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing](#). In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Towards Finding Appropriate Responses to Multi-Intents - SPM: Sequential Prioritisation Model

Jakob Landesberger

Speech Technology, Daimler AG
jakob.landesberger@daimler.com

Ute Ehrlich

Speech Technology, Daimler AG
ute.ehrlich@daimler.com

Abstract

Speech is an easily accessible and highly intuitive modality of communication for humans. Maybe that is the reason why people have wanted to talk to computers almost from the moment the first computer was invented. Today several consumer-level products developed in the last few years have brought inexpensive voice assistants into everyday use. The problem is that speech interfaces are mostly designed for certain simple commands. However, talking about several things in one utterance can make a dialogue more efficient. To find the appropriate reaction to such utterances, we propose prioritising one task according to certain criteria. Our sequential prioritisation model defines a six-step approach to address this problem.

1 Introduction

Utterances in dialogue serve often more than one communicative function. Like giving feedback about the understanding of a question and answering the question in a single utterance. The ability of humans to easily process such multiple communicative functions and to react accordingly, allows for a swift and effective communication (Lemon et al., 2002). This multifunctionality comes in a variety of forms. According to Allwood (1992), multifunctionality can be sequential or simultaneous. He gives an example where A's utterance contains the functions *feedback giving*, *request*, *request*, *request*, *statement*, and *response elicitation* in a sequential way.

*A: Yes! Come tomorrow. Go to the church!
Bill will be there, OK?* (Allwood 1992)

Bunt and Romary call these functional features such as request, statement, or promise dialogue acts and propose a formally definition:

A dialogue act is a unit in the semantic description of communicative behaviour produced by a sender and directed at an addressee, specifying how the behaviour is intended to influence the context through understanding of the behaviour. (Bunt 2005)

Following the idea of multifunctionality, Bunt (1989, 2009) proposes the dynamic interpretation theory (DIT) which distinguishes dialogue acts in 10 dimensions where participation in a dialogue is viewed as performing several activities sequential and parallel. The First dimension is called Task/Activity. A dialogue act is labelled as Task/Activity if its performance contributes to performing the task or activity underlying the dialogue. Other dimensions cover dialogue acts like discourse structuring, turn management, or management of social obligations.

Utterances containing at least two sequential dialogue acts labelled as Task/Activity, which contributes to two different tasks or activities, are often called multi-intents (MI). Several Researchers used this expression in a human-machine interaction context. Kim et al. (2017) and Shet et al. (2019) propose algorithms to distinguish and segment such MIs like the utterances from speaker B and C.

B: Find the Big Bang Theory tv show and play it.

C: What is the genre of big bang theory? Tell me the story about it.

Such MIs are a useful mechanism to make a dialogue more efficient. Especially during demanding tasks like driving a car, it can be useful to talk about several things at once, to get back to the main task as fast as possible.

If both Tasks require further clarification, it can be difficult to define a proper reaction for a spoken dialogue system. Answering with a MI, too, can

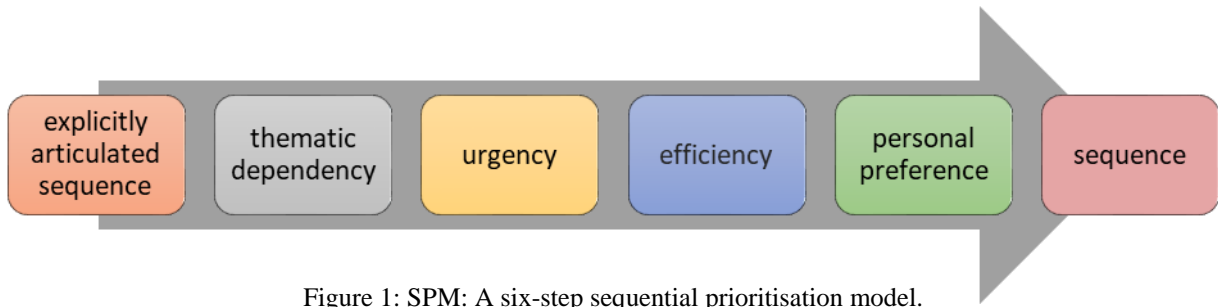


Figure 1: SPM: A six-step sequential prioritisation model.

produce long utterances, which can be cognitively very demanding. Prioritising a certain Task has to be logical and comprehensible. While a human conversation partner can easily decide if answering with a MI is appropriate and if not, identify the most important task, prioritise it and postpone the less important task, computers lack those skills. Therefore, we propose a six-step sequential prioritisation model (SPM) (see Figure 1).

2 Sequential Prioritisation Model:

Explicitly articulated sequence:

The first step is checking if the user mentions an explicit order of tasks. Speaker D explicitly structures the conversation by saying which part of the utterance he wants to take on first.

D: Call my mom and first tell me what the weather's gonna be like.

Thematic dependency:

If no discourse structuring hints are given in the speaker's utterance, there can be dependencies, which predefine the order of sequence.

E: I want to take a break. I am hungry. Isn't Berlin coming soon?

Speaker E wants to take a break and if Berlin is near, it seems like a good opportunity to stop there. Nevertheless, if Berlin is too far away E may need something to eat earlier. Therefore, before looking for restaurants the first task to approach is the last part of the utterance.

Urgency:

If none of the above mentioned criteria is met, there is a chance that one task is more urgent than the other is.

F: What do I do with the zucchini? Oh, the pan is hot. What is coming in now?

A task is urgent if the task has to be completed in a short amount of time, because if not, it loses

relevance or other negative consequences occur. Certainly, speaker F would be frustrated if the first mentioned task is considered before the second one. Urgency seems to play an especially important role in an environment with rapidly changing situations.

Efficiency:

If the tasks are both not urgent and equally important, maybe one of the tasks can be done faster e.g. because it needs less turns to complete. If the first part of G's utterance is done first, the speech channel is blocked by the call. Therefore, the second task cannot be completed until the call has ended.

G: Call my mom and tell me what the weather's gonna be like.

Personal Preference:

If a user-model is present which represents the likings and preferences of the user, the task that is preferred by the user can be prioritised.

Sequence:

The last strategy is the fall back solution, where the system talks about the tasks in the sequence they were mentioned.

3 Conclusion

To overcome the Problem of creating cognitively too demanding dialogues, while reacting to MIs, we present a six-step sequential prioritisation model. Each step defines criteria for the prioritisation of one task and has to be considered before going on to the next one.

Our future research will deal with testing and evaluating the model in real world scenarios with a special focus on the role of urgency. Additionally, we will research the role of explicit discourse structuring in the system's response to clarify the decision in a logical and comprehensible way.

References

- Allwood, J. (1992): On dialogue cohesion. Gothenburg Papers in Theoretical Linguistics 65. Gothenburg University, Department of Linguistics.
- Bunt, H. (1989): Towards a dynamic interpretation theory of utterances in dialogue. In H. Bouma and B. Elsendoorn (eds) Working models of human perception. New York: Academic Press, pp. 419 – 456.
- Bunt, H. (1994): Context and Dialogue Control. *Think Quarterly* 3 (1), 19–31.
- Bunt, H. (2005): A framework for dialogue act specification. Proceedings of SIGSEM WG on Representation of Multimodal Semantic Information.
- Bunt, H. (2009): The DIT++ taxonomy for functional dialogue markup. D. Heylen, C. Pelachaud, R. Catizone and D. Traum (eds.) Proc. AMAAS 2009 Workshop “Towards a Standard Markup Language for Embodied Dialogue Acts”, Budapest, May 2009.
- Bunt, H. (2011). Multifunctionality in dialogue. *Computer Speech & Language*, 25(2), pp. 222-245.
- Kim, B., Ryu, S., & Lee, G. G. (2017): Two-stage multi-intent detection for spoken language understanding. *Multimedia Tools and Applications*, 76(9), pp. 11377-11390.
- Lemon, O. et al. (2002): Multi-tasking and collaborative activities in dialogue systems. In Proceedings of the Third SIGdial Workshop on Discourse and Dialogue.
- Shet, R., Davcheva, E., & Uhle, C. (2019): Segmenting multi-intent queries for spoken language understanding. *Elektronische Sprachsignalverarbeitung 2019*, 141-147.

Tense use in dialogue

Jos Tellings **Martijn van der Klis** **Bert Le Bruyn** **Henriëtte de Swart**
Utrecht University Utrecht University Utrecht University Utrecht University
j.l.tellings@uu.nl m.h.vanderklis@uu.nl b.s.w.lebruyn@uu.nl h.deswart@uu.nl

Abstract

This paper reports on parallel corpus research that shows that there are differences in tense use in written texts between parts that represent dialogue, and parts that narrate the story. This calls for further study of tense use in dialogue, both in written representations, and in spoken dialogue. Yet, in the dialogue semantics literature almost no prior work exists that is devoted to tense use, neither from a formal, nor from a computational angle. We argue that this gap in dialogue research should be filled by further empirical investigation, as well as the development of computational tools for automated annotation of tense and temporal structure in dialogue. This will not only help understand how speakers track the temporal structure of dialogue, but also give theoretical linguistic literature on tense a wider empirical and computational dimension.

1 Differences in tense use between dialogue and narrative

We (Le Bruyn et al., 2019) investigated cross-linguistic variation of tense use by looking at a parallel corpus based on the novel *Harry Potter and the Philosopher's Stone* (HP) and its translations in other languages. It will suffice here to consider only English, and select two chapters from the novel: a more narrative-oriented chapter (chapter 1), and a more dialogue-oriented one (chapter 17). After separating dialogue from narrative in the text, occurrences of the present perfect and the simple past were counted:

	PERFECT	PAST
Narrative	0	600
Dialogue	41	163

Table 1: Tense uses in chapters 1 and 17 of HP

The results show that the PERFECT is not used in

the narrative part, but only in the dialogue parts. One hypothesis for this striking contrast between dialogue and narrative is that it has to do with temporal orientation. The dialogues are more likely to contain utterances of what is currently going on (relative to the story time), whereas the narrative parts tell a story that happened in the past. The traditional view is that the English PERFECT conveys current relevance; this would explain the occurrence of PERFECTS in here-and-now-oriented dialogue, and no occurrences in past-oriented narrative. This leads to the testable prediction that dialogues with a different temporal orientation have a different tense use.

2 Further investigation using the HP corpus

In order to test this hypothesis and its predictions, further empirical investigation is needed, as well as a way to formalize and quantify the notion of ‘temporal orientation’ that was used informally above. As for the empirical part, we start by looking at more data from the HP corpus than in Le Bruyn et al. (2019). Chapters 16 and 17 both contain dialogues, but chapter 16 is more present-oriented than chapter 17.

	PRESENT	PAST	PERFECT
Ch 16 dialogue	182	53	14
Ch 17 dialogue	126	129	22

Table 2: Raw data for tense use in chs 16 and 17 of HP.

The present orientation of Chapter 16 is confirmed by the higher PRESENT : PAST ratio. However, the number of PERFECTS is lower in Chapter 16 than in 17, whereas the hypothesis predicts a higher number. In order to further investigate these preliminary findings, and the consequences for the ‘current relevance’ view of the PERFECT, we need a

more fine-grained analysis of temporal orientation and tense use in dialogue. We also need to be able to scale up, and consider additional and larger dialogue corpora than just HP. Both goals require appropriate computational tools, as discussed in section 3.

3 Development of computational tools for annotating tense in dialogue

Speakers keep track of temporal orientation by parsing temporal expressions and aspect/tense inflection on verbs. Tools to automatically annotate these two categories already exist, but were not designed for dialogue. Therefore we will first provide an evaluation of how currently available tools perform on dialogue, and what improvements are needed.

Evaluation: The required computational steps can be divided into (i) syntactic parsing of tense categories and temporal expressions; and (ii) recognition of temporal links and event structure.

We include in our evaluation some tools that do the first task only: TMV, an annotator for tense/mood; SitEnt, a classifier for aspectual event type; and PerfectExtractor, software developed by one of the authors in our research project (van der Klis). A tool that is designed to do both (i) and (ii) is TARSQI (Verhagen and Pustejovsky, 2012), a toolkit that annotates texts in the ISO-TimeML format, and automatically recognizes events, times, and temporal links between them.

Since TARSQI looks the most promising, we started with that: we applied it to a set of written representations of dialogues (note that TARSQI was originally designed for the newswire domain). We found two major problems. First, with respect to task (i), it fails to recognize basic facts about English tense constructions, for example *have*+participle combinations are not recognized as a single perfect construction when non-adjacent (e.g. in questions: *Has John gone?*).

Second, in the domain of dialogue, the distinction between assertions and questions is of crucial importance. However, TARSQI does not annotate for speech act type, and therefore the time link it correctly ascribes to (1a) ($\text{time}(e_{\text{book-reading}}) \subseteq \text{yesterday}$) is also assigned to (1b) in which the time link is not asserted but presupposed, and to (1c) in which the establishment of the link depends on the answer. So, TARSQI also has problems with task (ii) in the specific domain of dialogue.

- (1) a. Ed read a book yesterday.
- b. Which book did Ed read yesterday?
- c. Q: Did Ed read a book yesterday?
A: Yes. / No.

Dialogue acts: From the evaluation it follows that we should take the internal structure of dialogue seriously in our analysis. This structure typically comes in the form of annotation for Dialogue Acts (DAs), covering question-answer contrasts, but several other details in addition. In order to illustrate the virtues of a DA-based analysis of tense in dialogue, we ran a pilot study by analyzing the Switchboard corpus, which is manually annotated for DA. Because TARSQI fails here, we ran our PerfectExtractor to extract PERFECTS from the corpus. Results (see poster) show a high occurrence of PERFECTS in questions, which underlines the significance of the above remarks on computational tools having problems with question acts.

The pilot study also indicates the limitations of the Switchboard corpus. Several taxonomies of DAs contain tags relating to Topic Management, but the one used in Switchboard, DAMSL, does not (see Petukhova and Bunt, 2009). Topic Management annotation is relevant because of linguistic work that claims that the PERFECT is used in cases of “topic negotiation” (Nishiyama and Koenig, 2010) and “topic closure” (Goutsos, 1997). With dialogue data annotated for Topic Management, we are able to assess those claims in a larger empirical and computational setting.

One way to go is to use systems for automatic DA recognition (which have received a lot of attention recently, e.g. Chen et al., 2018; Kumar et al., 2018; Cerisara et al., 2018) with a taxonomy including Topic Management tags. This allows us to scale up by looking at other datasets of spoken and written dialogue that are not annotated for DA yet. The time and event annotation capacities of systems such as TARSQI are useful, but need to be improved on dialogue-specific acts. This will bring temporal annotation to dialogue, an important step toward formalizing and quantifying the notion of temporal orientation as used above in section 1.

Finally, the development of such a system will benefit a range of other applications that require access to the temporal structure of dialogue, for example in human-machine interaction settings.

References

- Christophe Cerisara, Pavel Král, and Ladislav Lenc. 2018. On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech & Language*, 47:175–193.
- Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via CRF-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234. ACM.
- Dionysis Goutsos. 1997. *Modeling Discourse Topic: sequential relations and strategies in expository text*. Ablex Publishing Corporation, Norwood, NJ.
- Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Thirty-Second AAAI Conference on Artificial Intelligence*. arXiv:1709.04250 [cs.CL].
- Bert Le Bruyn, Martijn van der Klis, and Henriëtte de Swart. 2019. The Perfect in dialogue: evidence from Dutch. In press, *Linguistics in the Netherlands*.
- Atsuko Nishiyama and Jean-Pierre Koenig. 2010. What is a perfect state? *Language*, 86(3):611–646.
- Volha Petukhova and Harry Bunt. 2009. Dimensions in communication. TR, Tilburg University.
- Marc Verhagen and James Pustejovsky. 2012. The TARSQI Toolkit. In *LREC*, pages 2043–2048.

Software

- TMV: <https://clarin09.ims.uni-stuttgart.de/tmv/>
- SitEnt: <http://www.coli.uni-saarland.de/projects/sitent/page.php>
- PerfectExtractor: <https://github.com/UUDigitalHumanitieslab/perfectextractor>

Shared Gaze toward the Speaker and Grounding Acts in Native and Second Language Conversation

Ichiro Umata

Interaction Design Group
KDDI Research, Inc., Tokyo, Japan
ic-umata@kddi-research.jp

Koki Ijuin

The National Institute of Advanced
Industrial Science and Technology
koki-ijuin@aist.go.jp

Tsuneo Kato and Seiichi Yamamoto

Department of Information Systems Design
Doshisha University, Kyoto, Japan
{tsukato, seyamamo}@mail.doshisha.ac.jp

Abstract

The relation between shared gazes toward the current speaker and grounding acts were analyzed from the viewpoint of floor apportionment in Native (L1) and Second language (L2) conversation. Although the shared gaze phenomenon showed common characteristics between L1 and L2 conversations, there are one notable difference: in floor hold utterances, Continue (cf. (Traum, 1994)) utterances were drawing the listener’s visual attention in L1, whereas Initiate (cf. (Traum, 1994)) utterances were in L2.

1 Introduction

In multimodal interactions, the non-verbal cues have been considered particularly important in grounding, i.e. establishing a given piece of information as part of common ground (Clark, 1996). Among nonverbal cues, gaze has been observed to play an important role in communication, such as by expressing emotional states, exercising social control, highlighting the informational structure of speech, and speech floor apportionment (Argyle et al., 1968) (Duncan Jr., 1972) (Holler and Kendrick, 2015) (Kendon, 1967) (Umata et al., 2018) (Ijuin et al., 2018). In this study, we examine shared gaze toward the current speaker from the next speaker and the silent third participant from the viewpoints of floor apportionment and grounding acts defined by (Traum, 1994) in L1 and L2 conversations. The results of correlation analysis of gazes showed both common and different features between the two language conditions. As a common feature, there were shared gaze in floor switch utterances other than acknowledge utterances. As a different feature, there were shared gazes only in continue utterances in L1, whereas only in initiate utterances in L2.

2 Data

We analyzed data from conversations in a mother tongue and those in a second language made by the same interlocutors (for details, refer to (Yamamoto et al., 2015)). The data contains face-to-face three-party conversation in L1 (Japanese) and in L2 (English). We analyzed data from the goal-oriented task in L1 and L2 (20 conversations for each) in this study. Three sets of NAC EMR-9 head-mounted eye trackers and headsets with microphones recorded their eye gazes and voices. A trained annotator annotated the utterances with Grounding Act tags established by (Traum, 1994) for 20 groups of goal-oriented conversations (Umata et al., 2016).

3 Analyses and Results

We conducted correlation analysis of the gazes toward the current speaker (CS) from the next speaker (NS) and the silent third participant (SP) for major 4 grounding acts (*Initiate (init)*, *Continue (cont)*, *Acknowledge (ack)*, and *Acknowledge and Initiate (ack init)*). We used the average of gazing ratios based on Ijuin et al. as indices for the following analyses of gaze (Ijuin et al., 2018). The participant roles were classified into three types: CS as the speaker of the utterance, NS as the participant who takes the floor after the current speaker releases the floor, and SP who is not involved in speaking at that time. The average of role-based gazing ratios is defined as:

Average role-based gazing ratio (gazing ratio):

$$= \frac{1}{n} \sum_{i=1}^n \frac{DG_{jk(i)}}{DSU_{(i)}} \times 100 \text{ (\%)}$$

where $DSU_{(i)}$ and $DG_{jk(i)}$ represent the duration of the i -th utterance and the duration of participant j gazing at participant k during that utterance,

Lang.	GA	ρ	p
L1	<i>init</i>	.805**	.000
L1	<i>cont</i>	.660**	.002
L1	<i>ack</i>	.409	.073
L1	<i>ack init</i>	.579**	.007
L2	<i>init</i>	.594**	.006
L2	<i>cont</i>	.687**	.001
L2	<i>ack</i>	.152	.523
L2	<i>ack init</i>	.632**	.004

Table 1: Correlation of gazes in floor switch

respectively. A role-based gazing ratio is calculated for each group: i.e. a single gaze ratio is computed for each session, and for each relation.

3.1 Shared Gazes in Floor Switch Utterances

We formulated the following hypotheses for shared gazes toward the current speaker in floor switch utterances:

- H1:** In floor switch utterances, the next speaker and the silent third participants would try to obtain the speaker’s nonverbal cues from the visual channel, resulting in frequent shared gaze.
- H2:** There would be little shared gaze toward the current speaker in *ack* utterances where the speaker only acknowledges the previous speaker’s utterances without adding any new piece of information.

The results of Spearman’s correlation analyses are as in Table 1 (The correlation coefficients with their false discovery rates (FDR) $q < .01$ are marked with “**”).

The result showed there were strong correlations other than *ack* utterances, supporting our hypotheses H1 and H2.

3.2 Shared Gazes in Floor Hold Utterances

We formulated the following hypotheses for shared gazes toward the current speaker in floor hold utterances:

- H3:** In floor hold utterances, the speaker’s nonverbal cues would be not as salient as floor switch utterances, resulting in less shared gaze toward the current speaker.

The results of Spearman’s correlation analyses are as in Table 2.

Lang.	GA	ρ	p
L1	<i>init</i>	.090	.705
L1	<i>cont</i>	.583**	.001
L1	<i>ack</i>	-.272	.246
L1	<i>ack init</i>	.128	.591
L2	<i>init</i>	.705**	.001
L2	<i>cont</i>	.309	.185
L2	<i>ack</i>	.323	.164
L2	<i>ack init</i>	.110	.655

Table 2: Correlation of gazes in floor switch

Our hypothesis H3 was partially supported: the results suggest less shared gaze in floor hold utterances. There were, however, high correlations in *cont* in L1, and in *init* in L2, suggesting that the speaker was drawing the listeners’ shared attention in these utterances, and the attention drawing utterance categories were different in these two language conditions.

4 Discussion and Future Work

The analysis of shared gazes in floor switch utterances supported our hypotheses: the speaker gathered shared attention of the listeners other than *ack* utterances. For floor hold utterances, however, the result showed differences between L1 and L2: in floor hold utterances, *cont* utterances were drawing the listener’s visual attention in L1, whereas *init* utterances were in L2. Although the cause of this difference is not clear, one factor that might have affected the listeners’ shared attention is the difference in linguistic proficiency: the speakers might have tended to start their long-lasting speech turns with cushioning utterances with less information content in L1, whereas they could not use such rhetoric in L2 where their linguistic proficiency was not high. Further analyses of utterance content would be required to elucidate the cause of this phenomenon. The current analyses were conducted for Japanese and English conversations as L1 and L2 for each, and extending the analyses to other languages will also be necessary.

References

- Michael Argyle, Mansur Lallijee, and Mark Cook. 1968. The effects of visibility on interaction in a dyad. *Human relations*, 21:3–17.

- Herbert H. Clark. 1996. *Using language*. Cambridge University Press.
- Starkey Duncan Jr. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.
- Judith Holler and Kobin H. Kendrick. 2015. [Unaddressed participants’ gaze in multi-person interaction](#). *Frontiers in psychology*, 6.
- Koki Ijuin, Ichiro Umata, Tsuneo Kato, and Seiichi Yamamoto. 2018. Difference in eye gaze for floor apportionment in native- and second-language conversations. *Journal of Nonverbal Behavior*, 42:113–128.
- A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.
- David R. Traum. 1994. *A computational theory of grounding in natural language conversation*. Ph.D. thesis, University of Rochester.
- Ichiro Umata, Koki Ijuin, Mitsuru Ishida, and Seiichi Yamamoto. 2016. Quantitative analysis of gazes and grounding acts in l1 and l2 conversations. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pages 4249–4252.
- Ichiro Umata, Koki Ijuin, Tsuneo Kato, and Seiichi Yamamoto. 2018. Floor apportionment and mutual gazes in native and second-language conversation. In *Proceedings of the international conference on multimodal interaction (ICMI2018)*, pages 334–341.
- Seiichi Yamamoto, Keiko Taguchi, Koki Ijuin, Ichiro Umata, and Masafumi Nishida. 2015. Multimodal corpus of multiparty conversations in l1 and l2 languages and findings obtained from it. *Language Resources and Evaluation*, 49:857–882.

A Taxonomy of Real-Life Questions and Answers in Dialogue

Maxime Amblard, Maria Boritchev, Marta Carletti, Lea Dieudonat, Yiting Tsai

LORIA, UMR 7503, Université de Lorraine, CNRS, Inria

Nancy, France

{maxime.amblard, maria.boritchev}@loria.fr

{martacarletti1993, leadieudonat}@gmail.com

yi-ting.tsai5@etu.univ-lorraine.fr

Abstract

We present a taxonomy of questions and answers based on real-life data extracted from spontaneous dialogue corpora. This classification allowed us to build a fine-grained annotation schema, which we applied to several languages: English, French, Italian and Chinese.

1 Introduction

Nowadays, most spoken dialogue systems focus on task-based communication (making reservations, getting information, *etc.*). Annotations are often limited to domain-specific purposes. Many dialogues, especially task-oriented ones, are annotated with speech acts, which are a powerful tool to detect questions' and answers' intentions. A tradition of question and answers modelling inspired by logic approaches has been introduced by (Asher and Lascarides, 2003). From a more linguistic point of view, (Ginzburg and Sag, 2000) presents a detailed study of questions coupled with insights on their answers.

As most annotations are highly specific to a task, they fail to account for the complexity of spontaneous dialogues. Our schema is designed to handle phenomena encountered in real-life conversations. We worked on corpora of transcriptions of spontaneous dialogues, mainly in English (Norrick, 2017). We produced an annotation schema that we tested on French (ATILF, 2018), Italian (Sciubba et al., 2014) and Chinese (University, 2015). In this short paper, we focus on questions and answers classification (sect. 2) and on their combinations (sect. 3).

2 Questions and answers classification

We classify the questions and the answers according to their *form* and their *function*, following (Freed, 1994; Blandón et al., 2019). We do not pretend to be exhaustive here as answers can take

arbitrary forms following the non-verbal context of the dialogue. This taxonomy presents the main types of answers one can encounter in real-life corpora of transcribed oral conversations. The form of an utterance is defined by its syntactic form – such as syntactic inversions – and the lexical items that it contains (*wh*-words, 'yes', 'no', *etc.*). The function of an utterance is close to the concept of Austin's illocutionary force (Austin, 1975): it is defined by the intention of the speaker. Our taxonomy takes root in a previous classification schema where questions and answers were classified according to a mixture of form and function (Blandón et al., 2019). In this annotation schema we want to keep the form and the function of questions and answers separate.

In Table 1, we sum up the possible forms and functions for questions and answers. We assume that the interpretation of answers' forms (upper-right) and questions' functions (lower-left) do not need to be developed here. If we look at question forms, *disjunctive* questions can be *inclusive* or *exclusive* (resp.), depending on the interpretation of 'or': 'Do you want sugar or **milk** in your coffee?' vs 'Do you want sugar or **stevia** in your coffee?'. Here, the interpretation of 'or' depends on its arguments. Questions can be *auxiliary-deontic* ('Can you hand me the salt?') or *auxiliary-epistemic* ('Can you swim?') depending on the auxiliary they contain.

Finally, answers functions can vary a lot. Some are lexical, such as *give feature*, proposed in Boritchev (2017) (adapted from Jurafsky and Martin 2000), which corresponds to an answer to a *wh*-question ('Where do you live?'/ 'In Paris.'). Others correspond to an action, such as *perform* ('Can you hand me the salt?'/ '...'/ 'Thank you.').

3 Combining questions and answers

Questions and answers interact with each other. After an analysis of them in isolation, we consider

	Questions	Answers
Form	Yes/No, Wh, Disjunctive-Inclusive, Disjunctive-Exclusive, Auxiliary-Deontic, Auxiliary-Epistemic	Yes/No, Wh, Uncertain, Unknown
Function	Completion Suggestion, Phatic, Ask_Confirmation, Ask_Feature, Ask_Performance, Reported Speech (RS)	Refuse, Accept, Phatic, Give_Confirmation, Give_Uncertainty, Give_Unknown, Reported Speech (RS), Give_Feature, Perform, NONE

Table 1: Forms and Functions of Questions and Answers

how their association works and how it can result in comprehension. To do so, we introduce the notions of *symmetry* and *mismatch*. An answer is symmetric (see ex. 1) to its question when the semantic or syntactic requirements imposed by the question are fulfilled by the answer. If it is not the case, it is asymmetric (see ex. 2).

Example 1 Symmetry of form and function

A: *Why are you crying?*

B: *Because I hurt myself.*

In this example, the question is of Wh-form and its function is Ask_Feature. As the answer starts by ‘Because’, it is classified as of Wh-form and its function Give_Feature. Therefore, the semantic requirement imposed by the question is fulfilled by the answer.

Example 2 Asymmetry of form and function

A: *so- wh- where can you move to?*

B: *Well...you know...I don’t even know where I’m living next year.*

In ex. 2, the question is of Wh-form and its function is Ask_Feature. Yet, the answer is fuzzy and is classified as of Uncertain form and Give_Uncertainty function. Therefore, the syntactic requirement is not fulfilled.

Next, we define the notions of *mismatch of form (resp. function)*: when there is an asymmetry of form (resp. function) between a question and its answer, a mismatch of form (resp. function) occurs if and only if the form (resp. function) of the given answer doesn’t fall under one of the forms (resp. functions) accepted by the question. The identification of compatible questions and answers goes through tables of compatibility. They map the forms and functions that can combine with each other (in both cases of symmetry and asymmetry). In Table 2, question forms are associated with a set of answer forms that do not trigger a mismatch. Table 3 presents compatibilities of functions.

Q_Forms	Expected answer forms
Yes-no	{ Yes/No, Uncertain, Unknown }
Wh	{ Wh, Uncertain, Unknown }
Disj._Inclusive	{ Yes/No, Uncertain, Unknown }
Disj._Exclusive	{ Wh, Uncertain, Unknown }
Aux._Deontic	{ Yes/No, NONE, Performance }
Aux._Epistemic	{ Yes/No, Uncertain, Unknown }

Table 2: Compatibility form

Q_Function	Expected answer function
Completion Suggestion	{ Refuse, Accept, Phatic, Give_Confirmation }
Phatic	{ Refuse, Phatic, Give_Confirmation, Report, NONE }
Ask_Confirmation	{ Refuse, Accept, Give_Uncertainty, Give_Unknown, Give_Confirmation }
Ask_Feature	{ Give_Feature, Give_Uncertainty, Give_Unknown }
Ask_Performance	{ Perform, NONE, Give_Unknown, Give_Uncertainty, Accept }
RS	{ Phatic, Reported, NONE }

Table 3: Compatibility function

4 Conclusion

This taxonomy of questions and answers allowed us to produce an annotation schema. We tested it on English, French, Italian and Chinese corpora.¹ We were able to tag a wide range of questions and their possible answers. The notion of mismatch allowed us to detect cases of indirect answers and distinguish them from cases where no answers were given. Following this process, we are also able to combine sequences of questions and answers in coherent blocs that constitute negotiation phases (Boritchev and Amblard, 2018).

¹See our poster for results.

References

- Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.
- ATILF. 2018. [Tcof : Traitement de corpus oraux en français](#). ORTOLANG (Open Resources and TOols for LANGuage) –[www.ortolang.fr](#).
- John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.
- María Andrea Cruz Blandón, Gosse Minnema, Aria Nourbakhsh, Maria Bortichev, and Maxime Amblard. 2019. Toward dialogue modeling: A semantic annotation scheme for questions and answers. In *The 13th Linguistic Annotation Workshop (The LAW XIII)*.
- Maria Boritchev. 2017. Approaching dialogue modeling in a dynamic framework. Master’s thesis, Université de Lorraine.
- Maria Boritchev and Maxime Amblard. 2018. [Coffee or tea? Yes](#). SEMDIAL 2018 (AixDial) - The 22nd workshop on the Semantics and Pragmatics of Dialogue. Poster.
- Alice F. Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of Pragmatics*, 21(6):621 – 644.
- Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations: the form, meaning, and use of English interrogatives*. CSLI Publications, Stanford.
- Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, Upper Saddle River, NJ, USA.
- Neal Norrick. 2017. [SCoSE part 1: Complete conversations](#). English Linguistics, Department of English at Saarland University.
- Maria Eleonora Sciubba, Stefania Marzo, and Elwys De Stefani. 2014. Strengthening students interactional competence in Italian L2 by exploiting a corpus of video-recorded spontaneous interactions. ALIAS–Archivio di LInguA Spontanea. In *Euro-Call 2014, Date: 2014/08/20-2014/08/23, Location: Groningen*.
- HongKong Polytechnic University. 2015. PolyU corpus of spoken Chinese. Department of English.

Pattern Recognition is Not Enough: Representing Language, Action and Perception with Modular Neural Networks

Simon Dobnik

Department of Philosophy, Linguistics
and Theory of Science (FLoV)
University of Gothenburg, Sweden
simon.dobnik@gu.se*

John D Kelleher

Information, Communications
and Entertainment Institute (ICE)
Dublin Institute of Technology, Ireland
john.d.kelleher@dit.ie

Abstract

Current deep learning approaches to modelling of spatial language in generating image captions have shortcomings because they are focused on recognition of visual patterns. The multiplicity of factors that influence spatial language which also include aspects of interaction between speakers and between speakers and their environment invites a modular approach where the solution can be built in a piece-wise manner and then integrated. We call this approach where deep learning is assisted with domain knowledge expressed as modules that are trained on data a top-down or mechanistic approach to otherwise a bottom-up phenomenological approach.

In recent years deep learning approaches have made significant breakthroughs. An exciting aspect of deep learning is learning inter/multi-modal representations from data that includes discrete information (e.g. words) and continuous representations (e.g. word embeddings and visual features), such as those used in automatic image captioning systems. A number of shortcomings with current deep learning architectures have been identified with respect to their application to spatial language such as “the chair is to the left and close to the table” or “go down the corridor until the large painting on your right, then turn left”. For example, in (Kelleher and Dobnik, 2017) we argue that contemporary image captioning networks have been configured in a way that they capture visual properties of objects (“what” in terms of (Landau and Jackendoff, 1993)) rather than spatial relations between them (“where”). Consequently, within the captions generated by these systems the

relation between the preposition and the object is not grounded in geometric representation of space but only in the linguistic sequences through the decoder language model where the co-occurrence of particular words in a sequence is estimated.¹ This is because neural networks are typically used as generalised learning mechanisms that learn with as little supervision through architecture design as possible. We call this data-driven approach a *bottom-up* or *phenomenological approach*. The problem is that the chosen architecture may not be optimal for every aspect of the cognitive representations that we want to learn.

We do not argue that language model is not informative for predicting spatial relations. In addition to (i) scene geometry (Logan and Sadler, 1996; Dobnik and Åstbom, 2017) they also rely on (ii) perspective and perceptual context (Kelleher and Kruijff, 2005; Dobnik et al., 2015), (iii) functional world knowledge about dynamic kinematic routines of objects (Coventry et al., 2005), and (iv) interaction between agents through language and dialogue and with the environment through perception (Schutte et al., 2017; Dobnik and de Graaf, 2017). In (Dobnik et al., 2018) we show that a language model is useful in predicting functional relations between objects. The system can learn something about object interaction without visually observing these objects and such knowledge is used as background knowledge when generating and interpreting spatial descriptions. The information expressed in a language model or visual features of the scene is therefore just one of the modalities that must be taken into account. This provides a challenge for computational modelling

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.

¹The over-reliance of deep learning models on the language model has been criticised recently for example, in relation to visual question answering and an attempts have been made to make the systems give a greater weight to images in predicting the caption, for example by balancing different answers in datasets (Agrawal et al., 2017).

of spatial descriptions because (i) it is difficult to provide and integrate that kind of knowledge and (ii) its contextual underspecification. A computational system taking into account these meaning components in the context would be able to understand and generate better, more human-like, spatial descriptions and engage in more efficient communication in the domain of situated agents and humans. Furthermore, it could exploit the synergies between different knowledge sources to compensate missing knowledge in one source from another (Schutte et al., 2017).

In (Dobnik and Kelleher, 2017) we argue that the multiplicity of factors that influence spatial language invites a modular approach where the solution can be built in a piece-wise manner and then integrated (Feldman, 1989; Regier, 1996; Andreas et al., 2016; Johnson et al., 2017). We call this approach where deep learning is assisted with domain knowledge expressed as modules that are trained on data a *top-down* or *mechanistic approach*. One challenge to spatial language is the lack of an overarching theory explaining how these different factors should be integrated but (Herskovits, 1987) and (Coventry et al., 2005) appear to be promising candidates. Early work on neural networks includes some examples of neural models that could provide a basis for the design of specific modules. For example, (Regier, 1996) captures geometric factors and paths of motion. The system in (Coventry et al., 2005) processes dynamic visual scenes containing three objects: a teapot pouring water into a cup and the network learns to optimise, for each temporal snapshot of a scene, the appropriateness score of a spatial description obtained in subject experiments. The idea behind these experiments is that descriptions such as *over* and *above* are sensitive to a different degree of geometric and functional properties of a scene, the latter arising from the functional interactions between objects. The model is split into three modules: (i) a vision processing module that deals with detection of objects from image sequences that show the interaction of objects, (ii) an Elman recurrent network that learns the dynamics of the attended objects in the scene over time, and (iii) a dual feed-forward vision and language network to which representations from the hidden layer of the Elman network are fed and which learns how to predict the appropriateness score of each description for each temporal con-

figuration of objects. Each module of this network is dedicated to a particular task: (i) to recognition of objects, (ii) to follow motion of attended objects in time and (iii) to integration of the attended object locations with language to predict the appropriateness score, factors that have been identified to be relevant for computational modelling of spatial language and cognition in previous experimental work (Coventry et al., 2005). The example shows the effectiveness of representing networks as modules and their possibility of joint training where individual modules constrain each other.

The model could be extended in several ways. For example, contemporary CNNs and RNNs could be used which have become standard in neural modelling of vision and language due to their state-of-the-art performance. Secondly, the approach is trained on a small dataset of artificially generated images of a single interactive configuration of three objects. An open question is how the model scales on a large corpus of image descriptions (Krishna et al., 2017) where considerable noise is added: the appearance and location of objects is distorted by the angle at which the image is taken. Furthermore, there are no complete temporal sequences of objects and the corpora mostly do not contain human judgement scores on how appropriate a description is given an image. Finally, (Coventry et al., 2005)’s model integrates three modalities used in spatial cognition, but as we have seen there are several others. An important aspect is grounded linguistic interaction and adaptation between agents. For example, (Lazari-dou et al., 2016) describe a system where two networks are trained to perform referential games (dialogue games performed over some visual scene) between two agents. In this context, the agents develop their own language interactively. An open research question is whether parameters such as frame of reference intended by the speaker of a description could also be learned this way.

Due to their dependence on several modalities spatial descriptions therefore provide a good test-bed for the requirements of modelling language, action and perception with neural networks. While it is hard to capture these modalities with a general learning framework, using our expert domain knowledge and splitting the networks into modules that can be specialised for a purpose reduces the complexity of the learning task and makes it more tractable.

References

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. [Don't just assume; look and answer: Overcoming priors for visual question answering](#). *arXiv*, arXiv:1712.00377 [cs.CV]:1–15.
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. [Learning to compose neural networks for question answering](#). In *Proceedings of NAACL-HLT 2016*, pages 1545–1554, San Diego, California. Association for Computational Linguistics.
- Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. [Spatial prepositions and vague quantifiers: Implementing the functional geometric framework](#). In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.
- Simon Dobnik and Amelie Åstbom. 2017. [\(Perceptual\) grounding as interaction](#). In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.
- Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. [Exploring the functional and geometric bias of spatial relations using neural language models](#). In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Simon Dobnik and Erik de Graaf. 2017. [KILLE: a framework for situated agents for learning language through interaction](#). In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 162–171, Gothenburg, Sweden. Northern European Association for Language Technology (NEALT), Association for Computational Linguistics.
- 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 and John D. Kelleher. 2017. [Modular mechanistic networks: On bridging mechanistic and phenomenological models with deep neural networks in natural language processing](#). In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017)*, Gothenburg, 12–13 June 2017, volume 1, pages 1–11, Gothenburg, Sweden.
- Jerome A. Feldman. 1989. Structured neural networks in nature and in computer science. In Rolf Eckmiller and Christoph v.d. Malsburg, editors, *Neural Computers*, pages 17–21. Springer, Berlin, Heidelberg.
- Annette Herskovits. 1987. *Language and Spatial Cognition*. Cambridge University Press, New York, NY, USA.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, C. Lawrence Zitnick, and Ross B. Girshick. 2017. [Inferring and executing programs for visual reasoning](#). *arXiv*, arXiv:1705.03633v1 [cs.CV]:1–13.
- John D. Kelleher and Simon Dobnik. 2017. [What is not where: the challenge of integrating spatial representations into deep learning architectures](#). In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017)*, volume 1, pages 41–52, Gothenburg, Sweden.
- John D. Kelleher and Geert-Jan M. Kruijff. 2005. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, pages 1–7, Aberdeen, Scotland. Association for Computational Linguistics.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *International Journal of Computer Vision*, 123(1):32–73.
- Barbara Landau and Ray Jackendoff. 1993. “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. [Multi-agent cooperation and the emergence of \(natural\) language](#). *arXiv*, arXiv:1612.07182v2 [cs.CL]:1–11.
- Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.
- Terry Regier. 1996. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, Massachusetts, London, England.
- Niels Schutte, Brian Mac Namee, and John D. Kelleher. 2017. Robot perception errors and human resolution strategies in situated human–robot dialogue. *Advanced Robotics*, 31(5):243–257.

Investigating Variable Dependencies in Dialogue States

Anh Duong Trinh[†], Robert J. Ross[†], John D. Kelleher[‡]

[†] School of Computer Science

[‡] Information, Communications & Entertainment Institute

Technological University Dublin

ADAPT Centre, Ireland

anhduong.trinh@mydit.ie, {robert.ross, john.d.kelleher}@dit.ie

Abstract

Dialogue State Tracking is arguably one of the most challenging tasks among dialogue processing problems due to the uncertainties of language and complexity of dialogue contexts. We argue that this problem is made more challenging by variable dependencies in the dialogue states that must be accounted for in processing. In this paper we give details on our motivation for this argument through statistical tests on a number of dialogue datasets. We also propose a machine learning-based approach called energy-based learning that tackles variable dependencies while performing prediction on the dialogue state tracking tasks.

1 Introduction

Dialogue Systems have a wide application in the modern world to assist users with conversational activities. Among dialogue processing tasks dialogue state tracking is the process of identifying user intents within the dialogue contexts. Generally task-oriented dialogue systems define dialogue states as a combination of slot-value pairs. We argue that there exist relationships among the slots, that must be taken into account in the dialogue state tracking process to reflect the natural human way of processing information.

The idea of leveraging variable dependencies in the dialogue state tracking process is not new to the research community. There have been several published works around this phenomenon such as in the multi-task learning model (Trinh et al., 2018), the language modelling tracker (Platek et al., 2016), Conditional Random Fields (Kim and Banchs, 2014), Attention-based Sequence-to-Sequence model (Hori et al., 2016), and the work by Williams (2010). We find that these approaches are good at leveraging variable dependencies at different stages of the architecture.

In this paper we perform statistical tests on spoken dialogue data of Dialogue State Tracking Challenge (DSTC) series including the second challenge (Henderson et al., 2014a) and the third challenge (Henderson et al., 2014b). We demonstrate that there exist strong dependencies between dialogue slots that validate the motivation for our research direction. Moreover, we present the energy-based learning approach to predicting dialogue states while accounting for the variable relationships.

2 Categorical Data Analysis

To investigate the presence or not of variable dependencies, we perform statistical tests pairwise on labels for bivariate statistics. The chosen method is Pearson’s chi-square test, which is effective for categorical data. There exist several measurements of association strength between variables directly related to the chi-square test statistics such as ϕ coefficient, contingency coefficient C , and Cramer’s V . These measures are scaled between 0 and 1 indicating that 1 is the perfect relationship and 0 is no relationship between variables.

We report the statistics of DSTC2 and 3 data in table 1. The variable dependencies are reported with the chi-square test-based Cramer’s V coefficient.

In the result we observe that all statistical significance values $p < 0.05$, that confirms the existence of variable dependencies within dialogue data. We also find that these dependencies are stable strong ($V \geq 0.15$).

To expand on this, let us consider the case of the DSTC2 data. Here, the analysis shows that slot *food* is strongly dependent on slots *price range* and *area* in the domain. This implication indicates that when processing a restaurant search query, the

DSTC2				DSTC3				
	food	price	area		food	price	area	type
food	-			food	-			
price	0.305	-		price	0.248	-		
area	0.269	0.214	-	area	0.163	0.232	-	
				type	0.300	0.195	0.220	-

Table 1: Statistical tests on DSTC2 & 3 data. The results are reported with the Cramer’s V coefficient.

system should not process *food* without considering *price range* or *area* and vice versa. For example, a query such as “*expensive French food in city centre*” should return more results than “*expensive fast food*”. On the other hand, the relationship between *price range* and *area* is weaker than with slot *food*, but still relatively strong.

Overall, the data analysis results validate our motivation of accounting variable dependencies in dialogue state predictions. By adding these dependencies as extra information in the interpretation process, we argue that we can enhance the dialogue state tracker on tracking more challenging situations.

3 Energy-based Dialogue State Tracker

Given the strong dependencies seen between these dialogue state tracking variables, we argue that it is important that any tracking of dialogue state variables should take such dependencies into account as to ignore these dependencies is to assume an Independence that is not valid. To that end we propose a machine learning-based approach, that are notable for tackling the associations between variables, to the dialogue state tracking task. Currently we are investigating the appropriateness of this approach to the dialogue state tracking challenge series.

The core of our on-going work is based on a structured prediction methodology based on Energy-Based Learning. Energy-based models are notably good at structured prediction tasks such as in our case where there are dependencies between a set of predicted variables (LeCun et al., 2006).

The main concept of this method is to associate a structured output Y and the input vector X with a scalar energy value $E = E(X, Y)$ and to measure their goodness of fit using an appropriate loss function $L(E, E^*)$ on those energy values (figure 1).

Currently we are developing energy-based dialogue state tracking models based on a number of

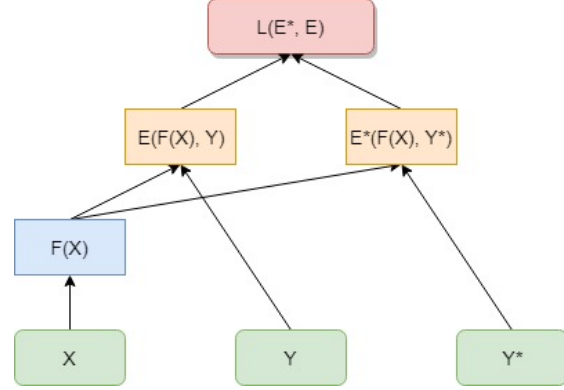


Figure 1: An example of Energy-Based Model, that consists of a feature network $F(X)$, an energy function $E(F(X), Y)$, and an objective function $L(E, E^*)$, where X is input variable, $F(X)$ is a feature representation generated by a feature network, Y is predicted output variable, and Y^* is a gold standard label output variable.

energy-based architectures such as Structured Prediction Energy Networks (SPEN) (Belanger and McCallum, 2016; Belanger et al., 2017) and Deep Value Networks (DVN) (Gygli et al., 2017). Following these approaches we build our energy networks on top of a LSTM-based (Hochreiter and Schmidhuber, 1997) analyser that builds a feature representation for individual dialogue turns.

4 Conclusion

To date our approach has shown a lot of promise in improving on models where variable dependencies are otherwise ignored. In details our energy-based tracker outperforms a LSTM-based multi-task model (Trinh et al., 2018) on both DSTC2 & 3 main tasks. The SPEN methodology helps to improve DSTC2 performance measured with accuracy metric by 3%, while the DVN algorithm increases DSTC2 result by 5% and DSTC3 by 9%.

The observed improvement is achieved mainly due to the energy function and inference process of the energy-based learning approach that takes advantage of target variable dependencies.

Acknowledgments

This research was supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

- David Belanger and Andrew McCallum. 2016. [Structured Prediction Energy Networks](#). In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48.
- David Belanger, Bishan Yang, and Andrew McCallum. 2017. [End-to-End Learning for Structured Prediction Energy Networks](#). In *Proceedings of the 34th International Conference on Machine Learning*.
- Michael Gygli, Mohammad Norouzi, and Anelia Angelova. 2017. [Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs](#). In *Proceedings of the 34th International Conference on Machine Learning*.
- Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The Second Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2014 Conference*, pages 263–272.
- Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014b. The Third Dialog State Tracking Challenge. In *Proceedings of 2014 IEEE Workshop on Spoken Language Technology*, pages 324–329.
- Sepp Hochreiter and Jurgen Schmidhuber. 1997. [Long Short-Term Memory](#). *Neural Computation*, 9(8):1735–1780.
- Takaaki Hori, Hai Wang, Chiori Hori, Shinji Watanabe, Bret Harsham, Jonathan Le Roux, John R. Hershey, Yusuke Koji, Yi Jing, Zhaocheng Zhu, and Takeyuki Aikawa. 2016. Dialog State Tracking With Attention-Based Sequence-To-Sequence Learning. In *Proceedings of 2016 IEEE Workshop on Spoken Language Technology*, pages 552–558.
- Seokhwan Kim and Rafael E. Banchs. 2014. Sequential Labeling for Tracking Dynamic Dialog States. In *Proceedings of the SIGDIAL 2014 Conference*, pages 332–336.
- Yann LeCun, Sumit Chopra, Raia Hadsell, Marc’Aurelio Ranzato, and Fu Jie Huang. 2006. [A Tutorial on Energy-Based Learning](#). *Predicting Structured Data*.
- Ondrej Platek, Petr Belohlavek, Vojtech Hudecek, and Filip Jurcicek. 2016. [Recurrent Neural Networks for Dialogue State Tracking](#). In *Proceedings of CEUR Workshop, ITAT 2016 Conference*, volume 1649, pages 63–67.
- 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, SEMDIAL*, pages 132–145.
- Jason D. Williams. 2010. [Incremental Partition Recombination For Efficient Tracking Of Multiple Dialog States](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5382–5385.

“What are you laughing at?”

Incremental processing of laughter in interaction

Arash Eshghi
Heriot-Watt University
a.eshghi@hw.ac.uk

Vladislav Maraev
University of Gothenburg
vladislav.maraev@gu.se

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

Julian Hough
Queen Mary University of London
j.hough@qmul.ac.uk

Chiara Mazzocchi
Université de Paris
chiara.mazzocchi@live.it

Abstract

In dialogue, laughter is frequent and can precede, follow or overlap what is laughed at. In this paper, we provide a preliminary, but unitary formal account of how forward- & backward-looking laughter are processed and integrated, using Dynamic Syntax which already has well-motivated accounts of repair, split utterances and feedback.

1 Introduction

In dialogue, laughter is very frequent and can constitute up to 17% of the total duration of conversation (in French part of the DUEL corpus, [Tian et al., 2016](#)). Following the terminology from conversational analysis ([Glenn, 2003](#)), we employ the term *laughable* to refer to what the laughter is pointing at, without making any claims about its possible humorous content.

According to preliminary work on the sequential distribution of laughter ([Tian et al., 2016](#)), 90% of laughables are present in the conversation in which they occur and can be ‘laughed about’ more than once. Laughter can precede, follow or overlap the laughable, with the time alignment between the laughter and laughable dependent on who produces the laughable and the form of the laughter. Laughter can interrupt either one’s own or one’s conversational partners’ utterances and this interruption does not necessarily occur at phrase boundaries (contra [Provine \(1993\)](#), e.g. ‘She is a his long-term heh friend’).

In this paper, we present a *unitary* (if preliminary) account of how laughter can be processed and integrated, following Dynamic Syntax ([Kempson et al., 2001, 2016](#), henceforth DS) accounts of repair in dialogue ([Hough, 2014; Eshghi et al., 2015](#)) and *Feedback Relevance Spaces* ([Howes and Eshghi, 2017a](#)). This account focuses on what laughter is doing as opposed to trying to

determine its meaning (c.f. [Ginzburg et al. \(2015\); Mazzocchi et al. \(2018\)](#)). Much like repair and feedback, laughter can occur sub-sententially and can be categorised as forward-looking or backward-looking. We model it analogously to pronouns, which can also be backward-looking (anaphoric) or forward-looking (cataphoric). Just as with pronouns, the laughable can come from linguistic material, or something non-linguistic in the context (as e.g. when we laugh at someone slipping on a banana peel).

2 Laughter in Dynamic Syntax (DS)

We are here using DS-TTR, and the formula decorations are record types ([Cooper and Ginzburg, 2015; Purver et al., 2011](#)). Space constraints do not allow us to introduce the DS machinery (see [Kempson et al., 2016; Cann et al., 2005a; Eshghi et al., 2012](#)); so we proceed directly to the analysis. We treat different types of laughter including forward-looking & backward-looking laughter *uniformly* as anaphoric. Akin to pronouns, this is done by projecting on the node under development, a formula meta-variable, together with a requirement for a fixed formula to be found ($? \exists x. Fo(x)$). The difference with pronouns is that laughter provides the additional semantic information that the laughable – the ‘referent’ to be identified – is laughed at. This extra semantic information is provided on a DS LINKED tree, linked to the node under development, with its root content later *conjoined* with that of the laughable at a later point when LINK-EVALUATION applies (see Fig. 2). Fig. 1 thus specifies a single lexical entry for laughter.

Paired with the LATE-***-ADJUNCTION mechanism in DS – used to model right-periphery phenomena, such as short answers to WH-questions (see [Gargett et al. \(2009\) & Cann et al. \(2005b\)](#), chapter 5) – this provides all that is needed for the incremen-

<i>laughter</i>	IF	$?Ty(X)$ $\neg(\downarrow_L)\exists x.Tn(x)$
	THEN	$make(\downarrow_L)$ $go(\downarrow_L)$ $put(Ty(X))$ $put(Fo(\left[\begin{array}{ll} head & : X \\ p=laughable(head) & : t \end{array} \right]))$ $go(\uparrow_L)$ $put(? \exists x.Fo(x))$ $put(Fo(U))$
	ELSE	ABORT

Figure 1: Lexical Entry for $\langle laughter \rangle$

tal interpretation of forward- and backward- looking laughter, whether the laughter occurs *locally* or is more distant from it, much like how anaphora and cataphra are modelled in DS.

Fig. 2 illustrates the process of parsing a forward-looking laughter, where the laughter is immediately followed by the laughable, “a telescope” — here we only illustrate the $Ty(e)$ subtree under development, which is attached to a larger tree with root node $Ty(t)$. Initially, the laughter token annotates the pointed node of $?Ty(e)$ with a metavariable ($Fo(U)$), and the attendant formula requirement, then LINKING off of that node to project the laughter’s semantic information on the LINKED tree. This leads to a type-complete node, but one which still requires a fixed formula value. Without the process of LATE- $*$ -ADJUNCTION, the parsing of the follow-up NP would be precluded. However, LATE- $*$ -ADJUNCTION allows an *unfixed node* to be introduced immediately below the $Ty(e)$ node, with the pointer moving onto to this unfixed node (connected with the dashed line). This then allows the follow-up NP, “a telescope” to be parsed as normal, leading to the bottom tree in Fig. 2. This is followed by steps of MERGE and LINK-EVALUATION, integrating the content of the laughter with the laughable NP, and allowing the parse to continue as normal.

Discussion Our model is couched purely in processing terms: it remains agnostic about the meaning of laughter, which can be determined by other factors such as intonation, social context and common ground. A reasonable approach to tackle this issue is to extend the account of integrating laughter into dialogue grammar (Ginzburg et al., 2015).

If no appropriate laughable is found, there is the possibility of clarification interaction (e.g.

“What’s funny?”). However, clarification requests of laughter are rare (Mazzocchi et al., 2018), suggesting that what counts as a laughable is a very widely applicable notion such that the laughter can almost always be resolved to some laughable.

Laughter by another may also serve as positive signal of understanding, i.e. have a *grounding effect* (Clark, 1996). Within the DS-TTR model, this grounding effect is also captured *for free* following the DS model of feedback in conversation such as backchannels & clarification requests (Eshghi et al., 2015; Howes and Eshghi, 2017b); this is because backward-looking laughter is treated as a *continuation* or *completion* (Howes, 2012). See Eshghi et al. (2015) for details.

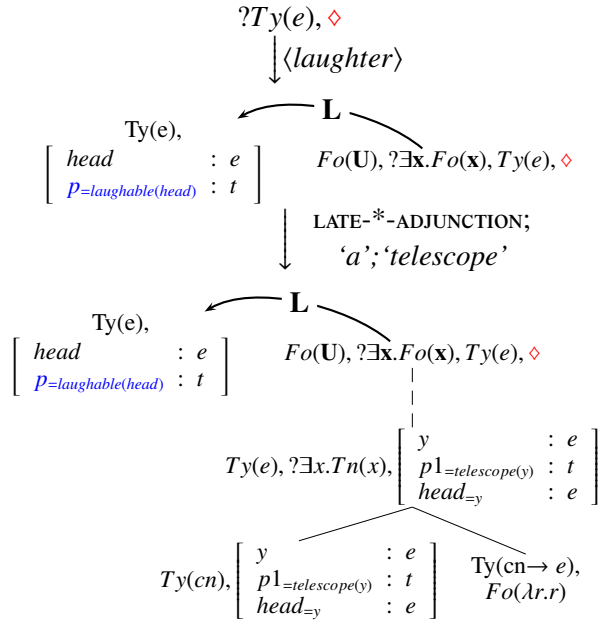


Figure 2: Processing “... $\langle laughter \rangle$ a telescope”

We have also not provided an account of how laughter is distributed syntactically in conversation. We plan to conduct further research investigating how the grammar of a languages provides opportunities for laughter using data with precise laughter annotation collected in the DUEL (French, Chinese and German, Hough et al., 2016) and NOMCO (Nordic languages, Navarretta et al., 2012) projects. We hypothesise that just as with patterns of repair, which vary across languages (Rieger, 2003) because of the specific features of the language (e.g. English allows self-repairs which repeat the determiner before a noun, but this strategy is not available for languages without determiners as separate words, such as Persian) there will be different patterns of laughter placement in different languages, constrained by the unfolding structure of the linguistic input.

References

- Ronnie Cann, Tami Kaplan, and Ruth Kempson. 2005a. Data at the grammar-pragmatics interface: the case of resumptive pronouns in English. *Lingua*, 115(11):1475–1665. Special Issue: On the Nature of Linguistic Data.
- Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005b. *The Dynamics of Language*. Elsevier, Oxford.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.
- Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. *The Handbook of Contemporary Semantic Theory*, pages 375–407.
- Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. [Conversational interactions: Capturing dialogue dynamics](#). In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.
- Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matt Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguistics.
- Andrew Gargett, Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, and Yo Sato. 2009. Grammar resources for modelling dialogue dynamically. *Cognitive Neurodynamics*, 3(4):347–363.
- Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*.
- Phillip Glenn. 2003. *Laughter in interaction*, volume 18. Cambridge University Press.
- Julian Hough. 2014. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.
- Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.
- Christine Howes. 2012. *Coordination in dialogue: Using compound contributions to join a party*. Ph.D. thesis, Queen Mary University of London.
- Christine Howes and Arash Eshghi. 2017a. Feedback relevance spaces: The organisation of increments in conversation. In *IWCS 2017 12th International Conference on Computational Semantics Short papers*.
- Christine Howes and Arash Eshghi. 2017b. [Feedback relevance spaces: the organisation of increments in conversation](#). In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.
- Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikiriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–275.
- Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.
- Chiara Mazzocconi, Vladislav Maraev, and Jonathan Ginzburg. 2018. Laughter repair. In *Proceedings of SemDial 2018 (AixDial)*, pages 16–25.
- Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, and Patrizia Paggio. 2012. Feedback in nordic first-encounters: a comparative study. In *LREC*, pages 2494–2499.
- Robert R Provine. 1993. Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology*, 95(4):291–298.
- Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK.
- Caroline L Rieger. 2003. Repetitions as self-repair strategies in english and german conversations. *Journal of Pragmatics*, 35(1):47–69.
- 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.

Exploring Lattice-based Models of Relevance in Dialogue for Questions and Implicatures

Julian Hough¹ and Andrew Lewis-Smith²

¹Cognitive Science Group, ²Theory Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
{j.hough, a.lewis-smith}@qmul.ac.uk

Abstract

We present work in-progress on modelling relevance in dialogue for questions and implicatures, setting out a formal programme of work on reducing the redundancy which classical logic introduces into proofs. To do this we firstly propose the use of relevance logics, then set out a lattice-theoretic generalisation of Knuth’s and Hough and Purver’s models of questions and answers to achieve Belnap’s First-degree Entailment.

1 Introduction

Formalizing what a relevant contribution consists of in a dialogue and particularly what constitutes a relevant answer to a question is now a classical problem for formal dialogue modelling. It has enjoyed a range of existing major treatments, clearly defined as a formal challenge from Grice et al. (1975) onwards and made into a sub-discipline of pragmatics with Relevance Theory (Sperber and Wilson, 1986).

Relevance was born out of Grice’s original theory of *implicature*, where speakers implicate hidden meaning which hearers can make sense of as in (1) from (Davis, 2014).

Alan: Are you going to Paul’s party?
Barb: I have to work. (1)

While a literal interpretation of Barb’s contribution would not permit it to be judged a relevant answer to Alan’s question, the unspoken meaning that she cannot attend is recoverable. Deriving from Grice’s account, as Davis (2014) notes, “Neo-Gricean theories have modified Grice’s principles to some extent, and Relevance theories replace them with a principle of communicative efficiency. The problems for such principle-based theories include overgeneration, lack of determinacy, clashes, and the fact that speakers often have other goals.” We add to this criticism

the failure to give *real-valued* relevance measures to contributions, especially for answers to questions, though see (Hough and Purver, 2017) for one such approach in progress. In the current models the short polar answers ‘yes’ and ‘no’ would have the same degree of relevance as Barb’s actual answer above, which is unintuitive.

2 Implicature with relevance logic

Here we explore some formal models of relevance agnostic to a theory of intention recognition, but which maintain the principle of least effort and maximising relevance in communication. To do this we look beyond classical logical approaches and move to a *relevance logic* approach. We furthermore explore how real-valued relevance measures of answers to questions could be incorporated into such a framework through lattice theory. We are aiming for a model which would put a real value on the degree of relevance of the contribution if certain reasoning is applied to yield the unsaid meaning and implicature.

Relevance in relevance logics is understood as ensuring every premise in a derivation is used in a proof. This has a connection in theoretical computer science to relevant type systems (Walker, 2005) and numerous engineering applications, e.g. (Cheng, 2004) or (Bruns and Huth, 2011).

In our example (1) we assume Alan and Barb would both have access to a general reasoning rule that may be available as a resource or reasoning pattern like (2).

$$\begin{aligned} &X \text{ is working at time } T \rightarrow \\ &\neg X \text{ can go to a party at time } T \\ &(\text{work-party exclusion rule}) \end{aligned} \quad (2)$$

This rule tells us when someone is working they cannot attend a party (fairly reasonable consideration for most, unless one works in e.g cater-

1. Barb can go to a party at time $T \vee \neg$ Barb can go to a party at time $T_{\{1\}}$ - *question*
2. Barb is working at time $T_{\{2\}}$ - *statement*
3. \neg Barb can go to a party at time $T_{\{2\}}$ - *instantiation of work-party exclusion rule applied to 2*
4. ResolveQuestion(1,3) $\{1,2\}$ - *question resolution of question 1 by statement 3*

Figure 1: Deriving an implicated answer to a question by Relevance Logic proof.

ing, clown acts, etc.). With this rule to hand, in the spirit of (Breitholtz, 2014), we can derive a proof of the implicature that Barb cannot go to the party at that time which can resolve Alan’s question, as shown in Fig. 1. We use Mares (2004)’s logical notation where the curly brackets containing the indices of the premises used in that line. The proof in Fig. 1 shows how both premises are used to derive the conclusion in line 4, which itself uses the implicature in line 3.

While this seems better than a classical logic approach because redundancy is minimized, the problem remains that we still don’t have a handle on a real-valued relevance which could lead to a computational model of selecting relevant rules.

3 Towards Relevance Logic Lattices for Real-valued Relevance

To model the real-valued relevance of answers to questions and implicatures, we look to work by Knuth (2005) and (Hough and Purver, 2017) whereby a boolean algebra statement lattice like that in Fig. 3 in the Appendix allows real-valued probabilities to be assigned to the atoms of the lattice and then consequently to the joins of those elements. Questions are derived from this lattice as the joins of all the downsets of these elements. In such a framework in our example in Fig. 1, a relevance value is contingent on the the real-valued inclusion of statement 3 in statement 2 on the lattice after the application of the ‘work-party exclusion rule’ – if this is sufficiently high, we could rule this a relevant application of the general rule in order to derive 4.

While this seems to give us what we want, a problem of relevance remains, but this time in terms of the available answers to questions: in Knuth’s analysis, all questions can trivially evaluate to \perp . In fact \top in Knuth’s analysis is co-extensive with the entire space of questions and answers, which is counter-intuitive for any question with any content that does not involve asking whether something is true or false. We propose adopting a different underlying algebra

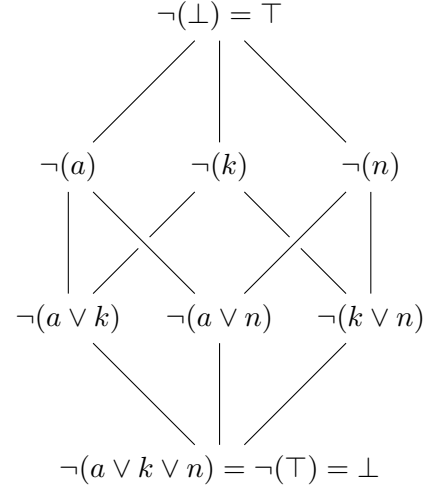


Figure 2: De Morgan lattice

which helps block these issues, and seems to capably model relevance both as a conversational implicature and as a logical consequence relation. We believe this can be achieved through a De Morgan lattice like Fig. 2 where the trivial results can be minimized and we can achieve a Relevant logic known as First-Degree Entailment (FDE). (Belnap, 1977) and their collaborators (Anderson et al., 2017) show how this can be achieved– see Fig. 4 Appendix for an illustration.

4 Future Work

In future work we would like to leverage the power of Knuth’s work on probability and information theory with question and statement lattices and the De Morgan lattices described above for deriving a real-valued relevance of a contribution resolving the central issue. We have evidence that Knuth’s approach can be generalised and De Morgan algebras are, in addition to being the backbone of FDE described above, investigated in fuzzy logic circles– for example forming an adequate algebraic semantics for a Lukasiewicz logic (Nguyen and Walker, 1996).

References

- Alan Ross Anderson, Nuel D Belnap Jr, and J Michael Dunn. 2017. *Entailment, Vol. II: The Logic of Relevance and Necessity*, volume 5009. Princeton University Press.
- Nuel D. Belnap. 1977. *A Useful Four-Valued Logic*, pages 5–37. Springer Netherlands, Dordrecht.
- Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis.
- Glenn Bruns and Michael Huth. 2011. Access control via belnap logic: Intuitive, expressive, and analyzable policy composition. *ACM Trans. Inf. Syst. Secur.*, 14(1):9:1–9:27.
- Jingde Cheng. 2004. Temporal relevant logic as the logic basis of anticipatory reasoning-reacting systems. In *AIP conference Proceedings*, volume 718, pages 362–375. AIP.
- Wayne Davis. 2014. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2014 edition. Metaphysics Research Lab, Stanford University.
- H Paul Grice, Peter Cole, Jerry Morgan, et al. 1975. Logic and conversation. 1975, pages 41–58.
- Julian Hough and Matthew Purver. 2017. Probabilistic record type lattices for incremental reference processing. In *Modern perspectives in type-theoretical semantics*, pages 189–222. Springer.
- Kevin H Knuth. 2005. Lattice duality: The origin of probability and entropy. *Neurocomputing*, 67:245–274.
- Edwin D Mares. 2004. *Relevant logic: A philosophical interpretation*. Cambridge University Press.
- Hung T. Nguyen and Elbert A. Walker. 1996. *A First Course in Fuzzy Logic*. CRC Press, Inc., Boca Raton, FL, USA.
- Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA.
- David Walker. 2005. Substructural type systems. *Advanced Topics in Types and Programming Languages*, pages 3–44.

Appendix

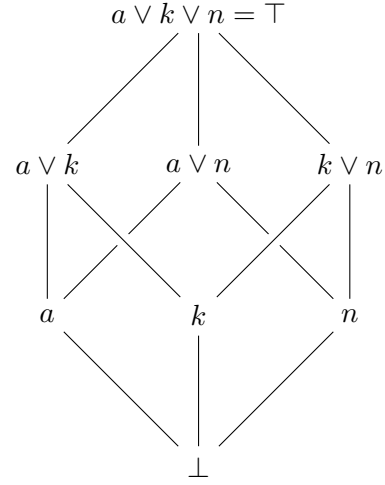


Figure 3: A Knuth-style lattice of statements for a Boolean algebra

$$\begin{array}{c}
 \frac{\Gamma \vdash \phi \quad \Gamma \vdash \psi}{\Gamma \vdash \phi \wedge \psi} \wedge I \\
 \\
 \frac{\Gamma \vdash \phi_i}{\Gamma \vdash \phi_1 \vee \phi_2} \vee I \quad (i \in \{1, 2\}) \\
 \\
 \frac{\Gamma \vdash \phi_1 \wedge \phi_2}{\Gamma \vdash \phi_i} \wedge E \quad (i \in \{1, 2\}) \\
 \\
 \frac{\Gamma \vdash \phi \vee \psi \quad \Delta, \phi \vdash \chi \quad \Delta, \psi \vdash \chi}{\Gamma, \Delta \vdash \chi} \vee E \\
 \\
 \frac{\Gamma \vdash \neg \neg \phi}{\Gamma \vdash \phi} \neg \neg E \\
 \\
 \frac{\Gamma \vdash \phi}{\Gamma \vdash \neg \neg \phi} \neg \neg I \\
 \\
 \frac{\Gamma \vdash \neg(\phi \vee \psi)}{\Gamma \vdash \neg \phi \wedge \neg \psi} \text{DeMorgan(i)} \\
 \\
 \frac{\Gamma \vdash \neg(\phi \wedge \psi)}{\Gamma \vdash \neg \phi \vee \neg \psi} \text{DeMorgan(ii)}
 \end{array}$$

Figure 4: FDE

Interactive visual grounding with neural networks

José Miguel Cano Santín¹ Simon Dobnik^{1,2} Mehdi Ghanimifard^{1,2}

¹Department of Philosophy, Linguistics and Theory of Science (FLoV)

²Centre for Linguistic Theory and Studies in Probability (CLASP)

University of Gothenburg, Sweden

¹jmcs990@gmail.com ²{simon.dobnik,mehdi.ghanimifard}@gu.se

Abstract

Training strategies for neural networks are not suitable for real time human-robot interaction. Few-shot learning approaches have been developed for low resource scenarios but without the usual teacher/learner supervision. In this work we present a combination of both: a situated dialogue system to teach object names to a robot from its camera images using Matching Networks (Vinyals et al., 2016). We compare the performance of the system with transferred learning from pre-trained models and different conversational strategies with a human tutor.

1 Introduction

Robotic systems need to acquire constantly new knowledge about their environment and the objects present in it as well as knowledge that they receive by interacting with their conversational partners. In grounding what they see in language they can benefit a lot by taking into account *how* such information is presented to them by the context of their perceptual environment and a human tutor (Skočaj et al., 2010). To this end, our situated dialogue system implements (i) different *dialogue interaction strategies* that exploit linguistic and interactive information from the dialogue and (ii) different *perceptual classification strategies* that exploit the interaction of the agent with its environment, namely the context of the previous objects and exposure to background knowledge. While our system includes both (i) and (ii) the focus of this paper is (ii). In particular, we examine how a deep neural network model can benefit from knowledge pre-trained on large data offline to learn new objects with little online data but with contextually provided background categories.

2 Method

Our situated dialogue system is based on the KILLE setup designed by Dobnik and de Graaf

(2017). It consists of a stationary Kinect v1 sensor in front of which a small individual object can be presented. The camera is connected to a Robot Operating System (ROS) framework (Quigley et al., 2009) running on Ubuntu 16.04 system using the *Freenect* driver. Our Python script which is implemented as a node within the ROS community takes care of both the object recognition and the dialogue management.

Dialogue Interaction Strategies Our situated dialogue system uses two different strategies to learn objects. The human tutor can present the object (e.g. *This is an apple*), in which case it will save the object in the dataset and, if necessary, it will retrain the model to learn a new label or ask for more about the object if it does not have enough images to learn. The robot can also be queried (e.g. *What is this?*). In this case, the robot will attempt to recognise the object presented and answer with the highest scoring label and the certainty of the guess depending on the score of that label (e.g. *This is an apple; I think this is an apple; I don't know what is this. Please, tell me.*). Then, the user can give feedback and confirm if the guess is true or tell the system the true label.

Interactive Visual Classification Our deep neural model consists of two independent modules. Firstly, there is a stack of the VGG16 CNN layers (Simonyan and Zisserman, 2014) pre-trained on ImageNet (Russakovsky et al., 2015). Here we test the advantages and disadvantages of the learning transferred from a large dataset to encode the images that the robot will perceive with its camera. Secondly, we present our implementation of Matching Networks (Vinyals et al., 2016) in a robot scenario with interactive grounding, which is the main contribution of this work. The idea of the matching network algorithm is to rapidly train a neural network from few observations. Each training instance consists of few (k)

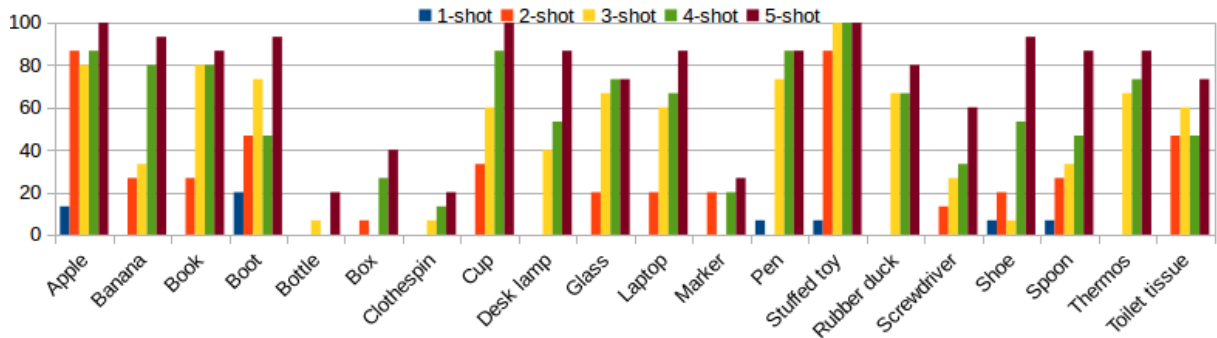


Figure 1: Results on learning new labels. The k -shot learned label is specified under the x axis and each of the bars represent the accuracy of the classification from 1-shot (left) to 5-shot (right).

images of each labelled class (n) that make up the support set S and a target image t belonging to one of these classes. The objective of training is to predict the class of the target image and therefore learn how to discriminate images of different classes. The contribution and the novelty of our work is the application of this model in an interactive robot training scenario where the classes of objects in the support set can be thought of as contextually present objects related to the current task in which the robot is being trained.

3 Experiments

Baseline In order to validate our implementation of the matching networks we use 20 categories from the test split of miniImageNet, which was created from ImageNet by Vinyals et al. (2016). Figure 2 shows that the accuracy of our system increases considerably when adding more images per label, as well as it becomes more difficult to classify correctly with more labels. An important aspect that needs to be considered in an interactive scenario is how long training and application of the model takes. Encode time does not seem to increase much with more images, while train time of the matching network is more clearly affected. However, being able to train a model for 20 labels in about 15 seconds and achieving an accuracy of 74.2% seems very reasonable.

Learning a New Class of Objects The objective of this experiment is to test how many images the system needs to see to learn new class labels. We collected a new small dataset of 20 categories and 20 images per category taken with the robot’s camera as the support set and target images. We simulate the learning process by training matching networks on 19 labels with five images each, which represent the categories that the robot al-

5 labels	1-shot	5-shot	10-shot
Accuracy	75.8%	89.8%	98.8%
Encode time	1.12s	1.63s	2.15s
Train time	1.43s	3.57s	7.27s
20 labels	1-shot	5-shot	10-shot
Accuracy	52.5%	74.2%	82.6%
Encode time	1.41s	1.93s	2.39s
Train time	3.26	12.15s	25.99s

Figure 2: Baseline results on miniImageNet. Encode time is the number of seconds to encode the support set (S) images with VGG16. Train time is the number of seconds to train the matching networks.

ready knows, and then adding the remaining label to the support set for training each model which is learned with 1 to 5 images in each case. Then, we evaluate the recognition accuracy of the new label on the remaining 15 images of the same label for each of the models. Figure 1 shows that four to five images are necessary for most of the labels to have a reliable object recognition. Also, some labels are clearly easier to learn than others, e.g. bottle, box, clothespin and marker vs. apple, book and stuffed toy, which did not need more than three images to get to about 80% accuracy.

4 Conclusion and Future Work

The observed results are promising and the system could be extended in multiple ways, which is the focus of our ongoing work. For instance, by using offline pre-trained knowledge also for the matching networks, new interactive strategies with the robot, attention over the visual regions objects to avoid the influence of the background objects and trying different techniques for selecting the images of the support set.

Acknowledgments

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

- Simon Dobnik and Erik de Graaf. 2017. [Kille: a framework for situated agents for learning language through interaction](#). In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 162–171, Gothenburg, Sweden. Association for Computational Linguistics.
- Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. 2009. [Ros: an open-source robot operating system](#). In *ICRA Workshop on Open Source Software*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [ImageNet Large Scale Visual Recognition Challenge](#). *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Karen Simonyan and Andrew Zisserman. 2014. [Very Deep Convolutional Networks for Large-Scale Image Recognition](#). *arXiv e-prints*, page arXiv:1409.1556.
- Danijel Skočaj, Miroslav Janiček, Matej Kristan, Geert-Jan M. Kruijff, Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich. 2010. [A basic cognitive system for interactive continuous learning of visual concepts](#). In *ICRA 2010 workshop ICAIR - Interactive Communication for Autonomous Intelligent Robots*, pages 30–36, Anchorage, AK, USA.
- Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. [Matching Networks for One Shot Learning](#). *arXiv e-prints*, page arXiv:1606.04080.

Bouletic and Deontic Modality and Social Choice

Sumiyo Nishiguchi

Center for Language Studies
Otaru University of Commerce, Japan
nishiguchi@res.otaru-uc.ac.jp

Abstract

This paper argues that bouletic and deontic modality has decision makers and involves social choice. Bouletic modal is participant-internal while possibly intransitive and disconnected.

1 Decision Maker in Bouletic/Deontic Modality

The state of *wanting something* reflects personal preference and involves personal decision making. In that sense, wanting act follows the Condition of Liberalism, which says that, no matter how other people oppose, personal decisions can be made on certain matters. A weak Condition of Liberalism à la (Sen 1970, 1979) is that each individual is entirely decisive in the social choice over at least a pair of alternatives. It is that everyone has a say on something no matter what other people think. In actuality, what we want may not come out due to restrictions, but wanting something is a liberal act.

To put things in the possible world semantics (Lewis 1973, among others), in the best possible worlds for a decision maker, her wants are fulfilled. Her want-worlds are the subset of the worlds where her wants are fulfilled. The meaning of Oliver's utterance in (1a) is expressed as in (1b) which says that, in all the accessible worlds which accord with Oliver's wants at world w_c , he watches *Sword and the Stone*.

- (1) a. Oliver: I want to watch *Sword and the Stone*. (BNC KDE 1917)
- b. $\forall w.[\text{BOUL}_s(w)(w_c) \rightarrow \text{watch-Sword-and-the-Stone}(s,w)]$ (s: speaker, w: world, w_c : actual world, BOUL_x : bouletic accessibility relation of the individual x)

From the perspective of decision making, if the speaker is the agent of wanting something, the speaker is the single decision maker regarding her preference. If the first person plural *we* wants something unanimously, the group members including the speaker are the decision makers as in (2b).

- (2) a. Oliver: We want to watch *Sword and the Stone*.

b. decision maker = {I, hearer}

Even when Oliver wants something different, the wanters Bill's desire remains unaffected, in (3).

- (3) a. Oliver: I want to watch *Sword and the Stone*

Nicola: Hm

Bill: Yes and *Pinocchio* and *Scooby Doo* and *Robin Hood*

Oliver: I don't like

Bill: And *Batman* and *Robin* and *Rescuers* and *Ghostbusters*.

(BNC KDE 1917-1920)

b. decision maker = {Bill}

- c. $pP_b s \wedge sP_o p \rightarrow pP_s$ (p: *Pinocchio*, s: *Sword and the Stone*, b: *Bill*, o: *Oliver*, $xP_i y$: x is strictly preferred to y by i)

In contrast, the decision maker of deontic modals such as *must*, *should*, and *ought to* differs from the attitude holder. Traffic laws are imposed on public by the lawmakers: therefore, the decision makers are not drivers but a lawgiver in (4). If a teacher decides that Oliver should submit a homework, she is the decision maker of the deontic modal, in (5). The decision that Oliver should study Spanish may be imposed due to the linguistic situation of people in Chile in (6).

- (4) a. We should follow traffic lights.
 b. decision maker = $\{x: \text{lawmaker}(x)\}$
- (5) a. Oliver should submit her homework.
 b. decision maker = $\{x: \text{instructor-of-Oliver}(x)\}$
- (6) a. Oliver should study Spanish, to communicate in Chile.
 b. decision maker = $\{X \subseteq \text{people in Chile}\}$

2 Previous Analyses

(van der Auwera and Plungian 1998) classify participant-internal and participant-external modality. Ability modal like *can* and necessity modals like *need* are participant-internal—the ability and necessity originates in the participants. Even though volition or bouletic modality are excluded from the core of modality, bouletic modality appears to be participant-internal. In *Mary wants to play the piano*, the desire originates in the attitude holder *Mary*.

3 Incorporating Decision Makers

Now that bouletic and deontic modals depend on decision makers, the accessibility relations between possible worlds depend on decision makers. When the group preference is involved, the group members' social decision is reflected.

- (7) a. Oliver wants to watch Sword and the Stone.
 b. $\forall w. [\text{BOUL}_o(w)(w_c) \rightarrow \text{watch-Sword-and-the-Stone}(o)(w)]$
- (8) a. We want to watch Robin Hood.
 b. $\forall w. [\text{BOUL}_{s,h}(w)(w_c) \rightarrow \text{watch-Robin-Hood}(s,h)(w)]$ (s: speaker, h: hearer)
- (9) a. Oliver should submit homework.
 b. $\forall w. [\text{DEON}_i(w)(w_c) \rightarrow \text{submit-homework}(o)(w)]$

Such incorporation of modal judges may be reminiscent of (Stephenson 2007)'s analysis on epistemic modality, built on (Laserson 2005) on predicates of personal taste such as *fun* and *tasty*. I further claim that bouletic and deontic modals have decision makers. It is related to (von Finckel 1999) who incorporates the want argument α (cf. Kratzer 1981, Heim 1992).

4 Social Choice

Group decision is a social choice (Arrow 1963, Sen 1979, Chevalerey et al. 2007). The social choice function SCF returns a single choice, which is going to a movie. The decision may not be unanimous but follows Pareto principle, in that when nobody has contrary preference, the mass decision agrees with individual's preferences.

- (10) a. decision makers $I = \{o, b, n\}$
 b. alternatives $\chi = \{\text{Sword and the Stone, Pinocchio, Robin Hood}\}$
 c. A profile, a vector of linear orders, or preference $R = (R_o, R_b, R_n) \in L(\chi)^3$
 d. Social Choice Function $\text{SCW}(L(\chi)^3) = \{\text{Sword and the Stone}\}$

Also Independence of Irrelevant Alternatives is adhered because the relative ranking between Pinocchio and other alternatives only matters to the group decision even with cyclicity.

- (11) a. $sR_o pR_o r \wedge sR_b pR_b r \rightarrow sR pR r$
 b. $sR_o bR_o pR_o r \wedge bR_b sR_b eR_b h \rightarrow sR pR r$

The domain of the Social Choice Function should be restricted because the order may not be transitive. Subjective and personal preference can order alternatives cyclicly and intransitively as in (12a). Moreover, some elements in the domain may not be connected with preference relation. Some movies may not be compared with other movies. The utterance (12b) is perfectly plausible.

- (12) a. I want to watch Sword and the Stones more than Scooby Doo. I want Scooby Doo than Robin Hood but Batman to Sword and the Stones.
 b. I like Sword and the Stones better than Pinocchio. I do not know about the Rescuers.

Thus, deontic/bouletic modals have decision makers and bouletic modal is participant-internal. Bouletic modality can be intransitive and disconnected even though Pareto condition and IIA applies, in harmony with Arrow's Impossibility Theorem.

Acknowledgment

This work is partially supported by JSPS KAKENHI Grant Number 16K02643.

References

- Kenneth J. Arrow. 1963. *Social Choice and Individual Values*, 2 edition. Yale University Press, New Haven.
- Johan van der Auwera and Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology*, 2:79–124.
- Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. 2007. A short introduction to computational social choice. In *Proceedings of the 33rd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM-2007)*. Springer-Verlag, Berlin, Heidelberg.
- Kai von Fintel. 1999. NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics*, 16:97–148.
- Irene Heim. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9:183–221.
- Angelika Kratzer. 1981. The notional category of modality. In H.J. Eikmeyer and H. Rieser, editors, *Words, Worlds, and Contexts*, pages 38–74. de Gruyter, Berlin, New York.
- Peter Lasnik. 2005. context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, 28:643–686.
- David Lewis. 1973. *Counterfactuals*. Blackwell, Oxford.
- Amartya K. Sen. 1970. The impossibility of a paretian liberal. *Journal of Political Economy*, 78(1):152–157.
- Amartya K. Sen. 1979. *Collective Choice and Social Welfare*. North-Holland, Amsterdam.
- Tamara Stephenson. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, 30:487–525.

Towards a formal model of word meaning negotiation

Bill Noble, Asad Sayeed, and Staffan Larsson

Centre for Linguistic Theory and Studies in Probability (CLASP)

Department of Philosophy, Linguistics, and Theory of Science

University of Gothenburg, Sweden

{bill.noble, asad.sayeed, staffan.larsson}@gu.se

Abstract

We introduce a model of the interactive semantics of word meaning negotiation (WMN). We represent a WMN as a growing graph whose nodes are semantic *anchors* and edges are proposed (agent-specific) semantic *links* between them.

Word meaning negotiation is a conversational routine in which speakers explicitly discuss the meaning of a word or phrase. WMNs occur when one participant disagrees with or doesn't understand what a speaker meant by a particular word or phrase. Such a discrepancy represents a breakdown in the alignment of participants' lexico-semantic resources.

1 Background

Although WMN has not received a great deal of attention as such, it has been addressed in the language acquisition literature (e.g., Varonis and Gass, 1985; Clark, 2007) and in psycholinguistic research on semantic alignment (Brennan and Clark, 1996; Metzing and Brennan, 2003).

Myrendal (2015) gives an in-depth qualitative analysis of WMN in Swedish online discussion forums. We seek to model two key findings from that work. First, we aim to capture the distinction between WMNs originating in non-understanding (NON) and those originating in disagreement (DIN). Myrendal (2015, §3.4.1) finds that the source of the discrepancy plays an important role in the trajectory of the WMN. Second, we would like to define *semantic operations* (Myrendal, 2015, §4.5 & 5.6) as actions within the framework of our model and predict the results of those actions.

Along these lines, Larsson and Myrendal (2017) give a Type Theory with Records (TTR) formalization of updates carried out by semantic

operations. Where that formalization is restricted to updates resulting from *accepted* semantic operations in isolation, our model seeks to capture the interactive features of WMNs, including rejected proposals and sequences of semantic operations.

The Trigger-Indicator-Response (TIR) model (Varonis and Gass, 1985) captures the discourse structure of WMNs¹, which is an important prerequisite to understanding their semantics. It identifies three utterances that characterize WMNs: A *trigger*, by speaker *A*, which includes a lexical item (the *trigger word*) that is not understood by speaker *B*, an *indicator*, in which *B* signals their non-understanding (or disagreement) of the trigger word, and a *response*, in which *A* overtly acknowledges the non-understanding.

	Speaker	Utterance
U1	a	I have a whistle, 5 dollars...
U2	b	A whistle?
U3	a	It's to make noise with your mouth when you need help... do you know?
U4	b	Oh yeah, it's good.

Example 1: From Yanguas (2010, p. 78) with trigger (U1), indicator (U2), response (U3), and reply to the response (U4).

2 Model

We model a WMN as a growing, rooted, labeled, graph whose nodes are meaningful units called *semantic anchors*, and edges are proposed (speaker-specific) *links* between those anchors. Speaker contributions create new anchors, create links between anchors, and change the relation expressed

¹The TIR model was designed for NONs, though some of the same concepts carry over to DINs.

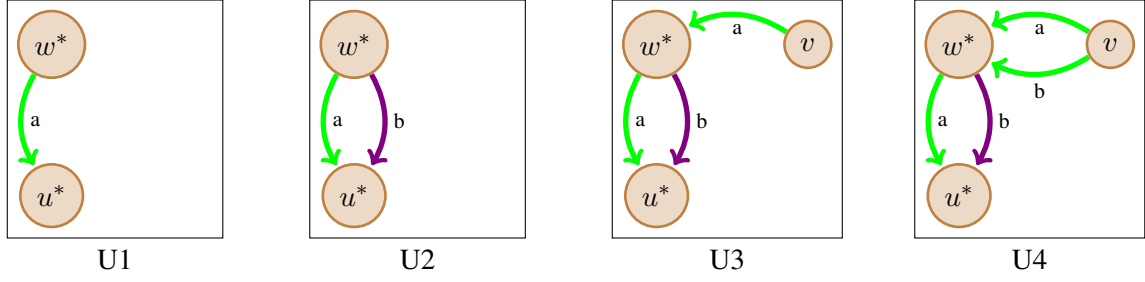


Figure 1: WMN model for Example 1. Link color indicates the semantic relation (green = +, violet = ?). w^* = “whistle”, u^* = the situation under discussion, v = “to make noise with your mouth...”

by existing links. In this way, we seek to capture the intuition that speakers jointly “triangulate” the meaning of a target term by situating it in relation to other agreed-upon meanings.

Formally, a WMN of consisting of T utterances between a set of speakers, S , about target term, w^* , is given by:

$$G^t = \langle N, w^*, L, \{R_a\}_{a \in S} \rangle_{t \leq T}$$

where N is the set of anchors introduced by the agents, L is a set of semantic relations, and each $R_a : N \times N \rightarrow L$ gives the kind of semantic relation (if any) posed by a .

For now, we assume three semantic relations: $L = \{+, -, ?\}$. Roughly, $R_a(u, v) = +$ means a asserts that u *applies to* v and $R_a(u, v) = -$ means that a asserts u *does not apply to* v . If a raises the question of the semantic relation between u and v without making an assertion, we write $R_a(u, v) = ?$. Note that this is a directed relation: $R_a(u, v) = +$ is different from $R_a(v, u) = +$, and links (possibly with different semantic relations) may exist in both directions. More precisely, we use $R_a(u, v) = +$ when a asserts that u is a *partial definition* (supplying necessary but not sufficient conditions) for v , or that v is an *example of* u .²

In contrast to Larsson and Myrendal (2017), this model captures WMNs at the level of *understanding* (Clark and Schaefer, 1989). Grounding at the level of *uptake* is achieved when $R_a(u, v) = R_b(u, v)$ for all $a, b \in S$.

2.1 Semantic operations

Speaker contributions can add any number of semantic anchors and/or links, or change the rela-

²Depending on the underlying semantic representation, this overloading may be problematic. In TTR, both partial definitions and (verbal) examples correspond to the *subtype* relation (\sqsubseteq), but examples given by demonstration are more adequately modeled by the *of type* relation ($:$).

tion expressed by existing links. As a result, G is monotone increasing, that is; for each $t \leq T$, $N^t \subseteq N^{t+1}$ and $\text{Dom}(R_a^t) \subseteq \text{Dom}(R_a^{t+1})$.

Now we can define some of the semantic operations from Myrendal (2015) in terms of the model:

- **exemplify** – u is an example of v
 - create a new anchor, u (the example)
 - create a link $R_a(v, u) = +$
- **explicate** – u is a (partial) definition of v
 - create a new anchor u (the explication)
 - create a link $R_a(u, v) = +$
- **endorse** – u is a v
 - create a link $R_a(u, v) = +$ between existing anchors u and v .
- **meta-linguistic CR** – what do you mean by u ?
 - create a link $R_a(u, v) = ?$ between existing anchors, u and v

This is not an exhaustive list, but demonstrates how semantic operations can be defined in terms of the atomic actions offered by the model.

3 Future Work

There are two main lines of future work. First, the model should define *semantic updates* based on the state of the graph (i.e., taking the entire sequence of semantic operations into account). This would achieve our goal of giving an interactive update semantics for word meaning negotiation. Second, we intend to develop an annotation schema for semantic operations from which we can derive the WMN graph. From there, we can test the adequacy of the model by making predictions about how agents will use negotiated terms in the future.

References

- Susan E Brennan and Herbert H Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.
- Eve V. Clark. 2007. Young Children's Uptake of New Words in Conversation. *Language in Society*, 36(2):157–182.
- Herbert H. Clark and Edward F. Schaefer. 1989. [Contributing to discourse](#). *Cognitive Science*, 13(2):259–294.
- 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.
- Charles Metzing and Susan E Brennan. 2003. [When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions](#). *Journal of Memory and Language*, 49(2):201–213.
- Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. PhD Thesis, University of Gothenburg, University of Gothenburg.
- E. M. Varonis and S. Gass. 1985. [Non-native/Non-native Conversations: A Model for Negotiation of Meaning](#). *Applied Linguistics*, 6(1):71–90.
- Inigo Yanguas. 2010. Oral Computer-Mediated Interaction between L2 Learners: It's about Time! *Language Learning & Technology*, 14(3):72–93.

Towards Multimodal Understanding of Passenger-Vehicle Interactions in Autonomous Vehicles: Intent/Slot Recognition Utilizing Audio-Visual Data

Eda Okur Shachi H Kumar Saurav Sahay Lama Nachman

Intel Labs, Anticipatory Computing Lab, USA

{eda.okur, shachi.h.kumar, saurav.sahay, lama.nachman}@intel.com

1 Introduction

Understanding passenger intents from spoken interactions and car’s vision (both inside and outside the vehicle) are important building blocks towards developing contextual dialog systems for natural interactions in autonomous vehicles (AV). In this study, we continued exploring AMIE (Automated-vehicle Multimodal In-cabin Experience), the in-cabin agent responsible for handling certain multimodal passenger-vehicle interactions. When the passengers give instructions to AMIE, the agent should parse such commands properly considering available three modalities (language/text, audio, video) and trigger the appropriate functionality of the AV system. We had collected a multimodal in-cabin dataset with multi-turn dialogues between the passengers and AMIE using a Wizard-of-Oz scheme via realistic scavenger hunt game.

In our previous explorations (Okur et al., 2018, 2019), we experimented with various RNN-based models to detect utterance-level intents (set destination, change route, go faster, go slower, stop, park, pull over, drop off, open door, and others) along with intent keywords and relevant slots (location, position/direction, object, gesture/gaze, time-guidance, person) associated with the action to be performed in our AV scenarios.

In this recent work, we propose to discuss the benefits of multimodal understanding of in-cabin utterances by incorporating verbal/language input (text and speech embeddings) together with the non-verbal/acoustic and visual input from inside and outside the vehicle (i.e., passenger gestures and gaze from in-cabin video stream, referred objects outside of the vehicle from the road view camera stream). Our experimental results outperformed text-only baselines and with multimodality, we achieved improved performances for utterance-level intent detection and slot filling.

2 Methodology

We explored leveraging multimodality for the NLU module in the SDS pipeline. As our AMIE in-cabin dataset¹ has video and audio recordings, we investigated 3 modalities for the NLU: text, audio, and video. For text (language) modality, our previous work (Okur et al., 2019) presents the details of our best-performing Hierarchical & Joint Bi-LSTM models (Schuster and Paliwal, 1997; Hakkani-Tur et al., 2016; Zhang and Wang, 2016; Wen et al., 2018) (H-Joint-2, see A) and the results for utterance-level intent recognition and word-level slot filling via transcribed and recognized (ASR output) textual data, using word embeddings (GloVe (Pennington et al., 2014)) as features. This study explores the following multimodal features:

Speech Embeddings: We incorporated pre-trained speech embeddings (Speech2Vec (Chung and Glass, 2018)) as features, trained on a corpus of 500 hours of speech from LibriSpeech. Speech2Vec² is considered as a speech version of Word2Vec (Mikolov et al., 2013) which is compared with Word2Vec vectors trained on the transcript of the same speech corpus. We experimented with concatenating word and speech embeddings by using pre-trained GloVe embeddings (6B tokens, 400K vocab, dim=100), Speech2Vec embeddings (37.6K vocab, dim=100), and its Word2Vec counterpart (37.6K vocab, dim=100).

Audio Features: Using openSMILE (Eyben et al., 2013), 1582 audio features are extracted for each utterance using the segmented audio clips from in-cabin AMIE dataset. These are the INTERSPEECH 2010 Paralinguistic Challenge features (IS10) including PCM loudness, MFCC, log Mel Freq. Band, LSP, etc. (Schuller et al., 2010).

¹Details of AMIE data collection setup in (Sherry et al., 2018; Okur et al., 2019); in-cabin dataset statistics in A.

²github.com/iamyuanchung/speech2vec-pretrained-vectors

Modalities	Features (Embeddings)	Intent Recognition			Slot Filling		
		Prec	Rec	F1	Prec	Rec	F1
Text	GloVe (400K)	89.2	89.0	89.0	95.8	95.8	95.8
Text	Word2Vec (37.6K)	86.4	85.2	85.6	93.3	93.4	93.3
Audio	Speech2Vec (37.6K)	85.1	84.4	84.5	93.2	93.3	93.1
Text & Audio	Word2Vec + Speech2Vec	88.4	88.1	88.1	94.2	94.3	94.2
Text & Audio	GloVe + Speech2Vec	91.1	91.0	90.9	96.3	96.3	96.3
Text & Audio	GloVe + Word2Vec + Speech2Vec	91.5	91.2	91.3	96.6	96.6	96.6

Table 1: Speech Embeddings Experiments: Precision/Recall/F1-scores (%) of NLU Models

Modalities	Features	Prec	Rec	F1
Text	Embeddings (GloVe)	89.19	89.04	89.02
Text & Audio	Embeddings (GloVe) + Audio (openSMILE/IS10)	89.69	89.64	89.53
Text & Video	Embeddings (GloVe) + Video_cabin (CNN/Inception-ResNet-v2)	89.48	89.57	89.40
Text & Video	Embeddings (GloVe) + Video_road (CNN/Inception-ResNet-v2)	89.78	89.19	89.37
Text & Video	Embeddings (GloVe) + Video_cabin+road (CNN/Inception-ResNet-v2)	89.84	89.72	89.68
Text & Audio	Embeddings (GloVe+Word2Vec+Speech2Vec)	91.50	91.24	91.29
Text & Audio	Embeddings (GloVe+Word2Vec+Speech2Vec) + Audio (openSMILE)	91.83	91.62	91.68
Text & Audio & Video	Embeddings (GloVe+Word2Vec+Speech2Vec) + Video_cabin (CNN)	91.73	91.47	91.50
Text & Audio & Video	Embeddings (GloVe+Word2Vec+Speech2Vec) + Video_cabin+road (CNN)	91.73	91.54	91.55

Table 2: Multimodal (Audio & Video) Features Exploration: Precision/Recall/F1-scores (%) of Intent Recognition

Video Features: Using the feature extraction process described in (Kordopatis-Zilos et al., 2017), we extracted intermediate CNN features³ for each segmented video clip from AMIE dataset. For any given input video clip (segmented for each utterance), one frame per second is sampled and its visual descriptor is extracted from the activations of the intermediate convolution layers of a pre-trained CNN. We used the pre-trained Inception-ResNet-v2 model⁴ (Szegedy et al., 2016) and generated 4096-dim features for each sample. We experimented with adding 2 sources of visual information: (i) cabin/passenger view from the Back-Driver RGB camera recordings, (ii) road/outside view from the DashCam RGB video streams.

3 Experimental Results

For incorporating speech embeddings experiments, performance results of NLU models on in-cabin data with various feature concatenations can be found in Table 1, using our previous hierarchical joint model (H-Joint-2). When used in isolation, Word2Vec and Speech2Vec achieves comparable performances, which cannot reach GloVe performance. This was expected as the pre-trained Speech2Vec vectors have lower vocabulary coverage than GloVe. Yet, we observed that concatenating GloVe + Speech2Vec, and further GloVe + Word2Vec + Speech2Vec yields better NLU results: F1-score increased from 0.89 to 0.91 for intent recognition, from 0.96 to 0.97 for slot filling.

³github.com/MKLab-ITI/intermediate-cnn-features

⁴github.com/tensorflow/models/tree/master/research/slim

For multimodal (audio & video) features exploration, performance results of the compared models with varying modality/feature concatenations can be found in Table 2. Since these audio/video features are extracted per utterance (on segmented audio & video clips), we experimented with the utterance-level intent recognition task only, using hierarchical joint learning (H-Joint-2). We investigated the audio-visual feature additions on top of text-only and text+speech embedding models. Adding openSMILE/IS10 features from audio, as well as incorporating intermediate CNN/Inception-ResNet-v2 features from video brought slight improvements to our intent models, reaching 0.92 F1-score. These initial results using feature concatenations may need further explorations, especially for certain intent-types such as stop (audio intensity) or relevant slots such as passenger gestures/gaze (from cabin video) and outside objects (from road video).

4 Conclusion

In this study, we present our initial explorations towards multimodal understanding of passenger utterances in autonomous vehicles. We briefly show that our experimental results outperformed certain baselines and with multimodality, we achieved improved overall F1-scores of 0.92 for utterance-level intent detection and 0.97 for word-level slot filling. This ongoing research has a potential impact of exploring real-world challenges with human-vehicle-scene interactions for autonomous driving support with spoken utterances.

References

- Yu-An Chung and James Glass. 2018. [Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech](#). In *Proc. INTERSPEECH 2018*, pages 811–815.
- Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. [Recent developments in opensmile, the munich open-source multimedia feature extractor](#). In *Proc. ACM International Conference on Multimedia, MM '13*, pages 835–838.
- Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Vivian Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. [Multi-domain joint semantic frame parsing using bi-directional rnn-lstm](#). ISCA.
- Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. 2017. [Near-duplicate video retrieval by aggregating intermediate cnn layers](#). In *International Conference on Multimedia Modeling*, pages 251–263. Springer.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, pages 3111–3119, USA.
- Eda Okur, Shachi H Kumar, Saurav Sahay, Asli Arslan Esme, and Lama Nachman. 2018. [Conversational intent understanding for passengers in autonomous vehicles](#). *13th Women in Machine Learning Workshop (WiML 2018), co-located with the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.
- Eda Okur, Shachi H Kumar, Saurav Sahay, Asli Arslan Esme, and Lama Nachman. 2019. [Natural language interactions in autonomous vehicles: Intent detection and slot filling from passenger utterances](#). *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP'14)*.
- Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. [The interspeech 2010 paralinguistic challenge](#). In *Proc. INTERSPEECH 2010*.
- M. Schuster and K.K. Paliwal. 1997. [Bidirectional recurrent neural networks](#). *Trans. Sig. Proc.*, 45(11):2673–2681.
- John Sherry, Richard Beckwith, Asli Arslan Esme, and Cagri Tanriover. 2018. [Getting things done in an autonomous vehicle](#). In *Social Robots in the Wild Workshop, 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2018)*.

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. [Inception-v4, inception-resnet and the impact of residual connections on learning](#). *CoRR*, abs/1602.07261.

Liyun Wen, Xiaojie Wang, Zhenjiang Dong, and Hong Chen. 2018. [Jointly modeling intent identification and slot filling with contextual and hierarchical information](#). In *Natural Language Processing and Chinese Computing*, pages 3–15, Cham. Springer.

Xiaodong Zhang and Houfeng Wang. 2016. [A joint model of intent determination and slot filling for spoken language understanding](#). In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16*, pages 2993–2999.

A Appendices

AMIE In-cabin Dataset: We obtained 1331 utterances having commands to AMIE agent from our in-cabin dataset. Annotation results for *utterance-level intent types, slots and intent keywords* can be found in Table 3 and Table 4.

AMIE Scenario	Intent Type	Utterance Count
Set/Change Destination/Route	SetDestination	311
	SetRoute	507
Finishing the Trip	Park	151
	PullOver	34
	Stop	27
Set/Change Driving Behavior/Speed	GoFaster	73
	GoSlower	41
Others (Door, Music, A/C, etc.)	OpenDoor	136
	Other	51
Total		1331

Table 3: AMIE In-cabin Dataset Statistics: Intents

Slot/Keyword Type	Word Count
Intent Keyword	2007
Location	1969
Position/Direction	1131
Person	404
Time Guidance	246
Gesture/Gaze	167
Object	110
None	6512
Total	12546

Table 4: AMIE In-cabin Dataset Statistics: Slots

Hierarchical & Joint Model (H-Joint-2): 2-level hierarchical joint learning model that detects/extracts *intent keywords & slots* using seq2seq Bi-LSTMs first (Level-1), then only the words that are predicted as *intent keywords & valid slots* are fed into Joint-2 model (Level-2), which is another seq2seq Bi-LSTM network for *utterance-level intent detection* (jointly trained with *slots & intent keywords*) (Okur et al., 2019).

Pronominal Ambiguity Resolution in Spanish Child Dialogue: A Corpus Based Developmental Language Acquisition Approach.

Martha Robinson
Department of Linguistics
University of Edinburgh
martha.robinson94@gmail.com

Abstract

This investigation is a CHILDES (MacWhinney, 2000) corpus based study of how 3-4 and 5-6 year old monolingual Spanish speaking children learn to disambiguate null and overt pronominal reference from the input they are exposed to whilst engaging in adult-child dyadic communicative acts, as they are involved in the active development of pragmatic awareness. It was found, that although there was no significant difference between both groups in terms of tokens belonging to either pronominalization strategy in the input the children received, there was, however, a difference in the types of lexical verbs and modes of child-adult interaction at each developmental step.

1 Introduction

In the literature, it is often mentioned that pronouns have no fixed reference and their interpretation is highly context dependent (Kempson et al., 2001). In fact, more recent psycholinguistic experimental work has shown that syntactic, pragmatic and discursive factors figure prominently in their interpretation (Stevenson et al., 1994; Kehler, 2002; Kehler et al., 2008; Kehler and Rohde, 2015; inter alia). From an early age, children are able to produce personal pronominal forms correctly and their acquisition has been said to be closely embedded in early communicative experiences (Salazar Orvig et al., 2010). Nevertheless, comparison of adults and older children has shown that there are considerable differences in the mastery of their appropriate use among children and that many so called failures are in fact, developmental (Song and Fisher, 2007).

2 Null and Overt Pronominals in Spanish.

From a typological perspective, Spanish has been described as belonging to the group of languages licencing a partial *pro*-drop strategy (and related to Italian, Portuguese, Catalan, among others). This means, that subject pronominal expressions tend to be phonetically null and, dropped readily. In the literature, this has been related rich verbal morphology. However, Spanish also presents an overt pronominal counterpart as shown in (a) with the third person stressed form *Él* (he).

- a. Felix_i le pegó a Max_j y luego Ø_i/ÉL_j le pegó a Pedro.
Felix hit Max and then Ø_i/HE_j hit Pedro.

Here, the null form in the second conjunct co-refers naturally with the higher subject *Felix*. In contrast, the overt stressed form *él* (he) shows a marked natural preference to attach to the lower object *Max*. The alternation between overt and null pronouns has been at the centre of a great deal of debate in the linguistics literature for decades, especially in these inter-sentential instances of anaphoric co-reference (RAE, 1999). In fact, a prevalent view is that they occur in complementary distribution and display a division of labour strategy, a position that has also been widely adopted in more recent experimental psycholinguistics literature (as in Carminati, 2002 for Italian and Alonso-Ovalle, et al., 2002 for Spanish, for example). Although (a) appears to be a clear example of a strict division of labour strategy, it is also the case, that these types of anaphoric relations are often affected by verb semantics as we see in (b) and (c) with IMPLICIT CAUSALITY verbs such as *asustar* (frighten) and *temer* (fear):

- b. María_i asusta a Ana_j porque Ø_i/ELLA_j es; antipática.
María frightens Ana because Ø_i/SHE_j is horrible.
c. María_i teme a Ana_j porque Ø_j/ELLA_j es antipática.
María fears Ana because Ø_j/SHE_j is horrible.

In contrast to (a) above with a verb like *pegar* (hit) where null and overt forms appear to enter into an either/or relation, here both are able to enter into parallel coreference relations. In (b) both forms co-refer with *María* and in (c) with *Ana*. And here, the fact that a null form can actually co-refer with a lower object and the overt form with the higher subject NP, is particularly unexpected and undermines the strict division of labour perspective often proposed in the literature. Other factors disputing this perspective are issues of dialectal variation and formal syncretism in certain verbal paradigms (especially between 1st and 3rd person singular verbal inflections), in spoken varieties of Spanish where the appearance of overt pronouns is favoured in order to disambiguate reference to the subject. And finally, more general pragmatic and discursive principles also figure prominently in Spanish (in line with Kehler, 2002; Kehler et al., 2008; Kehler and Rohde, 2015, inter alia, on pronominal co-reference in English).

Experimental work with children both in English and Spanish has shown that they display a marked preference for lower attachment coreferential interpretations (Hartshorne, et al., 2014, for English and Ruigendijk, et al., 2011 for Spanish). However, in a study involving 5-6 year old children and adults conducted by Kehler, Hayes and Barner, (2011) involving *Transfer of Possession* and *Transfer of Location* verbs in English, a main effect of verb type and age was found. Although children are already highly adept at knowing how discourse works from an early age (Clark, 2005), the anaphoric value of pronominal expressions is first acquired by being involved in dialogue, before it is extended to monological uses (Salazar Orvig et al., 2010). Therefore, since the fundamental skills underlying the communicative process develop gradually between the ages of 3 and 5, until these become the highly sophisticated conversational acts of adults, the main question here, is how children learning Spanish achieve this task from the input they are exposed to, especially since they have to acquire two pronominalization strategies. We contend here that it is not only the input children are exposed to, but also how children conduct themselves in communicative exchanges that aid the development of anaphoric interpretations.

3 Method

We took a sample of adult-child (investigator or parent) dyadic interactions from three Iberian Spanish monolingual child dialogue corpora and created a sub-corpus (27, 277 total number of words) based on 100 utterances per child (approx. 200 adult-child turns) on which we calculated their Mean Length Utterance (Brown, 1973) or MLU_w (ie. MLU measured in words, as discussed in Ezeizabarrena and García Fernández, 2018), totalling 3972 turns. The dialogues involved traditional interactive story telling sessions (we excluded monological narratives) as well as communicative acts recounting children's daily routines at school and at home. The children were then matched for socio-economic status, MLU_w and strict monolingual Spanish linguistic background (ie. excluding familial bilinguals and monolingual and multilingual children cohabiting in areas with linguistic minority languages). Samples were collected at two developmental steps 3-4 (N=10) and 5-6 (N=8) years old, identified as Group A and Group B respectively. We excluded adult-adult exchanges at this point of the investigation.

4 Results

In the first instance, we found an overwhelming marked preference for the null form of the pronominal in adult-child interactions in both groups with only a few instances of the overt form and this was fairly consistent among both age groups (Group A: null 87.6% vs overt 12.4% and Group B: null 89.25% vs overt 10.75%). This is consistent with the fact that in a language like Spanish, the null form is considered to be the default and the overt, the exception (albeit the overt appears in certain highly predictive contexts). For the overt pronominal we found that adult-child ratio showed no significant difference either (Group A: adult 48.95% vs child 51.05% and Group B: adult 48.54% vs child 49.5%) and this means that children are exposed to and have already learned of the availability of the overt form from a very young age. However, the difference can be seen in the types of lexical verbs and adult-child interactional strategies utilised whilst engaging in communicative exchanges which differ at both developmental steps, an issue that nevertheless, merits further investigation.

References

- Roger Brown. 1973. *A first Language: The early stages*. Cambridge, Harvard University.
- Maria Nella Carminati. 2002. The processing of Italian subject pronouns. PhD Thesis, University of Massachusetts, Amherst.
- Eve Clark. 2005. *Pragmatics and Language Acquisition*. The Handbook of Pragmatics. Laurence R. Horn and Gregory Ward (eds.), Blackwell.
- María José Ezeizabarrena and Iñaki García Hernández. 2017. Length of Utterance, in morphemes or in words?: MLU3-w, a reliable measure of language development in Early Basque. *Frontiers in Psychology*, 8, 2265: <https://doi.org/10.3389/fpsyg.2017.02265>
- Theres Gruter, Hannah Rohde and Amy J. Schafer. 2017. Coreference and discourse coherence in L2: The Roles of grammatical aspect and referential form. *Linguistic approaches to Bilingualism*, 7(2), 199-229.
- Joshua K. Hartshorne and Jesse Snedeker. (unknown date). *Integration of discourse and semantic structure in children's resolution of ambiguous pronouns*. https://d2dg4e62b1gc8m.cloudfront.net/pub_pdfs/HartshorneSnedeker_BUCLD09_abstract.pdf
- Joshua K. Hartshorne, Rebecca Nappa, and Jesse Snedeker. 2014. Development of the first-mention bias. *Journal of Child Language*, 42(2), 423-446.
- Andrew Kehler. (2002). *Coherence, reference and the theory of grammar*. Stanford, CA: CLSI Publications.
- Andrew Kehler, Laura Kertz, Hannah Rohde, Jeffrey L. Elman. 2008. Coherence and coreference revisited. *Journal of Semantics*, 25, 1-44.
- Andrew Kehler and Hannah Rohde. 2015. Pronominal reference and pragmatic enrichment: A Bayesian account. In Proceedings of the 37th Annual Conference of the Cognitive Science society, pp. 1063.
- Andrew Kehler, Emily Hayes and David Barner. 2011. *Pragmatically driven biases in children's pronoun interpretation*. Poster presented at the 24th CUNY Conference on Human Sentence Processing, March 24-26, 2011, Stanford.
- Ruth Kempson, Wilfried Meyer-Viol, Dov Gabbay. 2001. *Dynamic Syntax: The flow of language understanding*. Wiley-Blackwell.
- Evan Kidd, Andrew Stewart and Ludovica Serratrice. 2011. Children do not overcome lexical biases where adults do: the role of referential scene in garden-path recovery. *Journal of Child Language*, 38 (222-234).
- Brian MacWhinney. 2000. The CHILDES Project: Tools for analysing talk, 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.
- RAE: *Gramática Descriptiva del Español*. 1999. Violeta Demonte and Ignacio Bosque (eds.), Real Academia Española, colección Nebrija y Bello. Espasa Calpe.
- Esther Ruigendijk, Sergio Baauw, Shalom Zuckerman, Nada Vasic, Joke de Lange, Sergei Avrutin. 2011. A Cross-linguistic Study on the interpretation of Pronouns by Children and Agrammatic Speakers: Evidence from Dutch, Spanish and Italian. In Edward Gibson and Neal J. Pearlmutter (eds.), *The Processing and Acquisition of Reference*. MIT Press, Cambridge Massachusetts.
- Anne Salazar Orvig, Haydée Marcos, Aliyah Morgenstern, Rouba Hassan, Jocelyne Leber-Marin, Jacques Parès. 2011. Dialogical beginnings of anaphora: The use of third person pronouns before the age of 3. *Journal of Pragmatics* 42, 1842-1865.
- Hyun-joo Song and Cynthia Fisher. 2007. Discourse prominence effects on 2.5-year old children's interpretation of pronouns. *Lingua*, 117(11), 1959-1987.
- Stevenson, R., Crawley, R. & Kleinman, D. (1994): Thematic roles, focusing and the representation of events. *Language and Cognitive Processes*, 9, 519-548.

Eye gaze in interaction: towards an annotation scheme for dialogue

Vidya Somashekarappa, Christine Howes and Asad Sayeed

Centre for Linguistic Theory and Studies in Probability (CLASP)

Department of Philosophy, Linguistics and Theory of Science

University of Gothenburg

{vidya.somashekarappa, christine.howes, asad.sayeed}@gu.se

Abstract

This paper proposes an approach to annotating eye gaze in natural dialogue which takes into account both social and referential functions of eye gaze and how they interact. The goal of this research is to provide a basis for robot or avatar models which communicate with humans using multimodal natural dialogue.

1 Introduction

Linguists and psychologists have shown a long standing interest in non-verbal communication relating to speech and gesture, including eye-gaze, which is the focus of this work (Kendon, 1967; Argyle and Cook, 1976; Goodwin, 1980, 1981).

1.1 Social functions of eye gaze in dialogue

Argyle and Cook (1976) showed that listeners display longer sequences of uninterrupted gaze towards the speaker, while speakers tended to shift their gaze towards and away from the listener quite often. Later work has refined this observation, with, for instance (Rossano, 2012) noting that this distributional pattern is dependent on the specific interactional activities of the participants, for example, a more sustained gaze is necessary in activities such as questions and stories, since gaze is viewed as a display of attention and engagement. (Brône et al., 2017) also found that different dialogue acts typically display specific gaze events, from both speakers' and hearers' perspectives.

Unaddressed participants also display interesting gaze behaviour showing that they anticipate turn shifts between primary participants by looking towards the projected next speaker before the completion of the ongoing turn (Holler and Kendrick, 2015). This may be because gaze has a 'floor apportionment' function, where gaze aversion can be observed in a speaker briefly after taking their turn before returning gaze to their pri-

mary recipient closer to turn completion (Kendon, 1967; Brône et al., 2017).

1.2 Referential functions of eye gaze in dialogue

Previous studies have tended to focus on either social functions of gaze (e.g., turn-taking or other interaction management) or how gaze is used in reference resolution, with few researchers combining these.

The process of identifying application-specific entities which are referred to by linguistic expressions is reference resolution. One example is identifying an image on a display by referring to "the painting of a night sky". One area in which multimodal reference resolution has been previously studied is in the context of sentence processing and workload. For example, Sekicki and Staudte (2018) showed that referential gaze cues reduce linguistic cognitive load. Earlier work (e.g., Hanna and Brennan, 2007) showed that gaze acts as an early disambiguator of referring expressions in language.

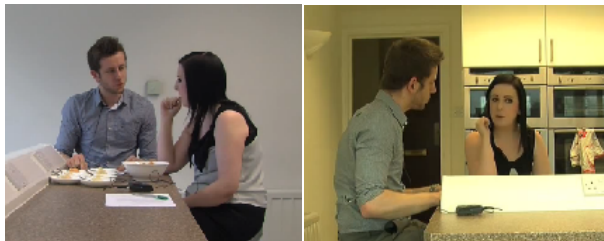
Campana et al. (2002) proposed to combine reference resolution component of a simulated robot with eye tracking information; they intended to deploy this on the International Space Station. Unfortunately, eye movements' integration with speech was not addressed. Also, eye gaze information was used only in case of inability to identify unique referenced objects. Zhang et al. (2004) implemented reference resolution by integrating a probabilistic framework with speech and eye gaze; results showed an increase in performance. They also found that reference resolution of eye gaze could also compensate for lack of domain modelling.

2 Research questions

- Annotation – is it feasible (and can we automate some or all of it using machine learning techniques)?
- Can we classify elements of the dialogue based only on gaze behaviours? (Dialogue acts? Turn-taking? Reference objects?)
- Can we come up with an implementable model of gaze in dialogue for a conversational robot or avatar to interpret human gaze behaviour and produce human-like gaze behaviour?

3 Data

The data used in this pilot come from the case study reported in [Lavia et al. \(2018\)](#). Data consists of videos of pairs of participants (staff at the Good Housekeeping Institute) taste-testing eight different kinds of hummus. Participants are seated at a right-angle to each other, with separate cameras and radio microphones capturing each participant (see figure 1), providing a clear recording of eye movements, facial expressions, gestures and speech.



(a) View from camera 1

(b) View from camera 2

Figure 1: The two camera views

3.1 Annotation

Multimodal video annotation software ELAN will be used for manual analysis. Each of the annotations are entered in tiers and are assigned to a particular time interval. Speech of each participant will be annotated in different tiers as Speech1 and Speech2 (contains transcription of speech and laughter). This will be followed by four additional tiers focusing on the eye gaze. A joint attention tier displays the information of participants looking at a particular object/place at the same time and what exactly they are paying attention to. Mutual gaze tiers records the eye gaze of participant 1 (P1) looking at participant 2 (P2) and vice versa.

The final two tiers are dedicated to random eye gaze information of each participant when they are not involved in Mutual gaze or Joint attention (Random1 and Random2 for participant 1 and participant 2 respectively).

Annotating the speech along with exclusive eye gaze data would help in understanding the dialogue acts elementary to the non verbal yet-obvious interpretations of speech such as referencing, providing subtle cues to organise and control communication, conveying feedback and coordinating turn taking behaviours during speech overlaps. It is also interesting to look into the influence of disagreement in the rating which is persist over the entire conversation influencing fairness and measure how much of this capitulate behaviour is observed through eye gaze. This could help us understand much more about coordinated opinions and gaze switching soon after joint attention.

The task in the video as mentioned earlier is to rate the various hummus. The eye gaze information linked with emotion driven attention could help explore more of the constantly changing opinion of a participant to go along with the partner's stronger perspective. Also, what are the eye movement patterns during such situations and how does it affect the entirety of rating.

4 Discussion

Looking at all the different forms of non verbal communication, eye gaze is very powerful, but even so, we are rarely consciously aware of it. But we are at the verge of breakthroughs in building virtual human avatars, and now, more than ever, it is important to have them behave in more natural ways. Another application example of where this might help is in the area of virtual teleconferencing, by using user gaze information to enhance participant interaction through the conferencing software user interface. As we have discussed above, there is still a need to expand the part of the multimodal dialogue systems literature that focuses on building effective computational models on how people make use of gaze in ordinary conversations.

References

Michael Argyle and Mark Cook. 1976. *Gaze and mutual gaze*. Cambridge University Press.

- Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*, 28(3):449–483.
- Ellen Campana, Beth Ann Hockey, Jason Baldridge, Roger Remington, John Dowding, and Leland S. Stone. 2002. [Using eye movements to determine referents in a spoken dialogue system](#). *Proceedings of the 2001 Workshop on Perceptive User Interfaces*.
- C. Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York.
- Charles Goodwin. 1980. Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological inquiry*, 50(3-4):272–302.
- Joy E. Hanna and Susan E. Brennan. 2007. [Speakers eye gaze disambiguates referring expressions early during face-to-face conversation](#). *Journal of Memory and Language*, 57(4):596 – 615. Language-Vision Interaction.
- Judith Holler and Kobin H Kendrick. 2015. Unaddressed participants gaze in multi-person interaction: optimizing reciprocity. *Frontiers in psychology*, 6(98):1–14.
- Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.
- Lisa Lavia, Harry J. Witchel, Francesco Aletta, Jochen Steffens, André Fiebig, Jian Kang, Christine Howes, and Patrick G. T. Healey. 2018. [Non-participant observation methods for soundscape design and urban planning](#). In Francesco Aletta and Jieliang Xiao, editors, *Handbook of Research on Perception-Driven Approaches to Urban Assessment and Design*. IGI Global.
- Federico Rossano. 2012. *Gaze behavior in face-to-face interaction*. Ph.D. thesis, Radboud University Nijmegen Nijmegen.
- Mirjana Sekicki and Maria Staudte. 2018. [Eye’ll help you out! How the gaze cue reduces the cognitive load required for reference processing](#). *Cognitive Science*, 42(8):2418–2458.
- Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, and Xiaoyang Mao. 2004. [Overriding errors in a speech and gaze multimodal architecture](#). pages 346–348.

