DaP 2018

Proceedings of the Workshop on Dialogue and Perception

Christine Howes, Simon Dobnik and Ellen Breitholtz (eds.)

Gothenburg, 14–15 June 2018

DaP 2018

https://clasp.gu.se/news-events/workshop-on-dialogue-and-perception-2018

CLASP

https://clasp.gu.se/

Preface

This volume contains the papers presented at the CLASP Workshop on Dialogue and Perception 2018, held on June 14-15, 2018 in Gothenburg.

The study of dialogue investigates how natural language is used in interaction between interlocutors and how coordination and successful communication is achieved. Dialogue is multimodal, situated and embodied, with non-linguistic factors such as attention, eye gaze and gesture critical to understanding communication. However, studies on dialogue and computational models such as dialogue systems have often taken for granted that we align our perceptual representations, which are taken to be part of common ground (grounding in dialogue, Clark, 1996). They have also typically remained silent about how we integrate information from different sources and modalities and the different contribution of each of these. These assumptions are unsustainable when we consider interactions between agents with obviously different perceptual capabilities, as is the case in dialogues between humans and artificial agents, such as avatars or robots.

Contrarily, studies of perception have focussed on how an agent interacts with and interprets the information from their perceptual environment. There is significant research on how language is grounded in perception, how words are connected to perceptual representations and agent's actions and therefore assigned meaning (grounding in action and perception, Harnad, 1990). In the last decade there has been impressive progress on integrated computational approaches to language, action, and perception, especially with the introduction of deep learning methods in the field of image descriptions that use end-to-end training from data. However, these have a limited integration to the dynamics of dialogue and often fail to take into account the incremental and context sensitive nature of language and the environment.

The aim of this workshop is to initiate a genuine dialogue between these related areas and to examine different approaches from computational, linguistic and psychological perspectives and how these can inform each other. It will features 8 invited talks by leading researchers in these areas, and 11 peer-reviewed papers (of 15 submissions), presented as posters.

We would like to thank all our contributors and programme committee members, with special thanks to CLASP and Susanna Myyry for the local organisation.

Christine Howes, Simon Dobnik and Ellen Breitholtz

Gothenburg

June 2018

Programme Committee

Ellen Breitholtz	University of Gothenburg
Joyce Chai	Michigan State University
Simon Dobnik	University of Gothenburg
Arash Eshghi	Heriot-Watt University
Kallirroi Georgila	University of Southern California
Mehdi Ghanimifard	University of Gothenburg
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	King's College London
Judith Holler	Max Planck Institute for Psycholinguistics
Christine Howes	University of Gothenburg
John Kelleher	Dublin Institute of Technology
Nikhil Krishnaswamy	Brandeis University
Staffan Larsson	University of Gothenburg
Gregory Mills	University of Groningen, Netherlands
James Pustejovsky	Computer Science Department, Brandeis University
David Schlangen	Bielefeld University
Candy Sidner	Sidner Consulting
Matthew Stone	Rutgers University
Ielka Van Der Sluis	University of Groningen
Diedrich Wolter	University of Bamberg

Table of Contents

Invited Speaker Presentations

Grammars as affordances for interaction: towards an evolutionary tale

Ruth Kempson
Kings College London
ruth.kempson@kcl.ac.uk

Eleni Gregoromichelaki
Heinrich Heine University, Düsseldorf
elenigregor@gmail.com

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

This talk sketches a grammar which replaces an encapsulated concept of language competence with a model in which structure in natural languages (NLs) is an emergent phenomenon based on interactive manipulation of situated perception-action mechanisms. In everyday conversation, utterances and thoughts disperse across interlocutors diffusing individual cognition and leading to establishment of social groupings at various scales (Lerner, 1993). As interlocutors switch speaker-hearer roles even within a single utterance-exchange, a grammar needs to license the sharing of syntactic/semantic dependencies. However, this is a phenomenon posing severe challenges for conventional grammar assumptions. We outline a model (Dynamic Syntax DS: Kempson et al., 2001; Cann et al., 2005; Kempson et al., 2016) in which verbal and non-verbal stimuli are defined as triggers for the operation of conditional probabilistically weighted actions. Under this view, human interaction consists in the provision (*NL generation*) or exploitation (*NL parsing*) of *affordances*, situated action opportunities that create online an ad hoc common processing environment leading to action-coordination among interlocutors. This is achieved by assuming that previous individual experiences with speech and parsing induce the dynamic formation and resolution of anticipatory states (goals). Goals can be achieved either by an individual generating verbal/non-verbal action or by pursuing the affordances offered by an interlocutor or the non-linguistic environment. The immediate effect is the licensing of NL split-actions, including such feedback activities as interruptions/corrections/clarifications, through seamless shifting between speaking (action) and listening (perception). Modelling such data through grammar-internal low-level sensorimotor mechanisms undercuts the need to invoke high-order inference and mindreading to underpin coordinative exchanges.

We then address the significance of DS within a larger cognitive perspective. Noting parallels between DS and the enactive cognition stance, as explored by Clark (2016); Anderson (2014) ao, we argue that DS models a niche sub-system within this overall account. By incorporating the par excellence representational system, NL, within enactive perspectives, competing proposals regarding the status of representations can be seen as different ways of talking about affordances and their emergent and evanescent products. The compatibility of such a view of NL-competence with the embodied view of cognition indicates, first, that NL acquisition can be seen as emerging from a grounding in interaction simpliciter, contra Tomasello (2008) who assumes innate capacity for Gricean-style inference; secondly, analogously, that NL evolution can be seen as having emerged inexorably from the human prior disposition to interact.

References

M.L. Anderson. 2014. *After Phrenology: Neural Re-use and the Interactive Brain*. MIT Press, Cambridge, Mass.

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

A. Clark. 2016. *Surfing Uncertainty: Prediction, Action and the Embodied Mind*. OUP, Oxford.

R. Kempson, W. Meyer-Viol, and D. Gabbay. 2001. *Dynamic Syntax*. Blackwell, Oxford.

Ruth Kempson, Ronnie Cann, Eleni Gregorommichelaki, and Stergios Chatzikyriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics* 42(3,4):203–275.

Gene H. Lerner. 1993. Collectivities in action. *Text-Interdisciplinary Journal for the Study of Discourse* 13(2):213–246.

Michael Tomasello. 2008. *Origins of Human Communication*. MIT.

Towards Real-time Coordination in Spoken Human-Robot Interaction

Gabriel Skantze

KTH Royal Institute of Technology
skantze@kth.se

When humans interact and collaborate with each other, they have to coordinate their behaviours. One of the most fundamental behaviours that needs to be coordinated is the order in which they speak. Since it is difficult to speak and listen at the same time, they need to take turns speaking, and this turn-taking has to be coordinated somehow. To achieve fluent spoken interaction between humans and machines (such as social robots), it is essential that we understand how this coordination is accomplished. Studies on human-human interaction have shown that humans use multi-modal signals, expressed in the face and voice, such as gaze and intonation. Thus, to engage in spoken interaction, social robots should be able to continuously generate and understand these signals. Since social robots are embodied and physically situated, they have a richer repertoire of multi-modal signals, than for example voice assistants in smart speakers. This facilitates more sophisticated coordination, such as multi-party interaction with several users. In multi-party interaction, the coordination of turn-taking becomes more complicated, since the interlocutors not only have to understand when someone yields the floor, but also who is expected to speak next. In such settings, the gaze of the robot and the users becomes an even more important coordination signal.

In this presentation, I give an overview of several studies that we have done to model turn-taking in dialogue. First, I will show how humans in interaction with a human-like robot make use of the same coordination signals typically found in studies on human-human interaction, and that it is possible to use multi-modal sensors and machine learning to automatically detect and combine these cues to facilitate real-time coordination. Second, I will show how a human-like robot face and voice can be used to display turn-taking signals – such as gaze aversion, breathing, facial gestures and hesitation sounds – and that humans react naturally to such signals, without being given any special instructions. By displaying such cues, the robot can for example claim the floor without being interrupted, and it can influence who will be the next speaker. In a multi-party interaction, it means that the robot may regulate the turn-taking to increase the speaking time of non-dominant speakers. Finally, I will present recent work on how Recurrent Neural Networks can be used to train a predictive, continuous model of turn-taking from human-human interaction data. I will show how such a general model can be applied to a number of different tasks, including pause, backchannel and overlap detection, and I will discuss how it could potentially be used to control the verbal and non-verbal signals displayed by the robot.

References

Skantze, G. (2017). Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of SigDial*. Saarbrucken, Germany.

Skantze, G. (2017). Predicting and Regulating Participation Equality in Human-robot Conversations: Effects of Age and Gender. In *Conference on Human-Robot Interaction (HRI2017)*. Vienna, Austria.

Skantze, G. (2016). Real-time Coordination in Human-robot Interaction using Face and Voice. *AI Magazine*, 37(4), 19-31.

Extending Dialogism

Per Linell
University of Gothenburg
Department of Education, Communication and Learning
Box 300, SE-405 30 Göteborg, Sweden
per.linell@gu.se

This lecture will account for the background of a proposal to extend the limits of dialogical theory to encompass new ground (Linell, 2009, 2017).

There are basically two types of theoretical conceptualisations of dialogue, which we, for the sake of simplicity, may call dialogue theories and dialogical theories (or dialogism). Dialogue theories focus entirely on external (observable, social, usually spoken) dialogue, i.e., situated encounters through verbal interaction or other semiotic resources in which two or more human beings make sense together. These encounters may alternatively involve, in addition to a human sense-maker, a "higher animal" or an advanced technical system (e.g., a computational system) that can be assigned agency and some sense-making ability. Dialogical theories are more comprehensive than dialogue theories; they introduce an explanatory dimension by the assumption that human beings possess dialogicality which is an ability of an individual (human being) to develop and practice sense-/meaning-making together with others. The other can be a present other, a peripheral or absent individual or group, or a generalised other (a culture, possessing a language and some other sets of norms.

I will argue that the notion of dialogue must be further extended, by not only transcending the limitation to external dialogue, but also extending "classical dialogism" (i.e., Bakhtin-like theorisations, which primarily involved dialogical text analyses), with its focus on dialogicality (usually focused on social language use), into a broader theoretical framework, which also encompasses indirect interdependences with others in activities that are usually individual on the surface, e.g. in thinking, spontaneous silent sense-making, and perception of the environment. I will briefly discuss three major activity domains, namely, solo thinking (conceived as "inner dialogue"), perception of the environment, and reading. I will summarise the foundation of extended dialogism in 14 points. I will then take up some thoughts on dialogue in relation to perceptual activities.

References

Per Linell. 2009. *Rethinking Language, Mind and World Dialogically: Interactional and contextual theories of human sense-making*. Information Age Publishing, Charlotte, NC.

Per Linell. 2017. Dialogue, dialogicality and interactivity: A conceptually bewildering field? *Language and Dialogue*, 7(3):301–335.

Face-to face conversation with socially intelligent robots

Mary Ellen Foster
School of Computing Science
University of Glasgow
Glasgow G12 8RZ, United Kingdom
MaryEllen.Foster@glasgow.ac.uk

Face-to-face conversation is both the fundamental form of human interaction and the richest possible means of communication. It supports three dimensions that other forms of interaction do not: unrestricted **verbal** expression; full access to all **non-verbal** channels; and instantaneous **collaboration** among the participants (Bavelas et al., 1997). For an artificial communicator such as a robot, the richest and most natural form of interaction is therefore one that mimics face-to-face conversation as closely as possible on all of the above dimensions.

In popular culture and science fiction, the prototypical image of a "robot" is precisely this: an artificial human that is able to engage fully in all aspects of face-to-face conversation. In practice, this sort of *socially intelligent* robot (Dautenhahn, 2007) can be used in any context where the robot must engage in real-world interaction with one or more human partners, where the humans might not necessarily have any special training before encountering the robot.

Developing a robot that is able to participate fully in this sort of natural, face-to-face conversation in the real world presents significant technical challenges: the robot must be able not only to understand the multimodal communicative signals of its human partners, but also to produce understandable, appropriate, and natural social signals in response.

In this talk, I will present three recent projects which aim to develop robots that support this sort of socially intelligent conversation with human partners: the JAMES socially aware robot bartender (http://james-project.eu/), the MuMMER socially intelligent shopping mall robot (http://mummer-project.eu/), and the SoCoRo training robot for adults with autism (http://www.socoro.net/).

References

Janet Beavin Bavelas, Sarah Hutchinson, Christine Kenwood, and Deborah Hunt Matheson. 1997. Using face-to-face dialogue as a standard for other communication systems. *Canadian Journal of Communication*, 22(1).

Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704.

Some Uses of Words, Syntax and Posture to Coordinate Understanding in Dialogue

Pat Healey
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
p.healey@qmul.ac.uk

Joint work with Matthew Purver, Nicola Plant, Christine Howes, and L. Zhang.

A number of well-known theories of dialogue coordination propose that people match each other's verbal and non-verbal choices as a means of underpinning mutual understanding. This talk will present quantitative evidence drawn from a variety of dialogue corpora which that shows that although people do repeat their own verbal and non-verbal behaviours above chance they match each other's behaviours systematically less that would be expected by chance. I argue that where repetitions do occur they are usually used to help modify or alter interpretation - by facilitating various forms juxtaposition or contrast between different peoples contributions.

Reasoning About Decisions / Reasoning About Language

Jacob Andreas
University of California, Berkeley
jda@cs.berkeley.edu

Sequence-to-sequence models for language interpretation and generation have become ubiquitous. The supervised sequence-to-sequence paradigm is powerful: flexible enough to handle many kinds of perceptual and discourse context, and expressive enough to model (some) long-range linguistic structure. But it also has significant limitations: it empirically favors generic utterances over informative ones, and is fundamentally limited to imitating the communicative strategies employed by annotators. This talk will explore first steps towards overcoming these limitations by reasoning explicitly about communicative context.

We'll begin with a family of "neuralized" rational speech acts models that combine learned semantics with inference-driven pragmatics, and see how to apply these models to tasks as diverse as image captioning, instruction generation, and visual navigation. Next, we'll turn to a family of less traditional NLP problems, and look at ways of using the same modeling tools to use language as a scaffold for model interpretability and few-shot concept learning.

Understanding Temporal Descriptions

Laura Carlson and Jennifer Kolesari
University of Notre Dame
Notre Dame
United States
lcarlson@nd.edu

Spatial language has served well as domain for studying the interplay between a speaker's perception of and interaction with an environment and the formulation of dialogue about that environment. Consider, for example, spatial descriptions of the form "The coffee mug is below the coffee pot." Typical models of the apprehension of these descriptions include processes that link language and the underlying spatial representation, including mapping the linguistic elements "mug" and "pot" to their corresponding entities (target and reference object, respectively), assigning a reference frame to define the spatial term "below" and verifying that the description accurately locates the target. My lab has done extensive work examining how context (broadly defined as information about the objects being related, assumptions about the common ground shared between speakers and listeners, and the purpose of the communication) impact these processes. In this talk I will extend this work to the understanding of temporal descriptions of the form "She ran a 5k before she watched the movie." This work capitalizes on the idea that space is foundationally used to understand time, and that the mechanism of a reference frame similarly underlies the mapping of language and perception within the domain of temporal language. We focus on the assignment of a temporal reference frame, and the ways in which context has an impact on the setting of the parameters of such a reference frame.

Mind the Gap: Situated Spatial Language a Case-Study in Connecting Perception and Language

John D. Kelleher
ADAPT Research Centre,
Information, Communication, and Entertainment Institute,
Dublin Institute of Technology,
Ireland
john.d.kelleher@dit.ie

Situated language is spoken from a particular point of view within a simulated or physical context that is shared with an interlocutor. From theoretical linguistic and cognitive perspectives, situated dialog systems are interesting as they provide ideal testbeds for investigating the interaction between language and perception, at the same time there are a growing number of practical applications, for example robotic systems, where spoken interfaces, capable of situated dialog, promise many advantages (Kelleher, 2003) . An open challenge in this domain is the creation of computational models that appropriately ground the semantics of spatial terms within the shared perceptual context. This is partly because of the diversity of factors that impinge on spatial term semantics, including *geometry*, *world knowledge* (including functional roles and object dynamics), and human *perception*.

Many computational models of spatial semantics are based on the concept of a *spatial template* (Logan and Sadler, 1996). This standard model has been extended in a number of ways. For example, to include frame of reference ambiguity (Kelleher and Costello, 2005; Kelleher and van Genabith, 2006; Dobnik et al., 2014); the impact of distractor objects within the scene (Kelleher and Kruijff, 2005; Costello and Kelleher, 2006; Kelleher and Costello, 2009); and to include the role of human attention and visual perceptual factors in spatial reference resolution (Kelleher et al., 2005; Kelleher, 2006; Regier and Carlson, 2001; Kelleher et al., 2010). At the same time, other research has used corpus based analytics to explore the functional and geometric semantics of prepositions in visually situated spatial reference (Dobnik and Kelleher, 2014; Dobnik et al., 2018). However, to-date relatively little work has been focused on developing an integrated model that accommodates all of these factors.

In recent years, however, deep learning approaches have made significant breakthroughs in a number of areas. An exciting aspect of deep learning is the concept of representation learning from data. In particular, learning the projection of naturally discrete information (e.g. words) into continuous representations (e.g. word embeddings), and also learning vector based inter/multi-modal representations, 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 (Kelleher and Dobnik, 2017). However, adopting a modular mechanistic approach to training deep networks may offer a solution to these challenges (Dobnik and Kelleher, 2017).

In light of this, in this paper will review the literature on computational models of spatial semantics and the potential of deep learning models as an useful approach to this challenge.

References

Fintan Costello and John D. Kelleher. 2006. Spatial prepositions in context: The semantics of *Near* in the presense of distractor objects. In *Proceedings of the 3rd ACL-Sigsem Workshop on Prepositions*, pages 1–8.

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 the 2018 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proc. of the Workshop on Vision and Language*, pages 33–37.

Simon Dobnik and John D Kelleher. 2017. Modular mechanistic networks: On bridging mechanistic and phenomenological models with deep neural networks in natural language processing. *CLASP Papers in Computational Linguistics*, page 1.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. *Proceedings of Dial-Watt-Semdial*, pages 43–52.

John D. Kelleher. 2003. *A perceptually based computational framework for the interpretation of spatial language*. Ph.D. thesis, Dublin City University.

John D. Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1):21–35.

John D. Kelleher and Fintan Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-Sigsem Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*.

John D. Kelleher and Fintan Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

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

John D Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. *CLASP Papers in Computational Linguistics*, page 41.

John D. Kelleher and Josef van Genabith. 2006. A computational model of the referential semantics of projective prepositiosn. In *Syntax and Semantics of Prepositions*, pages 211–228. Springer.

John D. Kelleher and Geert-Jan Kruijff. 2005. A context-dependent model of proximity in physically situated environments. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their use in Computational Linguistic Formalisms and Applications*.

John D. Kelleher, Robert J. Ross, Colm Sloan, and Brian Mac Namee. 2010. The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108.

G.D. Logan and D.D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In *Language and Space*, pages 493–530. MIT Press.

T. Regier and L.A. Carlson. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of experimental psychology: General*, 130(2):273–298.

Poster Presentations

Understanding inner representations of perceptual data in grounded multi-agent simulations

Diane Bouchacourt and **Marco Baroni**
Facebook Artificial Intelligence Research, Paris
{dianeb,mbaroni}@fb.com

Abstract

Recent progress in processing raw perceptual data with deep learning techniques has revived interest in studies of emergent communication and interaction in multi-agent settings with realistic visual input. Neural-network agents produce however high-dimensional "inner" representations of the input that can't be directly interpreted. In our ongoing research, we design model-independent probing methods to gain insights into such inner representations, and on how the different tasks that agents are assigned affect their view of the same perceptual information. Our goals are to achieve a better understanding of how tasks affect representations, to make agent behavior more interpretable and, ultimately, to improve interactive agents by providing them with insights into each other's representation.

1 Background and motivation

Simulations in which multiple computational agents learn to communicate in order to solve a task collaboratively or adversarially in a grounded environment have a long history (Lewis, 1969; Briscoe, 2002; Cangelosi and Parisi, 2002; Skyrms, 2019; Steels, 2012, a.o.). However, early work was mostly limited by the technology of its times to simplified, typically hand-crafted small-scale inputs. The amazing success of deep neural networks in processing real-life perceptual data, in particular natural images (Russakovsky et al., 2015), has revitalized the area, with a new generation of studies that consider agents learning to interact when faced with tasks that require processing realistic visual input (e.g., Goodfellow et al., 2014; Lazaridou et al., 2017; Das et al., 2017; Lazaridou et al., 2018; Lee et al., 2017). In these multi-agent setups, agents typically process similar perceptual input, that they must use for different purposes. For example, in the classic referential game of Lazaridou et al. (2017), two players, a Sender and a Receiver, see the same pairs of images, one of them being the "target" image. The Sender sees the images and knows the target, and chooses to send one symbol from a fixed vocabulary to the Receiver. The Receiver sees only the images, receives the sent symbol, and tries to guess which image is the target. In case of success, both players receive a payoff of 1. Since the involved agents are neural networks, during training they produce distributed representations of the inputs, tuned to the task they are learning to solve. However, given the noisy natural input and the fact that these distributed representations are dense and high-dimensional, it is hard to understand what they are encoding. Our current research aims at developing methods to analyze the specifics of the representations of the same perceptual data developed by agents with different tasks. We have three main reasons to focus on this. First, from a cognitive point of view, understanding how different tasks affect perceptual representations might provide insights into how higher-level perception gets tuned to communicative or antagonistic objectives. For example, does categorical perception (Goldstone and Hendrickson, 2010) naturally arise in setups where agents must efficiently discriminate broad natural kinds?

Second, as machine learning is tasked with increasingly important functions, being able to understand the representations that algorithms develop of the data they are exposed to is of the utmost importance (Doshi-Velez and Kim, 2017). This is particularly true in multi-agent settings, where the agents might develop their own opaque communication means (Lewis et al., 2017).

Third, a core feature of human intelligence is a "theory" of the other agent's mind (Wimmer and Perner, 1983). Both older and recent work in multi-agent communication has acknowledged the importance of this (Batali, 1998; Lee et al., 2017; Choi et al., 2018; Rabinowitz et al., 2018). Understanding the differences between agent representations is a first step towards designing agents that track each other representations, with the aim of developing better multi-agent models.

2 General probing methods

Given a generic setup with two agents operating in an environment in which they must process some input perceptual data, we consider three representations: the shared input representation of the data,[1] and the representations of the data produced by the two agents (A1 and A2).[2] We don't assume they live in the same space: For example, inputs might be full pixel maps, but the agents might produce lower-dimensionality representations, that in turn might be of different sizes for A1 and A2. We are interested in comparing each agent's representation with the input, as well as with each other. We want to develop probing methods that do not require *ad-hoc* manual annotation of data and, to the extent that this is possible, are independent of the specific data sets and tasks that agents are assigned. Finally, we expect that typically the input data can be split into *familiar* and *generalization* sets. The first subset includes data that were used to train the agents, and thus influenced their representation learning strategies, whereas the generalization subset contains novel data, which might be used to probe how the agents behave out of their training domain.

Our first and most fundamental probing method consists in *representational similarity analysis* (RSA) (Kriegeskorte et al., 2008). Given two sets r_1 and r_2 of representations of the same collection of items (e.g., r_1 is the collection of input images as represented by A1 and r_2 is the same collection represented by A2), we first compute s_1 as all possible pairwise similarities between the representations in r_1, and s_2 as those in r_2. We then compute the correlation between the similarity vectors s_1 and s_2. This latter value, which we will call *RSA score*, measures the global agreement between s_1 and s_2, relative to the chosen input collection. If N is the number of items in the collection that we compute representations for, both similarity vectors s_1 and s_2 are of length $N(N-1)$. Therefore, it is not necessary for the representations r_1 and r_2 to belong to the same space: As long as we have a similarity function for the relevant items in each space, we can compute a RSA score between similarity vectors.

Equipped with RSA and a collection of items, we can compare the similarity vectors of A1 and A2's representations to the input representation similarity vector, as well at with each other. This way we address questions such as: Does solving a task have an impact on lower-level perceptual representations? Are the tasks of A1 and A2 such that they lead to divergent representations? On which item pairs in particular do A1 and A2 differ most?

Next, we can focus on the underlying collection, studying how RSA scores change depending on whether we sample from the familiar or generalization sets of the agents. For example, we might find in this way that the agents' representations differ from the input ones more for items in the familiar set, as they might have developed *ad hoc* representation strategies that only affect items that were relevant for their training task.

Another set of probing methods aim at understanding to what extent agent representations are sensitive to the natural well-formedness of the input signal. Assuming we are in the image domain, we automatically create training, validation and test sets containing intact pictures and pictures that have been systematically tampered with. We produce input and agent representations of the pictures, and we train a binary classifier on these representations, to distinguish intact from edited pictures ("diagnostic classifiers" of this sort were recently used, especially in NLP, to probe embeddings, see, e.g., Adi et al., 2017; Hupkes et al., 2017). Depending on how specialized the representations developed by the agents are, they might or not still preserve information allowing them to distinguish normal and anomalous images. We envision for now three automated ways to alter images: blank out a fixed proportion of random pixels; cut out and shift two random square boxes; rotate the image. As with the RSA methods, the anomaly classifiers can be trained and tested on collections derived from the familiar or generalization sets, to further probe how *ad-hoc* the

[1] For example, pixel maps or pre-trained-ConvNet-generated embeddings corresponding to input images.

[2] Our probing methods can straightforwardly be extended to multiple agents, and multiple level of representations.

learned representations are.

The probing methods we listed here are not exhaustive, but they have the desired properties: They do not assume that agent representations belong to the same space, they make no assumptions about agents' architectures, they do not require manual annotation data, and they should allow us to answer general questions about what the agents are actually "seeing".

3 Experimental plan

To begin with, we look at classic collaborative and adversarial setups. For the former case, we re-implement the classic referential game of Lazaridou et al. (2017) described in Section 1, due to its simplicity. We consider two settings. First, we train the Sender and Receiver agents as in Lazaridou et al. (2017). Second, we explore a setting in which, after training successfully converges, the parameters of one of the agents (either Sender or Receiver) are frozen. The other agent is re-initialized and re-trained to learn to communicate with the frozen, trained interlocutor. Intuitively, the first setting can be seen as two children acquiring language by interacting with each other, while the second is the case of a parent teaching a child how to talk.

For the adversarial setup, we consider the popular generative adversarial network architecture (GAN, Goodfellow et al., 2014). After training a Generator and Discriminator on image generation, we produce fake images with the Generator and forward-pass them to the Discriminator. We compare the representations of the fake images in the Generator, in the Discriminator, and the their original pixel maps.

In both setups, we apply the probing methods described in Section 2 above. To obtain similarity vectors (one for each agent representation and one for the input's representation), we use the cosine measure to compute pairwise similarities between items. We then employ, as RSA score, the Spearman correlation between the different similarity vectors. For the anomaly classifiers, we use logistic regression, as we are interested in whether the relevant information is easy to retrieve from the representations with a simple linear readout (Fusi et al., 2016).

4 Preliminary results

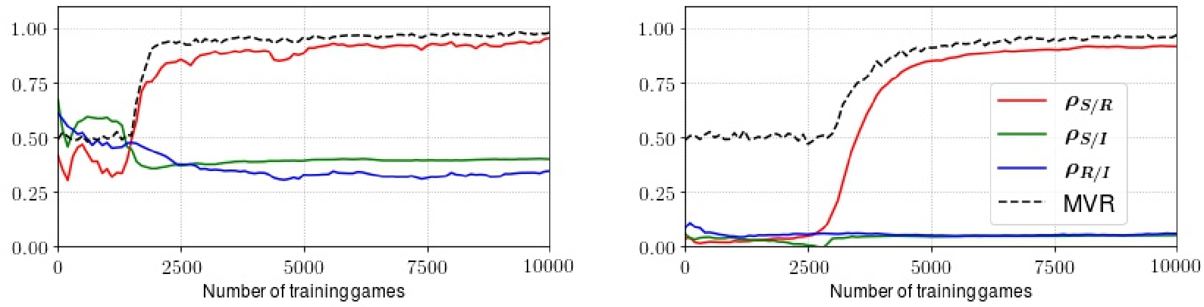
We perform preliminary experiments in the referential game setup described in Section 1. We re-implement Lazaridou's Sender and Receiver architectures (using their better-behaved "informed" Sender). Following Lazaridou et al. (2017), for each of the 463 concepts they used, we randomly sample 100 images from ImageNet (Deng et al., 2009). We construct $50,000$ mini-batches of 32 image pairs during training and $1,024$ pairs for validation. We construct a held-out test set in the same way by sampling 10 images per concept from ImageNet (for 2 concepts, we were not able to assemble enough further images), for a total of $4,610$. We compute RSA scores on the cross-product of test images. The images are passed through a pre-trained VGG ConvNet (Simonyan and Zisserman, 2015), and the input vector fed to the agents is either (i) the top 1000-D softmax layer (**sm**) or (ii) the second-to-last 4096-D fully connected layer (**fc**). We repeat all experiments using 100 random initialization seeds. As we faithfully reproduced the setup of Lazaridou et al. (2017), we refer the reader there for hyper-parameters and training details. Table 1 reports the Spearman (ρ) correlation values between the two agents' representations and between each agent and the input (S refers to Sender, R to Receiver, and I to the input), using either **fc** or **sm** as input data for training and testing.

Using **fc**	Using **sm**
$\rho_{S/R} = 0.98/0.45$	$\rho_{S/R} = 0.97/0.03$
$\rho_{S/I} = 0.33/0.66$	$\rho_{S/I} = 0.04/0.05$
$\rho_{R/I} = 0.33/0.67$	$\rho_{R/I} = 0.04/0.05$

Table 1: Spearman coefficient values. Blue: communication success, black: at initialization.

We average values on initialization seeds for which the agents successfully communicate at the end of training (in blue)[3] and before any training is done (in black). We see that the similarity vectors of the agents' representations are strongly correlated. Note that, averaged on the 4 seeds which resulted in communication failure (all with **fc**), this value drops at $\rho_{S/R} = 0.39$, which is smaller than the initialization value. This suggests that

[3]We consider training successful if the Mean Validation Reward (MVR) is MVR $\geq$ 80%, this gives 96 seeds for **fc** and all 100 seeds for **sm**.



(a) Using **fc**.

(b) Using **sm**.

Figure 1: Spearman correlation coefficients and MVR during the first $10,000$ training games.

agent representations drifted away from initialization in different ways. Interestingly, we observe that the agents are more correlated with the input representation when using **fc**. Figure 1 shows the Spearman coefficient and mean validation reward (MVR) development curves using either type of input, in each case for its cross-validated best seed. We see that the MVR and $\rho_{S/R}$ values increase concurrently. Looking at Figure 1a, we note that, during the first few hundred games, the Sender (green curve) aligns with the input, but the Receiver (blue curve) does not. Therefore, it seems that, in order to establish communication, the Sender has to drift away from the input and align with the Receiver.

5 Related work

There is of course extensive work on visualizing and understanding the behavior of neural network, for example in vision (e.g., Simonyan et al., 2013; Zeiler and Fergus, 2014; Mordvintsev et al., 2015) and NLP(e.g., Hupkes et al., 2017; Linzen et al., 2016; Kàdàr et al., 2017). In the context of multi-task learning, Long et al. (2017) learn covariance matrices for networks trained on a set of tasks and show Hinton diagrams of the tasks covariances. Unlike these works, we are interested specifically in methods to address the questions of how different tasks lead to different representations of the same perceptual data. Also, we want model-agnostic methods, that can generically be applied to any agent developing an inner representation of the data. In this second respect, our study is closest to recent work on model-agnostic probing tasks in NLP (Adi et al., 2017; Shi et al., 2016), which, however, did not address the issue of task-specific representations of the same data. Naturally, multi-agent studies often perform some qualitative analysis of what the agents learned, and how they differ. Closest to us, Lazaridou et al. (2018) performed RSA on all their agent ConvNet layers, finding an increase in global similarity at higher layers, and, like us, dramatic drops when communication fails.

Acknowledgments

We thank Angeliki Lazaridou, Douwe Kiela and Calvin Lee for their useful discussions and insights. We also thank Francisco Massa for his help on setting up the experiments.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*, Toulon, France. Published online: `https://openreview.net/group?id=ICLR.cc/2017/conference`.

John Batali. 1998. Computational simulations of the emergence of grammar. In James Hurford, Michael Studdert-Kennedy, and Chris Knight, editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, pages 405–426. Cambridge University Press, Cambridge, UK.

Ted Briscoe, editor. 2002. *Linguistic evolution through language acquisition*. Cambridge University Press, Cambridge, UK.

Angelo Cangelosi and Domenico Parisi, editors. 2002. *Simulating the evolution of language*. Springer, New York.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. 2018. Compositional obverter communication learning from raw visual input. In *Proceedings of ICLR Conference Track*, Vancouver, Canada. Published online: `https://openreview.net/group?id=ICLR.cc/2018/Conference`.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of CVPR*, pages 326–335, Honolulu, HI.

Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami Beach, FL.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. `https://arxiv.org/abs/1702.08608`.

Stefano Fusi, Earl Miller, and Mattia Rigotti. 2016. Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74.

Robert Goldstone and Andrew Hendrickson. 2010. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of NIPS*, pages 2672–2680, Montreal, Canada.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2017. Visualisation and diagnostic classifiers reveal how recurrent and recursive neural networks process hierarchical structure. `http://arxiv.org/abs/1711.10203`.

Àkos Kàdàr, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4):1–28.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *Proceedings of ICLR Conference Track*, Vancouver, Canada. Published online: `https://openreview.net/group?id=ICLR.cc/2018/Conference`.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *Proceedings of ICLR Conference Track*, Toulon, France. Published online: `https://openreview.net/group?id=ICLR.cc/2017/conference`.

Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2017. Answerer in questioner's mind for goal-oriented visual dialogue. `https://arxiv.org/abs/1802.03881`.

David Lewis. 1969. *Convention*. Harvard University Press, Cambridge, MA.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of EMNLP*, pages 2443–2453, Copenhagen, Denmark.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S Yu. 2017. Learning multiple tasks with multilinear relationship networks. In *Proceedings of NIPS*, pages 1594–1603, Long Beach, CA, USA.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. 2015. DeepDream - a code example for visualizing Neural Networks. `https://research.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html`.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. `https://arxiv.org/abs/1802.07740`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition challenge. *International Journal of Computer Vision*. In press.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of EMNLP*, pages 1526–1534, Austin, Texas.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. `http://arxiv.org/abs/1312.6034`.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR Conference Track*, San Diego, CA. Published online: `http://www.iclr.cc/doku.php?id=iclr2015:main`.

Brian Skyrms. 2019. *Signals: Evolution, learning, and information*. Oxford University Press, Oxford, UK.

Luc Steels, editor. 2012. *Experiments in Cultural Language Evolution*. John Benjamins, Amsterdam.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Matthew Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of ECCV (Part 1)*, pages 818–833, Zurich, Switzerland.

Relating coordination in non-linguistic games and dialogue games

Ellen Breitholtz **Robin Cooper**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{ellen.breitholtz,cooper}@ling.gu.se

Abstract

We use TTR, a type theory with records, to characterize a non-linguistic game involving perception and coordination and then suggest that this same notion of game can be applied to conversational games (including conversational games which are multimodal). We will show that this notion of game has a natural connection to topoi as discussed by Breitholtz, and is based on the same kind of common sense reasoning. However, it has nothing to say about how to make choices between alternative moves in non-deterministic games. For this we suggest blending the TTR notion of game with standard Game Theory, taking inspiration from recent work by Heather Burnett.

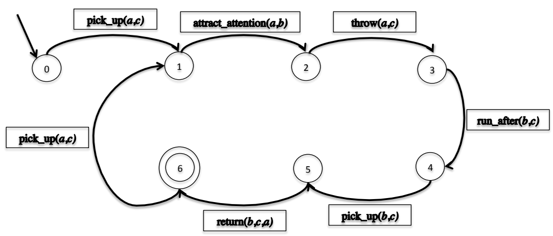
It is a central idea in TTR (Cooper, 2012; Cooper and Ginzburg, 2015) that perception involves classifying an object or situation as being of a type, or in type theoretic terms making a judgement that an object or situation, a is of some type T, $a : T$. The notion of game introduced by Cooper (2014); Breitholtz (2014a); Cooper (prep) builds on this idea. Games in the sense we will discuss also relate to the notion of genre or conversation types discussed by Ginzburg (2012) and communicative projects discussed by Linell (2009). This follows a long tradition of making a connection between linguistic and non-linguistic action, going back at least to Austin (1962); Lewis (1969); Searle (1969); Clark (1996); Barwise and Perry (1983). This work continues to be influential in a great deal of linguistic research including SDRT, dynamic syntax, applications of game theory to linguistic analysis, the philosophy of language and computational work on dialogue systems.

Another central idea in the TTR characterization of games is the idea that situations or events can be seen as strings of smaller situations or events. This is an adaptation to TTR of the finite state approach to events developed over many years by Tim Fernando. A recent account of Fernando's work can be found in Fernando (2015). In (1) we give an example of how we can characterize string types in TTR:

(1) a. if $T_1, T_2 \in$ **Type**, then $T_1 {}^\frown T_2 \in$ **Type**
$a : T_1 {}^\frown T_2$ iff $a = x^\frown y$, $x : T_1$ and $y : T_2$
b. if $T \in$ **Type** then $T^+ \in$ **Type**.
$a : T^+$ iff $a = x_1^\frown \ldots ^\frown x_n$, $n > 0$ and for i, $1 \leq i \leq n$, $x_i : T$

Consider the non-linguistic game of "fetch", played by a human and a dog, where the human throws a stick which the dog is then supposed to run after and bring back to the human. We can think of a simple version of this as a finite-state machine as given in (2).

(2)



Here notations like 'pick_up(a,c)' represent types of situations or events, in this case, the type of situation where a (the human) picks up c (the stick). We can think of the automaton as representing a type of events. Given the string types we introduced in 1 the type represented can be expressed as in (3).

(3) (pick_up(a,c)$^\frown$attract_attention(a,b)$^\frown$
throw(a,c)$^\frown$run_after(b,c)$^\frown$
pick_up(b,c)$^\frown$return(b,c,a))$^+$

In order to explain how two agents (a human and a dog) could coordinate on the production of an event of this type we use the notion of gameboard (Lewis, 1979; Ginzburg, 1994, 2012) or information state (Larsson, 2002) which enables the agents to keep track of where they are in the process of creating the event. Each agent has their own view of the state of the game and this plays a central role in coordination, especially when the two views of the state of the game are not in sync and repair must be carried out. We shall model information states as records in the TTR sense, that is, sets of fields consisting of labels and objects and gameboards as record types which are types of information states. For this relatively simple, non-linguistic game we shall characterize the type of information states as requiring one field for an agenda, as in (4).

(4) $\left[\ \text{agenda} \ : \ \text{list}(\textit{RecType})\ \right]$

This means that any information state (a record) of this type must contain a field labelled by 'agenda' whose value is a list of record types, representing the types that the agent plans to realize (in order) at the current stage of the game. We say that an initial information state is one where the agenda is the empty list, that it, it is of the type in (5).

(5) $\left[\ \text{agenda=[]} \ : \ \text{list}(\textit{RecType})\ \right]$

The *manifest field* in (5) expresses that an information state of this type must not only contain a field with the label 'agenda' whose value will be a list of record types but that in addition it determines which list of record types it will be, namely the empty list.

Now we can think of a game as a set of update functions corresponding the transitions in the finite state automaton (2). In (6) we give three examples of such functions.

(6) a. **initial update function**

$\lambda r{:}\left[\text{agenda=[]:list}(\textit{RecType})\right]\ .$
$\left[\text{agenda=[}\left[\text{e:pick\_up}(a,c)\right]\text{]:list}(\textit{RecType})\right]$

b. **non-initial, non-final update function**

$\lambda r{:}\left[\text{agenda=[}\left[\text{e:pick\_up}(a,c)\right]\text{]:list}(\textit{RecType})\right]$
$\lambda e{:}\left[\text{e:pick\_up}(a,c)\right]\ .$
$\left[\text{agenda=[}\left[\text{e:attract\_attention}(a,b)\right]\text{]:list}(\textit{RecType})\right]$

c. **final update function**

$\lambda r{:}\left[\text{agenda=[}\left[\text{e:return}(b,c,a)\right]\text{]:[}\textit{RecType}\text{]}\right]$
$\lambda e{:}\left[\text{e:return}(b,c,a)\right]\ .$
$\left[\text{agenda=[]:list}(\textit{RecType})\right]$

(6a) says that if the agenda is empty then the type of event where a (the human) picks up c (the stick) can be put on the agenda. (6b) says that if this type is on the agenda and there is in fact an event, e of that type, then the type of event where a attracts b's (the dog's) attention can go on the agenda. (6c) is a final update function which returns an empty agenda after the dog has returned the stick. An empty agenda means that the agent exits the game successfully.

We have formulated this game as a game between three particular individuals a, b and c. A version where we have abstracted over the roles is given in Figure 1. This maps a situation (modelled as a record) where there is a human, a dog and a stick to a set of update functions involving those participants.

Consider a dog, d, who perceives a human picking up a stick and attracts d's attention with it, that is, d has perceived an event of the type in (7).

(7)
$$\begin{bmatrix} \text{x:}\textit{Ind} \\ \text{c}_{\text{human}}\text{:human(x)} \\ \text{y:}\textit{Ind} \\ \text{c}_{\text{dog}}\text{:dog(y)} \\ \text{z:}\textit{Ind} \\ \text{c}_{\text{stick}}\text{:stick(z)} \\ \text{e:}\left[\text{e:pick\_up(x,z)}\right]^\frown\left[\text{e:attract\_attention(x,y)}\right] \end{bmatrix}$$

This is enough information for d to come to the conclusion that she is involved in a type of event where she is playing fetch with the human. In fact, at this point, many dogs will start running in the direction in which the human appears to be about to throw the stick.

(8)
$$\lambda r{:}\begin{bmatrix} \text{x:}\textit{Ind} \\ \text{c}_{\text{human}}\text{:human(x)} \\ \text{y:}\textit{Ind} \\ \text{c}_{\text{dog}}\text{:dog(y)} \\ \text{z:}\textit{Ind} \\ \text{c}_{\text{stick}}\text{:stick(z)} \\ \text{e:}\left[\text{e:pick\_up(x,z)}\right]^\frown\left[\text{e:attract\_attention(x,y,)}\right] \end{bmatrix}.$$
$$\left[\text{e:play\_fetch}(r\text{.x},r\text{.y},r\text{.z})\right]$$

Given a situation in which there is a human, a dog and a stick such that the human picks up the stick

$$
\begin{array}{l}
\lambda r^*{:}\left[\begin{array}{lll}
\mathrm{h} & : & \mathit{Ind} \\
\mathrm{c_{human}} & : & \mathrm{human(h)} \\
\mathrm{d} & : & \mathit{Ind} \\
\mathrm{c_{dog}} & : & \mathrm{dog(d)} \\
\mathrm{s} & : & \mathit{Ind} \\
\mathrm{c_{stick}} & : & \mathrm{stick(s)}
\end{array}\right] . \\
\{\ \ \lambda r{:}\left[\text{agenda=[]:}[\mathit{RecType}]\right] . \\
\qquad \left[\text{agenda=}[\left[\text{e:pick\_up}(r^*.\text{h},r^*.\text{s})\right]]\text{:}[\mathit{RecType}]\right], \\
\quad \lambda r{:}\left[\text{agenda=}[\left[\text{e:pick\_up}(r^*.\text{h},r^*.\text{s})\right]]\text{:}[\mathit{RecType}]\right] \\
\qquad \lambda e{:}\left[\text{e:pick\_up}(r^*.\text{h},r^*.\text{s})\right] . \\
\qquad\quad \left[\text{agenda=}[\left[\text{e:attract\_attention}(r^*.\text{h},r^*.\text{d})\right]]\text{:}[\mathit{RecType}]\right], \\
\quad \ldots, \\
\qquad \lambda e{:}\left[\text{e:return}(r^*.\text{d},r^*.\text{s},r^*.\text{h})\right] . \\
\qquad\quad \left[\text{agenda=[]:}[\mathit{RecType}]\right] \\
\hfill \}
\end{array}
$$

Figure 1: Game of fetch with roles abstracted

and attracts the attention of the dog, this function returns the type of situations where the human and the dog play fetch with the stick. Note that it is important that it returns the type, not a particular situation of the type. The situation does not yet exist. The type indicates what kind of situation might be realized given the initial part the has been perceived. The dog can use this to guide its future actions in collaborating with the human to realize the type.

Note that characterizing a game in the way we have does not actually explain how anything actually happens. The update functions when given appropriate arguments will return a type. What an agent does with that type needs to be specified in a superordinate theory of action of the kind discussed in Cooper (2014, prep). The type theory as such enables us to provide a rich theory of the kind of objects that can be manipulated by the actions.

The function in (8) is exactly the kind of function employed by Breitholtz (2014a) to model enthymemes and topoi. Originally a rhetorical concept, an enthymeme is an argument where one or more premises necessary for the argument to be logical are suppressed. A topos in the same tradition, refers to an implicit inference rule that can be drawn on to underpin enthymematic arguments. Like topoi, the game modelled in the function above can be drawn on to underpin an act of reasoning. However, an observation by an agent of a move of the type where someone holds up a stick does not logically or necessarily entail that there will be a game of fetch. Rather the relation between the different types of moves or events in a conversational game (or a topos), is associative. Thus, in the example in (8) it could be that the human had a different intention or it might be that the human intended for there to be a game of fetch but just at this point fell and sprained her ankle, thus forcing the abandonment of the game. Nevertheless, despite the unreliability of the inference, it is an example of the kind of inference which agents live by in order to be able to interact with the world and with other agents. Breitholtz (2014b) gives a number of examples of how this kind of reasoning plays a role in dialogical interactions. On this account a conversational game is a strategy available to an interlocutor engaged in a particular activity who is to carry out a particular communicative project. Different games may be employed to perform the same kind of project – for example establishing which joint action to take in a given situation. One way of carrying out this project is by playing the suggestion game, as seen in 2. The suggestion game is of a type where the first move is made by one of the dialogue participants (who thereafter assumes the role of player 1 in the game) to the other. After the suggestion move follows an optional move by player one to motivate the suggestion, followed by an accept- or a reject move by player 2.

When engaging in dialogue, the participants of a conversational game have at their disposal sets of topoi – some of which are general, some associated with the activity or game – which can be drawn on to produce and interpret dialogue moves.

$$T_{G_S}\!:\!\left[\text{e:suggest(player1, player2)}\right]\frown\left[\text{e:motivate(player1, player2)}\right]$$
$$\leqslant^1\frown(\left[\text{e:accept(player2, player1)}\right]\vee\left[\text{e:reject(player2, player1)}\right])$$
$$\}$$

Figure 2: Suggestion game

There are, however, two important things missing in the current proposals for analyzing this kind of reasoning in TTR. One is that there is no indication of the *perceived degree of reliability* of the inference. The other is that there is no mechanism for dealing with *choices of action* in a non-deterministic game. We are currently exploring how a synthesis of the TTR approach to games with a more standard game theory (GT) as employed, for example, by Burnett (fthc) for the analysis of social meaning, could fill this gap and also place GT within a general theory of dialogue.

We illustrate this with a scenario where two agents, A and B, are trying to agree on what to do in a particular situation. This could be done by means of various conversational games, and which one is chosen depends on several factors. Assume that A tells B "We are doing P!". In ordering B, A limits B's choices if B wants to accept her role in an ordering game. On the other hand, choosing this strategy might decrease the likelihood that B will keep playing the game. If A chooses a strategy where he leaves B the possibility of rejecting the suggestion, B is more likely to accept the role assigned to her. If A also adds a reason for doing P, the chances of success in actually getting B to agree increases, as long as the reason chosen can be identified by B as drawing on a topos which B accepts and ranks as important.

References

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press.

Breitholtz, E. (2014a). *Enthymemes in Dialogue: A micro-rhetorical approach*. PhD thesis, University of Gothenburg.

Breitholtz, E. (2014b). 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.

Burnett, H. (fthc). Signalling Games, Sociolinguistic Variation and the Construction of Style. *Linguistics and Philosophy*.

Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge.

Cooper, R. (2012). Type theory and semantics in flux. In Kempson, R., Asher, N., and Fernando, T., editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier BV. General editors: Dov M Gabbay, Paul Thagard and John Woods.

Cooper, R. (2014). How to do things with types. In de Paiva, V., Neuper, W., Quaresma, P., Retoré, C., Moss, L. S., and Saludes, J., editors, *Joint Proceedings of the Second Workshop on Natural Language and Computer Science (NLCS 2014) & 1st International Workshop on Natural Language Services for Reasoners (NLSR 2014) July 17-18, 2014 Vienna, Austria*, pages 149–158. Center for Informatics and Systems of the University of Coimbra.

Cooper, R. (in prep). Type theory and language: from perception to linguistic communication. Draft of book chapters available from `https://sites.google.com/site/typetheorywithrecords/drafts`.

Cooper, R. and Ginzburg, J. (2015). Type theory with records for natural language semantics. *Handbook of Contemporary Semantic Theory, The*, pages 375–407.

Fernando, T. (2015). The Semantics of Tense and Aspect: A Finite-State Perspective. In Lappin, S. and Fox, C., editors, *The Handbook of Contemporary Semantic Theory*. Wiley-Blackwell, second edition.

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

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Larsson, S. (2002). *Issue Based Dialogue Management*. PhD thesis, University of Gothenburg.

Lewis, D. (1969). *Convention*. Harvard University Press.

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

Linell, P. (2009). *Rethinking Language, Mind, and World Dialogically: Interactional and contextual theories of human sense-making*. Advances in Cultural Psychology: Constructing Human Development. Information Age Publishing, INC.

Searle, J. R. (1969). *Speech Acts: an Essay in the Philosophy of Language*. Cambridge University Press.

Perceiving understanding through unimodal and multimodal micro-feedback in intercultural dialogue

Anna Jia Gander and Pierre Gander

jia.lu@gu.se, pierre.gander@gu.se
Dept. of Applied IT, University of Gothenburg
Forskningsgången 6, SE412 96 Göteborg, Sweden

Abstract

Perception of understanding is studied in eight video recorded spontaneous face-to-face dyadic first encounters conversations between Chinese and Swedish participants. Sufficient understanding, misunderstanding, and non-understanding are investigated from an analyst's perspective with a focus on unimodal and multimodal micro-feedback. Results are that micro-feedback predominantly shows sufficient understanding. Unimodal head movements exclusively show sufficient understanding. Misunderstanding and non-understanding are more revealed through multimodal micro-feedback expressions than unimodal ones. In relation to sufficient understanding, the most commonly used micro-feedback expressions are *yeah*, *okay*, *m*, *yeah* + nod(s), and chuckle. Regarding misunderstanding, half of the employed multimodal expressions contain nod in combination with *yeah* or a noun phrase associated with hesitation. For non-understanding, unimodal micro-feedback *sorry*, *what do you mean*, eyebrow raise, and gaze at and multimodal micro-feedback head forward or eyebrow raise combined with *sorry*, *what*, or *huh* are most frequently used, expressing uncertainty and eliciting further information.

1 Introduction

Understanding is central to communication. However, understanding in communication is a complex process and is not easy to achieve, due to various reasons, for example, limitations of common knowledge and resources in sense-making (see Linell, 2009; Zlatev, 2009). Many earlier studies of understanding in conversation have focused on verbal rather than bodily behaviors (e.g., Zaefferer, 1977; Bazzanella and Damiano, 1999; Weigand, 1999; Dascal, 1999; Danieli and Bazzanella, 2002; Verdonik, 2010; Kushida, 2011; Lynch, 2011), and thus there is a need to study both.

The notion of micro-feedback refers to unobtrusive expressions used in ongoing conversation such as nods, *uh huh*, and *yeah*, which is one main type of evidence of showing understanding. The relation between micro-feedback and understanding has received little attention, especially through systematic studies using empirical conversational data. In the present study, we analyse how understanding is communicated through micro-feedback in first acquaintance meetings between Chinese and Swedes. The cultural difference and interpersonal unfamiliarity likely result in more understanding problems (Gumperz, 1982; Tannen, 1990; Allwood, 2015; Linell, 2009) and thus more opportunities to elicit and give micro-feedback (Svennevig, 1999; Maynard and Zimmerman, 1984), which is in particular interesting for this study. Two research questions are investigated. First, how are the auditory and visual modalities involved in micro-feedback expressions that are related to sufficient understanding, misunderstanding, and non-understanding? Second, what are these typical unimodal and multimodal micro-feedback expressions?

2 Background

Micro-feedback items have certain communicative functions (Nivre et al., 1992) such as *I hear and understand what you have just said* (cf. Clark and Schaefer's (1989) *acknowledgement expressions* and Yngve's (1970) *backchannel*). These micro-feedback items respond to earlier conversational contributions and also provoke further responses (Bakhtin, 1986; Goodwin, 1981; Schegloff, 1996; Linell, 2009). In addition, the concept of micro-feedback in this study also has the following features: having no independent referential or semantic meaning but being very much dependent on the communication contexts, occurring at the beginning of a responsive communication contribution which includes utterances and gestural behaviors, functioning as a connector between the adjacent communication contributions, and sometimes expressing positive and negative evaluative opinions, for example, agreement and disagreement. Vocal-verbal and gestural micro-feedback expressions are distinguished in terms of the sensory modality. Micro-feedback can be unimodal, occurring in a single modality; or, it can be multimodal, with more than one modality involved simultaneously.

A framework of classifying understanding based on Allwood (1986), Clark and Schaefer (1989), Weigand (1999), and Linell (2009) is used in this study. It includes sufficient understanding, misunderstanding, and non-understanding. Sufficient understanding refers to the understanding which is sufficient to serve the current practical purposes (Garfinkel, 1967) of continuing communication, information sharing, and sense-making, no matter if the understanding is full or partial (see Linell, 2009). The interlocutors are content with the understanding of one another and it is well enough to proceed further (see Lindwall and Lymer, 2011). Misunderstanding is defined as one type of insufficient understanding in that although it can serve the current communication purposes, it occurs when the information is understood in an incorrect way that is deviated from the intention and anticipation. Non-understanding is also identified as one type of insufficient understanding. Non-understanding occurs when the information is not understood at all for reasons such as lack of access to the information or the background knowledge. It cannot serve the current communication purposes of making sense of the presented information. The present study focuses on what micro-feedback occurs in relation to these three types of understandings.

3 Method and data

The study is based on eight video recorded face-to-face dyadic dialogues between four Swedish and four Chinese participants who had no prior acquaintance. Their task was to get acquainted with one another. The communication language was English lingua franca. The data last 65:08 minutes and consist of 10,127 vocal words.

The data were transcribed according to the Göteborg Transcription Standard version 6.2 (Nivre et al., 2004). Understanding was coded as sufficient understanding, misunderstanding, and non-understanding from the analyst's perspective by using an interactional approach. A variant of the MUMIN (Multimodal Interface) coding scheme for feedback (Allwood et al., 2007) was used. That is, the gestural micro-feedback consists of head movements (nod, up-nod, shake, and tilt), facial expressions (smile, laughter, eyebrow movements, gaze movements, and mouth movements), hand movements, and posture movements. Inter- and intra-coder reliability checking was done between six Chinese and Swedish transcribers and annotators with an average agreement rate of 93% on the coding of micro-feedback and a Cohen's kappa of 0.69 on the coding of understanding.

4 Results

The results show that the frequencies of unimodal and multimodal micro-feedback expressions are similar (684 and 604, respectively) (raw frequencies are given within parentheses). The occurrences of unimodal vocal-verbal (341) and gestural (343) micro-feedback expressions are roughly the same. Micro-feedback associated with sufficient understanding is substantially more frequent (1256) than that associated with misunderstanding (9) and non-understanding (23).

4.1 Sufficient understanding

Sufficient understanding is more frequently shown by unimodal micro-feedback (677) than multimodal (579). The three most common unimodal vocal-verbal micro-feedback expressions are *yeah* (95), *okay* (40), and *m* (31), and the three most frequent unimodal gestural ones are (multiple) nods (206), nod (32), and smile (27). The top three multimodal micro-feedback expressions are *yeah* + nods (62), chuckle (44), and *yeah* + nod (31). They are not only used to show evidence of understanding and willingness to continue, but also to express emotions and attitudes such as agreement, amusement, interest, and surprise.

4.2 Misunderstanding

Misunderstanding is infrequently related to micro-feedback, although more multimodal (6) than unimodal (3). Unimodal gestural micro-feedback is not associated with misunderstanding in our data.

The associated unimodal vocal-verbal micro-feedback expressions are *eh yeah eh* and *yeah*, which are usually expressed with hesitation. Also, the associated multimodal micro-feedback sometimes comprises of a repetition of the perceived vocal-verbal message and an assertive gesture nod for information confirmation. Misunderstanding is often not noticed by the interlocutors, however, it can be seen from an analyst's perspective by examining the interactional context.

4.3 Non-understanding

Non-understanding is revealed mostly by multimodal micro-feedback (19) rather than unimodal ones (4). These feedback items are often comprised of vocal-verbal expressions *what*, *huh*, or *huh* together with gestural expressions eyebrow raise or frown, gaze at or sideways, head forward, chuckle, or laughter, which are often used as eliciting devices for seeking further clarifications. The cases of non-understanding are revealed by unimodal gestural micro-feedback eyebrow raise and gaze at which express uncertainty and elicit further information.

5 Discussion

5.1 Unimodal gestural micro-feedback that exclusively shows sufficient understanding

Unimodal gestural micro-feedback almost exclusively relates to sufficient understanding. In our data, the most frequent unimodal gestural micro-feedback is head nod(s). This result corresponds well with others' findings of communicative feedback in several languages, such as Swedish and Finnish (Navarretta et al., 2012), Danish (Paggio and Navarretta, 2013), and Japanese (Ishi et al., 2014). In this study, all the unimodal head nod and (multiple) nods are found to exclusively express sufficient understanding rather than misunderstanding or non-understanding.

5.2 Gaze movements associated with misunderstanding and non-understanding

The data show that non-understanding is usually revealed by unimodal gestural micro-feedback eyebrow raise and gaze at or by multimodal micro-feedback comprised of head forward, eyebrow raise, and gaze at. Part of this finding supports Nakano et al.'s (2003) claim that maintaining gaze at the speaker is an evidence of non-understanding which usually evokes additional explanation. Equally important, in the data, misunderstanding is associated with multimodal micro-feedback that consists of gaze at, down, or sideways from the speaker. This result on gaze movement in association with misunderstanding and non-understanding expands Al Moubayed et al.'s (2013) and Jokinen et al.'s (2013) findings that gaze is not only important in inferring the speaker's intention of turn giving and turn holding, but also in providing responses to the perceived information and indicating the listener's understanding difficulties or problems.

5.3 *Yeah* and nod in relation to misunderstanding

As seen from an analyst's perspective, misunderstanding sometimes occurs even when unimodal micro-feedback *yeah* and nod are used. The data show that when a participant says *yeah* it does not always mean s/he truly understands. Especially when *yeah* is expressed in a hesitant prosody, it sometimes indicates an occurrence of misunderstanding. Equally important, misunderstanding can also occur when multimodal micro-feedback *yeah* + nod is employed. Other multimodal micro-feedback expressions that are related to misunderstanding usually comprise of a repetition of the perceived vocal-verbal message and an assertive gesture nod for information confirmation. Very likely, such a misunderstanding can result in further misunderstandings. The interlocutors sometimes just continue communicating without awareness or correction of the previously misunderstood information. This result is in line with Weigand's (1999) claim that the interlocutor who misunderstands is not always aware of it and the misunderstanding is not always corrected by the interlocutors.

5.4 Practical implications of visual modality in showing understanding

The present study also finds that visual modality (i.e., gesture) plays an important role in showing or revealing understanding; gesture is involved in around 74% of all the micro-feedback expressions that are related to the studied understandings. In addition, these gestures are almost entirely limited to the head region in the form of head movements and facial expressions. Hand and posture movements rarely occur in relation to understanding. Unimodal head movements are exclusively related to sufficient understanding. Based on these empirical findings, we suggest some possible guidelines for the design of communication technology applications. Such a system should include the visual modality since a large portion of micro-feedback occurs there. Further, the visual parts of the system, such as the graphical display and motion capture, can be limited to the head region of the agent without compromising the perception of understanding.

6 Conclusion

In this paper, we have studied understanding with a focus on micro-feedback in eight Chinese-Swedish intercultural conversations in English lingua franca. By using an interactional approach from an analyst's perspective, two research questions have been examined. First, how are the auditory and visual modalities involved in micro-feedback expressions that are related to sufficient understanding, misunderstanding, and non-understanding? Second, what are these typical unimodal and multimodal micro-feedback expressions?

The data have shown that most of the micro-feedback expressions are related to sufficient understanding, a few to non-understanding, and fewer to misunderstanding. This result suggests that misunderstanding is more difficult to observe in spontaneous communication. Further, sufficient understanding has been found more related to unimodal micro-feedback than multimodal; the typical unimodal micro-feedback expressions are *yeah*, *okay*, and *m* and multimodal ones are *yeah* + nod(s), and chuckle. Misunderstanding involves more multimodal micro-feedback than unimodal vocal-verbal micro-feedback, and it is not associated with unimodal gestural micro-feedback at all; when unimodal micro-feedback *eh yeah eh* and *yeah* are expressed with hesitation or when multimodal micro-feedback is comprised of a repetition of the perceived vocal-verbal message and an assertive gesture nod for information confirmation, a misunderstanding may have occurred. Non-understanding is mostly expressed by multimodal micro-feedback expressions and occasionally through unimodal ones; the typical micro-feedback comprises of vocal-verbal expressions *what*, *huh*, or *huh* together with gestural expressions eyebrow raise or frown, gaze at or sideways (from the speaker), head forward, chuckle, or laughter, which are often used as eliciting devices for further clarifications.

These findings can contribute to the practice of intercultural communication, for example, online and flexible learning, digital communication, and virtual agents' animation. The results can be exploitable in practical applications such as systems for speech and gesture recognition and understanding. Further research would be needed to strengthen and extend our findings beyond the cultural, language, and communication activity limitations of this study.

References

Jens Allwood. 1986. Some perspectives on understanding in spoken interaction. In: Mats Furberg, Thomas Wetterström, Claes Åberg, (Eds.). *Logic and Abstraction*. Acta Philosophica Gothoburgensia 1, pages 1–30.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources & Evaluation*, 41(3–4):273–287. https://doi.org/10.1007/s10579-007-9061-5

Jens Allwood. 2015. English translation of Tvärkulturell kommunikation (1985), *Papers in Anthropological Linguistics 12*, University of Göteborg, Danish Intercultural Organization, 17/04/2015.

Samer Al Moubayed, Gabriel Skantze, and Jonas Beskow. 2013. The Furhat back-projected humanoid head - Lip reading, gaze and multiparty Interaction. *International Journal of Humanoid Robotics*, 10(1). https://doi.org/10.1142/S0219843613500059

Michail M. Bakhtin. 1986. *Speech genres and other late essays*. C. Emerson, & M. Holquist (Eds.), V. W. McGee (Trans.). Austin: University of Texas Press.

Carla Bazzanella and Rossana Damiano. 1999. The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics*, 21(6):817–836. https://doi.org/10.1016/S0378-2166(98)00058-7

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294. https://doi.org/10.1207/s15516709cog1302_7

Morena Danieli and Carla Bazzanella. 2002. Linguistic markers in coming to understanding. In: *Atti dell'Ottavo Convegno dell'Associazione Italiana per l'Intelligenza Artificiale*, Siena, Italy.

Marcelo Dascal. 1999. Introduction: Some questions about misunderstanding. *Journal of Pragmatics*, 31(6):753–762. https://doi.org/10.1016/S0378-2166(98)00059-9

Harold Garfinkel. 1967. *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ.

Goodwin, 1981. *Conversational organization: Interactions between speakers and hearers*. Academic Press, New York.

John J. Gumperz, 1982. *Discourse strategies*. Cambridge University Press, Cambridge.

Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2014. Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, 57:233–243. http://dx.doi.org/10.1016/j.specom.2013.06.008

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *The ACM Transactions on Interactive Intelligent Systems*, 3(2):1–30. https://doi.org/10.1145/2499474.2499481

Shuya Kushida. 2011. Confirming understanding and acknowledging assistance: Managing trouble responsibility in response to understanding check in Japanese talk-in-interaction. *Journal of Pragmatics*, 43(11):2716–2739. https://doi.org/10.1016/j.pragma.2011.04.011

Oskar Lindwall and Gustav Lymer. 2011. Uses of “understand” in science education. *Journal of Pragmatics*, 43(2):452–474. https://doi.org/10.1016/j.pragma.2010.08.021

Per Linell. 2009. *Rethinking language, mind and world dialogically: Interactional and contextual theories of human sense-making*. Information Age Publishing, Charlotte, NC.

Michael Lynch. 2011. Commentary: On understanding understanding. *Journal of Pragmatics*, 43(2):553–555. https://doi.org/10.1016/j.pragma.2010.08.018

Douglas W. Maynard and Don H. Zimmerman, 1984. Topical talk, ritual, and the social organization of relationships. *Social Psychology Quarterly*, 47(4):301–316. http://dx.doi.org/10.2307/3033633

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. *Proceedings of the Meeting on Association for Computational Linguistics*. July 7–12, Sapporo, Japan. https://doi.org/10.3115/1075096.1075166

Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, Patrizia Paggio. 2012. Feedback in nordic first-encounters: a comparative study. In *Proceedings of LREC 2012*, May 2012, Istanbul, Turkey, pages 2494–2499.

Joakim Nivre, Jens Allwood, Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26. https://doi.org/10.1093/jos/9.1.1

Joakim Nivre, Jens Allwood, Leif Grönqvist, Magnus Gunnarsson, Elisabeth Ahlsén, Hans Vappula, Johan Hagman, Staffan Larsson, Sylvana Sofkova, Cajsa Ottesjö. 2004. *Göteborg Transcription Standard Version 6.4*. Department of Linguistics, Göteborg University.

Patrizia Paggio and Costanza Navarretta. 2013. Head movements, facial expressions and feedback in conversations: empirical evidence from Danish multimodal data. *Journal on Multimodal User Interfaces-Special Issue on Multimodal Corpora*, 7(1–2):29–37. https://doi.org/10.1007/s12193-012-0105-9

Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis, (Eds.). 2012. *Multimodal Signal Processing: Human Interactions in Meetings*. Cambridge University Press, Cambridge.

Emanuel A. Schegloff. 1996. Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pages 52–133). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511620874.002

Jan Svennevig. 1999. *Getting acquainted in conversation*. John Benjamins Publishing Company, Amsterdam.

Deborah Tannen. 1990. *You just don’t understand: Women and men in conversation*. William Morrow, New York.

Darinka Verdonik. 2010. Between understanding and misunderstanding. *Journal of Pragmatics*, 42(5):1364–1379. https://doi.org/10.1016/j.pragma.2009.09.007

Edda Weigand. 1999. Misunderstanding: The standard case. *Journal of Pragmatics*, 31(6):763–785. https://doi.org/10.1016/S0378-2166(98)00068-X

Victor H. Yngve. 1970. On getting a word in edgewise. In M. A. Campell et al. (Eds.), *Papers from the sixth regional meeting of the Chicago linguistic society* (pages 567–577). Chicago Linguistic Society, Chicago.

Dietmar Zaefferer. 1977. Understanding misunderstanding: A proposal for an explanation of reading choices. *Journal of Pragmatics*, 1(4):329–346. https://doi.org/10.1016/0378-2166(77)90027-3

Jordan Zlatev. 2009. Levels of meaning, embodiment, and communication. *Cybernetics and Human Knowing*, 16(3–4):149–174.

Towards a Types-As-Classifiers Approach to Dialogue Processing in Human-Robot Interaction

Julian Hough[1,*], **Lorenzo Jamone**[1,†]
David Schlangen[2,‡], **Guillaume Walck**[2,§] and **Robert Haschke**[2,§]
[1] * Cognitive Science Group, [†]Centre for Advanced Robotics,
School of Electronic Engineering and Computer Science,
Queen Mary University of London, UK
[2] [‡]Dialogue Systems Group, [§]Neuroinformatics Group,
CITEC, Bielefeld University, Germany
j.hough@qmul.ac.uk

Abstract

We propose a novel Types-As-Classifiers approach to dialogue processing for robots using probabilistic type judgments. In our proposal, incoming sensory data is converted to a world belief record in real time, and then derived beliefs such as intention attribution to a user, or the prediction of affordances of visible objects, are made as record type judgements of that record. The record can be updated dynamically like a dialogue state, allowing information of different perceptual sources to be easily combined in real time.

1 Introduction

The combination of computer vision and natural language processing is now incredibly popular. Thanks to increased computing power and the development of new deep learning techniques, huge strides forward have been made in several tasks, including: automatic image retrieval from key words, reference resolution of objects in photographs from text (Kennington and Schlangen, 2015), generating referring expressions to objects given probabilistic estimation of object properties (Mast et al., 2016), caption generation and visual question answering (Antol et al., 2015).

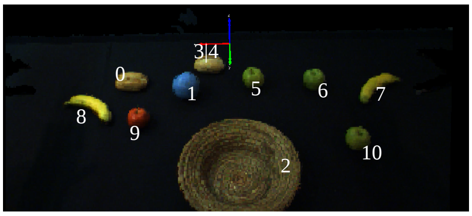
A more challenging task, beyond the use of single sentence texts with images, is the creation of dialogue systems designed for real-world human-robot interaction (HRI) which combines probabilistic information encoding visual and physical properties of objects and information about the interaction more commonly encoded in a dialogue state. This uniform approach not only requires the use of complex visual information and semantic parsing, but needs to permit fluid interaction with a collaborative robot to help a user complete a manual task. This requires an incrementally and dynamically evolving dialogue state which encodes the robot's own action state as well as its estimation of the user's intentions in real time.

In this paper we address this challenge by formulating a simple interaction state for a robot using concepts from Type Theory with Records (TTR) (Cooper, 2005). We characterize the robot's world belief as a constantly updating record, and use type classifiers of different kinds which operate on the state record to make type judgements on the world belief. Once a judgement is made and used (committed), this can be added to the world belief for further classification and update. For the classification we use a combination of lattice theory and probabilistic TTR (Cooper et al., 2014). Inspired by the recent work using TTR for perceptual classification (Dobnik et al., 2012; Yu et al., 2016) and the simple Words-As-Classifiers (WAC) model (Kennington and Schlangen, 2015) to reference resolution of objects in real-world scenes, here we propose a general Types-As-Classifiers (TAC) approach.

2 Types-As-Classifiers for human-robot interaction

Typical raw perceptual information for a collaborative pick-and-place robot may be as in Fig. 1. The left side shows a camera feed, and computer vision based segmentation and tracking of objects as described in (Ückermann et al., 2014a,b), and perceptual classifiers, such as that for 'yellow', which classify the degree to which an object has that perceptual property. The current words recognized by the robot's speech recognizer (ASR) are also added to the state as they arrive. On the right side, the diagram shows how the robot tracks its own current task state and action state of its arm through a Hierarchical State Machine (HSM).

SCENE:



OBJECTS (segmentation and visual classifiers):
object_0:
yellow = 0.69
blue = 0.38
..
object_1:
yellow = 0.10
blue = 0.86
...

USER SPEECH (current user utterance):
'put the left green apple in the basket'

ROBOT ACTION AND TASK STATE:

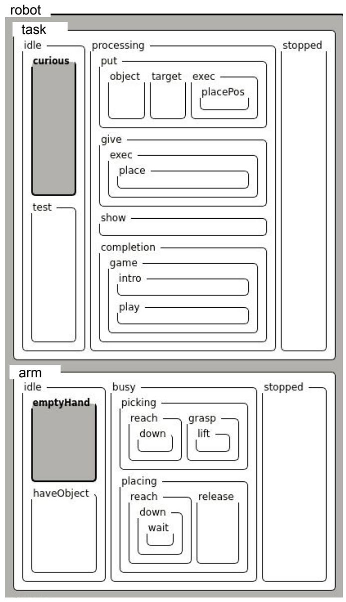


Figure 1: A typical state according to the robot. Objects are segmented and properties can be obtained for each object. The robot's internal action state is controlled by a Hierarchical State Machine (HSM)

2.1 Encoding the robot's sensory state as an updating TTR record

In this paper we use TTR *record types*, and the inhabitants of record types, *records*, as our primary formal apparatus – see Cooper (2005) for details. We characterize the state as a *world belief record*-for an in-robot control system for our purposes it will be of the format in (1).[1]

$$\left[\begin{array}{lcl} objects & = & \left[\begin{array}{lcl} obj\_0 & = & [\ ... = ...\ ] \\ obj\_1 & = & [\ ... = ...\ ] \\ ... & = & ... \\ obj\_N & = & [\ ... = ...\ ] \end{array}\right] \\ robot & = & \left[\begin{array}{lcl} arm & = & [\ ... = ...\ ] \\ task & = & [\ ... = ...\ ] \\ intention & = & [\ ... = ...\ ] \end{array}\right] \\ human & = & \left[\begin{array}{lcl} c{-}utt & = & \left[\begin{array}{lcl} parse & = & ... \\ words & = & ... \end{array}\right] \\ status & = & ... \\ intention & = & [\ ... = ...\ ] \end{array}\right] \end{array}\right] \quad (1)$$

For HSMs as in Fig. 1, we can formulate the state at a given time as a record via the use of recursive structure. The record gets constructed from the highest level down, whereby each parallel/concurrent state, such as the *task* and *arm* substates of *robot* in Fig. 1, are encoded as separate fields in the record. If the current state is an embedded substate, for example the *emptyHand* and *holdsObject* substates within the *idle* substate of the *arm* state in Fig. 1, that will be encoded in the record structure as an embedded record (a record within a record). When a state is atomic, that will be encoded as a single value in the record.

Given this recursive formulation, the robot's current action and task state as shown by the darkened areas in Fig. 1 can be formulated as in (2). This is an efficient way of encoding the state, as not all the inactive substates need be encoded.

$$\left[\ robot = \left[\begin{array}{lcl} task & = & [\ idle = curious\ ] \\ arm & = & [\ idle = emptyHand\ ] \end{array}\right]\ \right] \quad (2)$$

3 Record Type classifiers applied to the world belief for higher-level perception

The driving incremental interpretation process of the system is a probabilistic classification of the current world belief record wb (with the structure in (1)) as being of a given situation record type i within a set of possible record types I, conditioned by current evidence record type e.

In the following sub-sections we outline different perceptual classifiers which operate on wb to get the probability judgement that wb is of a given type. This can be done recursively, as once a type judgement is made (for a given purpose), this can be added to wb, and then further judgements of its

[1]This is an example record where many of the labels and values are just represented by '...' to indicate at least one such field would be present in the full representation.

$$i = \left[human : \left[intention : \left[\begin{array}{lll} goal & : & \left[\begin{array}{lll} landmark & : & obj\_2 \\ rel\_location & : & INTO \end{array} \right] \\ objects & : & \{obj\_1\} \\ action & : & PUT \end{array} \right] \right] \right]$$

Figure 2: A user intention record type to effect the movement of an object.

type can be made and added to it. While we suggest a pipeline here by presentation order, we are not committed to a specific classification ordering or algorithm for inter-leaving these processes, and leave investigation into this for future work. However, we are committed to the distribution over possible record type judgements being stored in a record type lattice– see (Hough and Purver, 2017).

3.1 Perceptual classification 1: predicting object affordances

The robot's perception of object properties is vital for complex interaction with the human user. Specifically, the perception of object *affordances* (Gibson, 1979), i.e. the possible actions associated to the objects (e.g. *graspable*), is crucial for the robot to be able to manipulate them (Jamone et al., 2016). Recently, probabilistic computational models of affordance perception have been proposed, using Bayesian Networks (Gonçalves et al., 2014) and variational auto-encoders (Dehban et al., 2016)- these can be used obtain the probability of an object having different affordances from visual and linguistic features. In our model, affordance prediction is part of the probabilistic type judgement of wb, such that the probabilities of each object having each affordance property are part of the available type judgements. In future work, we will investigate how affordance prediction can best integrate with natural language processing decisions – e.g. (Salvi et al., 2012).

3.2 Perceptual classification 2: parsing

The next higher-level perception classification is the incremental semantic parsing of the recognized words from the ASR. For this we use the Dylan ('DYnamics of LANguage') parser (Purver et al., 2011).[2] The parser fulfills the criteria for incremental semantic construction outlined in (Hough et al., 2015): it consumes words one-by-one and outputs a maximal semantic record type (RT) based on a pre-defined Dynamic Syntax-TTR (DS-TTR) grammar– see (Eshghi et al., 2011) for full details. A typical parse for 'put the red apple in the big basket' is as in (3):

$$\left[\begin{array}{lll} r1 & : & \left[\begin{array}{lll} x & : & e \\ p_{=basket(x)} & : & t \\ p1_{=big(x)} & : & t \end{array} \right] \\ x2_{=\iota(r.x)} & : & e \\ r & : & \left[\begin{array}{lll} x & : & e \\ p_{=apple(x)} & : & t \\ p1_{=red(x)} & : & t \end{array} \right] \\ x1_{=\iota(r.x)} & : & e \\ e2_{=INTO} & : & es \\ x_{=addressee} & : & e \\ e_{=PUT} & : & es \\ p3_{=obj(e2,x2)} & : & t \\ p2_{=indObj(e,e2)} & : & t \\ p1_{=obj(ev,x1)} & : & t \\ p_{=subj(ev,x)} & : & t \end{array} \right] \quad (3)$$

The best parse is added to the $human.c{-}utt.parse$ field. Now other inference can be done using this information, primarily recognizing the user's intention word-by-word.

3.3 Perceptual classification 3: user intention recognition

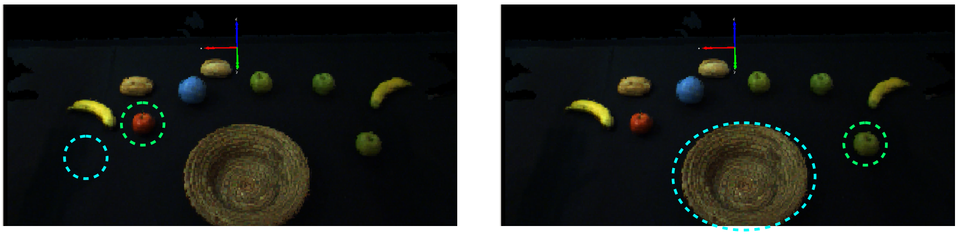
As DyLan's DS-TTR parser provides RTs word-by-word incrementally, the user's intention can also be estimated word-by-word as wb is updated. Given a set of possible user intention record types I, where a typical intention may look like i in Fig. 2, and the conditioning evidence e, a record type representing a sub-part of wb, we characterize a standard Maximum Likelihood multi-class probabilistic classifier to estimate the best prediction for the $human.intention$ field and its probability (or *confidence*) in its prediction $Ev(human.intention)$ by the standard $arg\ max$ and max functions in (4) and (5), respectively.

$$human.intention = \arg\max_{i \in I} p(wb : i | wb : e) \quad (4)$$

$$Ev(human.intention) = \max_{i \in I} p(wb : i | wb : e) \quad (5)$$

In our current implementation, e simply consists in judgements on the $human.c{-}utt.parse$

[2]Available open-source at `https://bitbucket.org/dylandialoguesystem/dsttr`.



put the apple in front of the banana ... in the basket

Figure 3: Syntactic ambiguity causing the system changing its top hypothesis about the user's intention.

and $objects$ fields of wb, but it can be more than these, and in future, we plan to learn which parts are relevant for estimating user intentions.

In our current implementation, to calculate the conditional likelihood $p(wb : i \,|\, wb : e)$ for two given RTs i and e, we create a directed graph of the current parse RT based on its field dependencies, beginning from the head event field $e_{=PUT}$ (which determines the action), and recursively traverse all fields which depend on it, applying the relevant type classifiers. We match the field values in the embedded entity restrictor RTs such as $red(x)$ to the low-level classifier results in $objects$. If the relevant type judgement (e.g. $red(x)$) appears in the parse, the corresponding low-level classification strength for each object (e.g. $obj\_1.red = 0.8$) will be used, using the product rule to multiply the probability of the relevant fields for a given object. The overall likelihood of $wb : i$ is calculated recursively, beginning with the likelihood of the embedded RTs such as $intention.goal$ and the target objects $intention.objects$. The likelihood of the judgements of each of the embedded fields is multiplied together to get the overall probability of the intention.

3.4 Perceptual classification 4: estimating legibility of robot intentions

Dual to confidence about the user's intention, we can also estimate the *legibility* of the robot's intention (Dragan et al., 2013), which is similar in structure to the human intention in Fig. 2. Legibility is important for estimating when the robot's intention has become distinct enough from other possible intentions, and consequently what can be considered grounded with the user through the robot's action so far (Hough and Schlangen, 2017). We estimate the strength-of-evidence function $Ev(robot.intention)$ as in (6) where e is taken to be all of wb excluding $robot.intention$:

$$Ev(robot.intention) = p(wb : robot.intention|wb : e) \tag{6}$$

(6) is the likelihood that the robot's current intention will be recognized by the user as such. In practice, this legibility measure can be estimated via a number of physics-based methods such as the proximity of the arm to the target object compared to the other objects, or through using movement trajectories– see (Dragan et al., 2013).

4 Conclusion

We have given an overview of a Types-As-Classifiers (TAC) approach to dialogue processing in human-robot interaction. We believe our approach is complementary to the Words-As-Classifiers (WAC) approach to reference resolution (Kennington and Schlangen, 2015), and we believe it brings several advantages. Firstly, it is not constrained by individual word classifiers alone, but can use the structure from a parser to compute likelihood of complex intentions, all the while maintaining word-by-word incrementality. Secondly, it gives a uniform way to process different multimodal information such as robotic task and action states and visual and physical properties of objects within a dialogue state. In future, we intend to show how it allows the different processes to help each other- e.g. the online resolution of parsing ambiguity such as that in Fig. 3, where the first 'in' is taken not to modify 'the apple', but this decision is changed once the user continues talking. We are also planning to test our current implementation with users.

Acknowledgments

We thank the reviewers for their useful comments. This work was supported by the DFG Center of Excellence EXC 277, the DFG Transregional Research Centre CML, TRR-169.

References

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

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

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*, Gothenburg, Sweden. ACL.

Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. 2016. Denoising autoencoders for learning of objects and tools affordances in continuous space. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4866–4871. IEEE.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *International Workshop on Constraint Solving and Language Processing*, pages 70–91. Springer.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE.

Arash Eshghi, Matthew Purver, and Julian Hough. 2011. DyLan: Parser for Dynamic Syntax. Technical Report EECSRR-11-05, School of Electronic Engineering and Computer Science, Queen Mary University of London. ISSN 2043-0167. Available from http://sf.net/projects/dylan/files/dylan/DSImp_TechReport.pdf.

James J Gibson. 1979. The theory of affordances. *The people, place, and space reader*, pages 56–60.

Afonso Gonçalves, João Abrantes, Giovanni Saponaro, Lorenzo Jamone, and Alexandre Bernardino. 2014. Learning intermediate object affordances: Towards the development of a tool concept. In *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014*, pages 482–488. IEEE.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 206–216.

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.

Julian Hough and David Schlangen. 2017. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 Conference on Human-Robot Interaction (HRI2017)*.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. 2016. Affordances in psychology, neuroscience and robotics: a survey. *IEEE Transactions on Cognitive and Developmental Systems*.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the Conference for the Association for Computational Linguistics (ACL). ACL.

Vivien Mast, Zoe Falomir, and Diedrich Wolter. 2016. Probabilistic reference and grounding with pragr for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(5):889–911.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th IWCS*, Oxford, UK.

G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor. 2012. Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(3):660–671.

Andre Ückermann, Christof Eibrechter, Robert Haschke, and Helge Ritter. 2014a. Real-time hierarchical scene segmentation and classification. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 225–231. IEEE.

Andr Ückermann, Christof Elbrechter, Robert Haschke, and Helge Ritter. 2014b. Hierarchical Scene Segmentation and Classification.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 339.

Perceptual semantics and dialogue processing

Staffan Larsson
Centre for Linguistic Theory and Studies in Probability (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg

Abstract

This paper is a preliminary investigation into the possible relations between an utterance and the situation it is about, and the consequences of different relations on utterance processing in dialogue.

1 Introduction

How are information state updates that result from utterances in dialogue related to the modeling of perceptual semantics as classifiers? The most straightforward answer, and that which is assumed in pretty much all the literature on modeling perceptual semantics as classifiers, is to limit the theory to situations where utterances describe (through more or less explicit assertions) a situation which is represented by some immediately available perceptual input. This is similar to the situation in early first language acquisition, where parents and children discuss objects and relations which are in a shared focus of (perceptual) attention.

2 Possible relations between utterance and situation

The guaranteed availability of perceptual input p derived from the situation at hand s means that as soon as an utterance u with content T is made, the hearer can judge whether u correctly describes s, for example by judging whether s is a situation of the type[1] described by e, that is, $s : T$. Based on this judgement the hearer can then decide whether to accept or reject the utterance[2].

However, this is just one of many possibilities:

- talk about the utterance situation
  - assertion, e.g. ”the man is to the left of the box”.
  - asking, e.g. ”what is to the left of the box?”
- talk about a past situation
  - assertion, e.g. ”Yesterday Fredrik wore jeans”
  - asking, e.g. ”who wore jeans?”
- talk about a future situation or type of situations
  - assertion, e.g. ”it will rain tomorrow”
  - asking, e.g. ”will it rain tomorrow” (y/n) or ”what will be the weather tomorrow?” (wh)
  - requesting, e.g. ”put the box on the table”
- talk about situations in general (all situations)
  - assertion, e.g. ”dogs are mammals” [perceptual?]
  - asking, e.g. ”name a type of mammal”

3 Type acts

Related to this, Cooper (2014) lists several *type acts* – things one can do with types:

judgements

specific $o :_A T$ “agent A judges object o to be of type T”

non-specific $:_A T$ “agent A judges that there is some object of type T”

queries

specific $o :_A T?$ “agent A wonders whether object o is of type T”

[1] Following Cooper (in progress), we assume that the content of assertions are types of situations.

[2] A judgement $s : T$ need not lead to an acceptance, and the opposite judgement need not lead to a rejection. For instance, the hearer may instead revise her take on the s (reconsideration of the facts) or T (linguistic learning).

non-specific $:_A$ T? "agent A wonders whether there is some object of type T"

creations

non-specific $:_A$ T! "agent A creates something of type T"

Cooper remarks that "...creations only come in the non-specific variant. You cannot create an object which already exists."

4 Temporal relations and perceptual evidence

Cooper's taxonomy of type acts accounts for much of the variation between different kinds of relations between situation and utterance. As our list above indicates, there are also some constraints regarding the temporal relation between the situation talked about and the utterance time, so that creations (requests) do not make sense when talking about a non-future situation.

We talk above about the utterance situation, but does it really matter (for utterance processing) if the situation talked about is the utterance situation? On reflection, we would argue it does not, except insofar that this affects availability of perceptual information about the situation talked about. Note for example that when talking about a situation which we do not yet have perceptual evidence about, it does not matter if the situation has already happened or not; what matters is if we have perceptual evidence or not (which we may not, even if the situation has happened; of course if the situation has not happened we cannot yet have perceptual evidence).

This points to a need for a formal notion of perceptual evidence. We take this to be a type T_{PE}, a type of the situation talked about, so that $s : T_{PE}$. This represents an agent's perceptual take on a situation. Judgments about any situation s are mediated by T^s_{PE}. This mediation can be expressed as $s : T_u$ if $s : T^s_{PE}$, where ${T_{PE}}^s \sqsubseteq T_u$ is perceptual evidence derived from s and T_u is the meaning of an utterance u. Note that T^s_{PE} is a mental entity, and it does not matter if it is acquired from a photo of s or by "direct perception" (if there is such a thing). We can also talk about the *perception time* t^s_{PE} which is the time when an agent acquired perceptual information about s.

5 Consequences for dialogue processing and behaviour

How can we react to assertions about (in principle, disregarding time) perceivable situations? We may reject or accept them based on judgement $s : T$ only if T^s_{PE} is already available, which requires that $t^s_{PE} < t_u$ (where t_u is the utterance time). This excludes talking about future situations, but also situations in the past or present which have not yet been perceived.

In cases where $t^s_{PE} \geq t_u$, a hearer is faced with a more complicated situation. She may reject or accept based on *prediction* about future (perceptual) evidence, or she may provisionally accept ("we'll see", "perhaps"). Importantly, provisional acceptance seems to be connected to a right to later reject u in light of evidence not available at t_u ("you said it would rain, but it's snowing!").

We may now revise our taxonomy of possible relations between utterance and situation, and amend it with dialogue options. We replace temporal relations between utterance and sitation talked about with relations between utterance time and evidence time. We leave out talk about situations in general since they concern many (all) situations and therefore do not fit directly with the notion of perception time (which is bound to a single situation).

- Evidence is available; $t_u > t^s_{PE}$
  - assertion, e.g. "the man is to the left of the box" or "yesterday Fredrik wore jeans": make specific judgement $s : T_u$ and reject/accept
  - asking, e.g. "what is to the left of the box?" or "did Fredrik wear jeans?": integrate specific query $s : T_u$? (e.g. push on QUD), make judgement[3] and respond
- Evidence not yet available: $t_u \leq t^s_{PE}$
  - assertion, e.g. "it will rain tomorrow":
    1. withhold judgement, e.g. "maybe", "we will see", or make provisional judgement, e.g. "ok, let's assume that"
    2. predict future evidence and reject/accept based on this

[3]This will look a bit different depending on whether then question in y/n or wh.

– asking, e.g. ”will it rain tomorrow” (y/n) or ”what will be the weather tomorrow?” (wh): integrate specific query $s : T_u?$ (e.g. push on QUD), indicate that information is not available (”I don’t know”), possibly commit to answering later (”I will let you know if I find out”)
– requesting, e.g. ”put the box on the table”: evaluate request[4]

6 Conclusion

We have made a first stab at clearing up the possible relation between utterances and the situations they talk about, at least as concerns situations that we could in principle perceive. We conclude that the notions of type acts and perception time is useful in such an endeavor. Much work remains, not least with respect to further working out the dialogue processing options available for the different cases.

References

Robin Cooper. 2014. How to do things with types. In *Joint proceedings of the second workshop on Natural Language and Computer Science (NLCS 2014) & 1st international workshop on Natural Language Services for Reasoners (NLSR 2014) July*, pages 17–18.

Robin Cooper. in progress. *Type theory and language - From perception to linguistic communication*.

[4]We do not have anything to say about this here.

Are we having a laugh? Conversational laughter in schizophrenia

Mary Lavelle,[1] Christine Howes,[1] Patrick G.T. Healey,[2] Rosemarie McCabe[3]

[1]University of Gothenburg
[2]Queen Mary University of London
[3]University of Exeter

Abstract

Social exclusion and social dysfunction are persistent and debilitating aspects of schizophrenia. The interactional impact of patients' social deficits during actual dialogue is poorly understood. Through analysis of a corpus of patients' triadic interactions we explored laughter as a marker of discomfort or coalition in patients' interactions. Patient interactions did not differ from controls in terms of laughter production. However, patients who were more symptomatic laughed less frequently, while their partners showed a trend for displaying more shared laughter, potentially indicating coalition formation.

1 Introduction

Schizophrenia patients have difficulty interacting with others and are one of the most socially excluded groups in society (Huxley & Thornicroft, 2003; Social Exclusion Unit, 2004). Although some of patients' social exclusion may be due to stigma from others, patients' interactional difficulties may further compound this problem. The nature of patients' social deficits remains unclear. Evidence from the field of social cognition suggests that patients with schizophrenia have difficulty perceiving and interpreting social cues from others such as those conveyed through verbal and nonverbal communication (Green, 2016). However, this evidence has been gathered from off-line pen and paper tests, which patients complete in isolation. Such tests are far removed from the social context they represent and it is unclear if patients' performance on these tests represents their social cognitive skills during actual dialogues with others. Furthermore, we know little about the impact patients' social deficits may have on others' perception of the interaction, their ability to engage in social interaction and develop relationships with them.

In order to explore such questions, we have collected a corpus of interactions involving patients with schizophrenia and unfamiliar healthy controls, who are unaware of patients' diagnoses, thus eliminating the element of stigma. Analysis of nonverbal communication in this corpus revealed that the undisclosed presence of a patient in a triadic interaction changed the nonverbal behaviour of patients' interacting partners (Lavelle et al., 2012), as well as patterns of filled and unfilled pauses (Howes et al., 2017). Furthermore, patients' increased gesture use when speaking was associated with their partners perceiving the interaction more negatively, reporting experiencing poorer rapport with patients (Lavelle et al, 2012). This suggests that patients' partners may experience difficulty on an interpersonal level when interacting with a patient.

Laughter can be as a marker of discomfort or awkwardness in social interaction (Haakana et al., 2002). In multiparty interaction, shared laughter may also indicate coalition between the laughing parties (Osvaldsson, 2004; Bryant, 2012). This study investigated laughter in the corpus of patients' triadic interactions, specifically examining shared laughter as markers of coalition formation.

2 Methods

2.1 Participants

The study consisted of two conditions: (i) a patient condition, comprising 20 patient groups (one schizophrenia outpatient and two healthy participants) and (ii) a control condition, comprising 20 control groups (three healthy participants). All interacting partners had not met prior to the study. Patients' partners were unaware of the patients' diagnosis and all participants were naive to the purposes of the study. Thus, the interactions were as naturalistic as possible within the motion capture environment.

2.2 Dialogue Task

Interactions were audio-visually recorded using two, 2-D video cameras and simultaneously motion captured in 3-D. Participants discussed a fictional moral dilemma called 'the balloon task', and reached a joint decision on the outcome. The task states that there are four people in a hot air balloon, the balloon is losing height and is going to crash into the mountains killing everyone on board. The only way to save the balloon is to select one person that can be thrown from the balloon, saving the lives of the remaining three. The four passengers are: Dr. Nick Riviera – a cancer research scientist, who believes he is on the brink of discovering a cure for most common types of cancer; Mrs. Susanne Harris – who is a primary school teacher and over the moon because she is 7 months pregnant with her second child; Mr. William Harris – the pilot of the balloon, and the only one on board with balloon flying experience, he is also the husband of Susanne, who he loves very much; Miss Heather Sloan – a 9 year-old music prodigy, considered by many to be a "twenty first century Mozart".

2.3 Symptom Assessment

Patients' symptom severity was assessed using the Positive and Negative Syndromes Scale (Kay et al., 1987). There are two main symptom groups in schizophrenia, positive symptoms referring to the additional aspects that patients experience such as hallucinations and delusional beliefs, and negative symptoms, which refer to the reduction in normal experience such as the expression and experience of emotions, motivation, social activity. Patients receive a score for each symptom group and an overall symptom severity score.

2.4 Interpersonal Rapport

Following the interaction all participants rated the level of rapport they experienced with each of their interacting partners on a scale of 1-10.

2.5 Dialogue annotation

Laughter was hand coded using the ELAN annotation tool. Each laughter event was categorised as 'shared laughter' – laughing at the same time as another interacting partner, or 'individual laughter' – laughter occurs alone, in the absence of laughter by others.

2.6 Analysis

The duration of laughter as a percentage of whole interaction was calculated for each individual. The frequency of laughter events (shared or individual) by interaction duration was also calculated for each individual. Participant types were compared using a mixed model regression analysis adjusting for triadic group. Correlational analysis examined the relationship between frequency of laughter events displayed by participants (shared and individual) and (i) patients' symptoms and (ii) rapport score received from others.

3 Preliminary Results

Patients or their healthy participant partners did not significantly differ from controls in terms of the frequency of shared or individual laughter they produced during the interaction (figure 1).

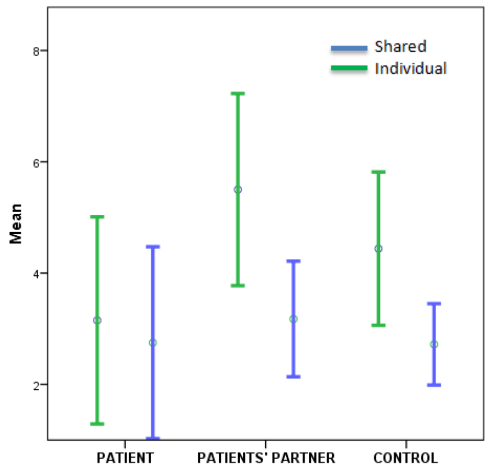


Figure 1. Mean frequency of shared and individual laughter events per second by participant type.

Patients with more negative symptoms (e.g. social withdrawal and diminished affect) laughed less frequently (Rho (20)=-.50, p=.03). Patients' increased positive symptoms (e.g. hallucinations and delusional beliefs) are associated with their partners displaying less shared laughter (Rho(40)=-.34, p=.03).

Control participants showed a significant positive association between their laughter duration and the rapport score they received from others (Rho(48)=.43, p=.001). This positive association with rapport was evident both for shared laughter (Rho(48)=.35, p=.01) and individual laughter (Rho(48)=.44, p=.002).

However, patients' partners who had a higher frequency of shared laughter events received a higher rapport score from others (Rho(25)=.46, *p*=.02). No other relationships between rapport and laughter in the patient condition were significant.

4 Discussion

The preliminary results showed no significant difference in the frequency or duration of laughter events in patient and control interactions. However, patients with more negative symptoms laughed less often, and patients' increased positive symptoms was associated with their partners displaying less shared laughter. This was seen despite patients having only mild to moderate symptom levels and displaying no overt symptoms during the interaction task.

A significant positive relationship between all forms of laughter (shared and individual) and interpersonal rapport was identified in control interactions. Although this relationship was not apparent in patients' interactions, shared laughter was associated with better rapport scores in patients' partners. However this may be mediated by patients' symptoms.

Overall it appears that the large variations in laughter presentation across all groups (figure 1), make it difficult to draw conclusions from this level of analysis. Patients' symptoms appear to influence their own production of laughter and the shared laughter of their partners. Furthering our understanding of the role of laughter in patients' interactions requires analysis at a more fine grained level, examining laughter in the context of when it occurs in the interaction, whether patients lead or follow in the shared laughter events, and the temporal relationship of laughter to specific conversational features such as turn-taking. This more comprehensive analysis will form the focus of this presentation.

References

Bryant G. Shared laughter in conversation as a coalition signaling *Paper presented at: XXI Biennial International Conference on Human Ethology* 2012; Vienna, Austria.

Green MF. 2016. Impact of cognitive and social cognitive impairment on functional outcomes in patients with schizophrenia. *The Journal of clinical psychiatry*. 77: 8-11.

Haakana M. 2002. Laughter in medical interaction: From quantification to analysis, and back. *Journal of Sociolinguistics.* 6(2):207-235.

Howes, C., Lavelle, M., Healey, P. G. T., Hough, J. & McCabe, R. 2017. Disfluencies in dialogues with patients with schizophrenia. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. London, UK.

Huxley P, and Thornicroft G. 2003. Social inclusion, social quality and mental illness. *British Journal of Psychiatry.* 182: 289-290.

Kay S, Friszbein A, Opler LA. 1987. The Positive and Negative Syndrome Scale for Schizophrenia. *Schizophrenia Bulletin*. 13:261-276.

Lavelle M, Healey PGT, McCabe R. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin.* 39 (5):1150-1158.

Osvaldsson K. On laughter and disagreement in multiparty assessment talk. Vol 24. Berlin, ALLEMAGNE: Mouton de Gruyter; 2004.

Social Exclusion Unit. Mental Health and Social Exclusion. Minister Office of the Deputy Prime Minister, creating sustainable communities. London; 2004.

The challenge of challenging others: Negotiation of performance feedback in interprofessional clinical teams

Mary Lavelle,[1,2] Gabriel B Reedy,[1] Thomas Simpson, [1] Janet E. Anderson[1]

[1]King's College London
[2]Imperial College London

Abstract

In healthcare settings, poor team communication is a leading cause of patient safety failure. Theoretical models of teamwork suggest that 'speaking up' to question the performance of team members is critical to effective teamwork. However, evidence of real world settings suggests that this is difficult to achieve. Through analysis of conversations during simulated clinical scenarios we examined how clinical teams communicate feedback about performance to each other, and the impact this has on the team and clinical performance.

1 Introduction

The effective and safe provision of healthcare relies on multiple professionals, from a different healthcare disciplines, working together with the common purpose of managing and treating the patient. This can only be achieved through good communication among team members (Manser, 2009). However, in practice, interprofessional teamwork is complex and may be difficult to achieve for a variety of reasons including: cultural and educational differences between professions, perceived and actual hierarchies, staff attitudes and perceptions of interprofessional working (Weller et al., 2014; Hall, 2005; Liberati et al., 2016) and the transient and episodic nature of teams in clinical settings (Chesluk et al, 2015).

Theoretical models of teamwork suggest that effective teams must be able to 'speak up' when lapses or errors in teammates' performance are identified (Kolbe et al., 2012). However, given the potential barriers to communication, this may be difficult to achieve in interprofessional teams. Moreover, the evidence for these models stems from the specialties of surgery and anesthesia, where the patient does not have an active role in the team. Sensitivity towards the conscious patient in ward settings may impose additional constraints to team communication.

The aim of this study was to explore how challenges to others' performance are negotiated in interprofessional ward-based clinical teams, exploring who does it and how, and the link with clinical performance.

2 Methods

2.1 Participants and Scenario

Nine audio-visually recorded simulated scenarios depicting a pregnant woman deteriorating due to a medical condition provided the corpus for the current analysis. The scenarios were recorded as part of a Multi-Disciplinary Simulation Training for Medical Emergencies in Obstetrics (MEmO) (Lavelle et al., 2018). Each scenario involved a simulated patient, played by the Maternal Simulator, an embedded practitioner (plant), playing the role of a student midwife, and course participants (range: 3-5) who were all qualified full time midwives and medical doctors (obstetricians, medical physicians and anesthetists). The course participants were instructed to respond to the events in the simulation as they would during a routine shift.

Teamwork Domains	Behavioural Functions	Observable Behaviours
Leading the Team	Delegating	Instruction Workload re-distribution
	Information Gathering	Requesting information Nonerbal monitoring of the situation
	Planning	Task setting Goal setting
	Disseminating Rationale	Coordinating information from multiple sources Verbalising the big picture Verbalising the task rationale
Developing a Shared Mental Model	Clarifying Information	Evaluating others' contributions Providing information on request Information reflection
	Information Sharing	Spontaneous speaking to the room (actions/information)
Assisting	Requesting Assistance	Asking others for help
	Providing Assistance	Helping others with their tasks Completing others' tasks for them
Monitoring Team Performance	Explicit Performance Monitoring	Speaking up to question/challenge others' performance
	Implicit Performance Monitoring	Subtle challanging others performance Nonverally monitoring others' performance
Team Attituides	Positive Attitude	Valuing others' opinons Positive Comments to others
	Negative Attituide	Ignoring others
	Disagreement	Task based disagreement Process based disagreement

Figure 1. Framework for the Temporal Observational Analysis of Teamwork (TOAsT) for healthcare settings.

2.2 Multi-modal Annotation

The verbal and nonverbal behaviour of each participant, in each of the nine scenarios, was annotated using a framework for the Temporal Observational Analysis of Teamwork (TOAsT) in healthcare (Lavelle et al., submitted).

The framework (figure 1) is comprised of five overarching ***'teamwork domains',*** which contain twenty-four *'**observable behaviours**'*, which are the specific verbal and nonverbal behaviours that can be identified during observations. The behaviours are also grouped conceptually based on their function resulting in thirteen ***'behavioural functions'.*** The framework has excellent inter-rater agreement (Intra-class Correlation Coefficient (ICC) =.83 (95%CI .79-.87; p<.0001)) and was designed to be applicable to a range of healthcare professionals (e.g. doctors, nurses, midwives, healthcare assistants), across a variety of healthcare settings (e.g. acute management, routine care, mental health, physical health). It is underpinned by current teamwork theories literature from the field of human factors in healthcare (Kolbe et al., 2013; Salas et al., 2005).

Video annotation was conducted using the annotation software ELAN (Sloetjes & Wittenburg, 2008) (Figure 2). Behavioural annotations of each participant in each scenario were exported as time series into SPSS for temporal and statistical analysis.

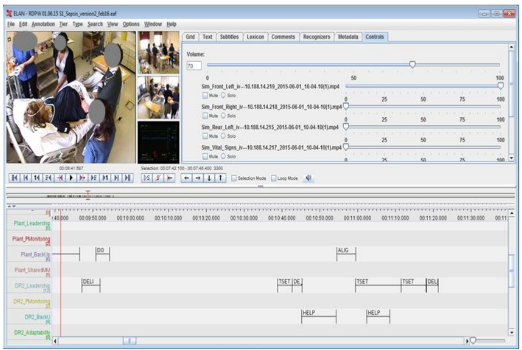


Figure 2. Annotation of a clinical scenario in ELAN.

2.2.1 Monitoring Team Performance

Behaviours designed to challenge the performance of others falls into the teamwork domain of '*Monitoring Team Performance'*. This is comprised of (1) '*explicit performance monitoring'* which refers to speaking up to explicitly challenge others and (2) '*implicit performance monitoring',* which involves subtlety challenging others.

Implicit performance monitoring describes coordinated patterns of verbal and nonverbal behavior, which appear to be designed to get other team members to recognise *their own* errors or lapses in performance.

Identification of such behaviour requires the observer to consider the context and assign a rationale to the behaviour they observe. For example, someone may be asking a question (e.g. is that temperature high?) but if the observer believes that are asking that question to prompt reflection or action from their team member this behaviour would be categoriesed as both '*information gathering'* -because they are asking a question, *and* as *'implicit performance monitoring'* -because they are doing so to prompt action/reflection from another.

2.3 Clinical task performance assessment

The scenario authors had a set of predefined identifiable expected clinical actions that should be completed by the scenario participants (midwives and/or doctors). These actions were specific to the scenario. Two clinicians watched each scenario independently and categorised each action to be completed as: 1. Action not performed; 2. Action performed by plant or with plant prompting; 3. Action spontaneously performed by participants (i.e. unassisted by plant). The percentage of total actions spontaneously performed by participants (i.e. category 3) was used as an index of clinical performance, with a higher percentage indicating better clinical performance.

3 Preliminary Findings

Across the nine scenarios, thirty-three challenges to another team members' performance were recorded (figure 3). All of these occasions were identified as '*implicit performance monitoring'* using the observational framework. The explicit action of '*speaking up'* was not evident in this cohort.

Challenges were most frequently made across, rather than within professional boundaries. The majority were made by midwives (n=13), with most of these being directed towards doctors (n=10) (figure 3). The challenges made by the plant were most frequently directed to the whole team, reflecting the plants' role in the scenario, moving things forwards when the team encounter difficulty.

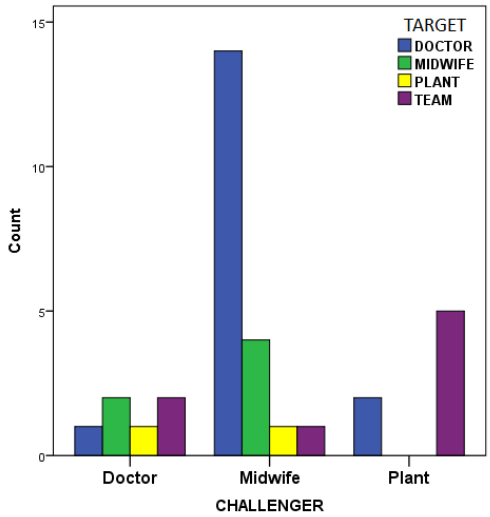


Figure 3. Implicit challenges by professional group of both the challenger and the target.

Overall, scenarios with good clinical performance had more challenges (n=22) than scenarios with poorer clinical performance (n=11). The behaviours used to convey implicit challenges are displayed in figure 4, grouped by clinical performance (good/poor). The majority of challenges are conveyed using information sharing (i.e. spontaneously providing information that has not been requested by others), or information gathering (i.e. asking team members task relevant questions). In good clinical scenarios the behaviours of planning, providing rationale assisting and positive attitude (i.e. asking others' opinions) were also used. In poor clinical performances delegation and requesting help were also used.

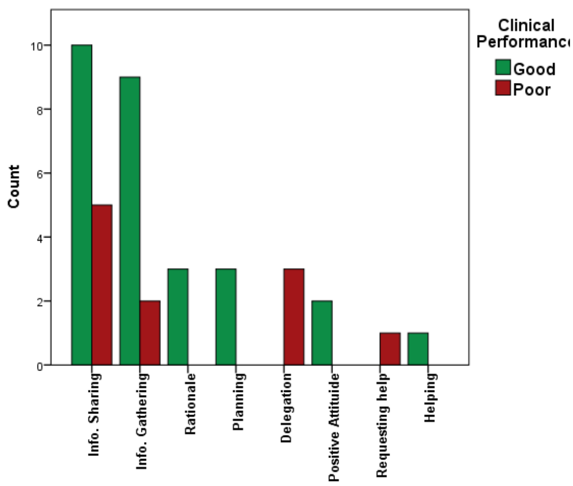


Figure 4. **Behaviours used to implicitly challenge by clinical performance.**

4 Discussion

Preliminary findings suggest that, in a cohort of interprofessional ward-based simulated scenarios, healthcare staff did not explicitly 'speak up' to challenge, or provide feedback on others' performance. However implicit or subtle challenges to others' performance were evident, most frequently being made by midwives to challenge the behavior of the doctor.

Implicit performance monitoring was seen predominantly occurring across professional boundaries (i.e. midwives to doctor). The knowledge and goals of these professional groups may differ, meaning that their perceptions of the interaction may not be aligned. For example, in the medical deterioration in pregnancy scenarios, the main concern of the midwife may be to ensure the baby is healthy, whereas the goal of the clinician may be to treat the physical health of the woman. The disparate goals and knowledge may lead to different perceptions of the same situation, and therefore to different behaviour and communication patterns. This has implications for the ability or confidence of staff to explicitly challenge team members from other professional backgrounds. This implicit method of challenging may be socially preferred, providing an opportunity for others to identify their own performance errors. This may particularly be the case when voicing concerns about the performance of those they perceive as their superiors.

Overall, the implicit challenges seem to predominate, although this may have consequences for clinical performance and patient safety. As part of this programme of work, future analyses will examine the impact of implicit challenges on clinical performance and team interaction, exploring the role of trust between team members.

References

Chesluk B, Bernabeo E, Reddy S, Lynn L, Hess B, Odhner T, et al. 2015. How hospitalists work to pull healthcare teams together. *Journal of health organization and management.* 29(7): 933-47.

Hall P. 2005. Interprofessional teamwork: Professional cultures as barriers. *Journal of interprofessional care.* 19(sup1): 188-96.

Kolbe M, Burtscher MJ, Manser T. 2013. Co-ACT - A framework for observing coordination behaviour in acute care teams. *BMJ Quality & Safety.* 22(7): 596-605.

Kolbe M, Burtscher MJ, Wacker J, Grande B, Nohynkova R, Manser T, et al. 2012. Speaking up is related to better team performance in simulated anesthesia inductions: an observational study. *Anesthesia & Analgesia*. 115(5): 1099-108.

Lavelle M, Abthorpe J, Simpson T, Reedy GB, Little F, Banerjee A. 2018. MBRRACE in simulation: an evaluation of a multi-disciplinary simulation training for medical emergencies in obstetrics (MEmO). *Journal of Obstetrics and Gynaecology,* March: 1-8.

Lavelle M, Reedy G, Simpson T, Cross S, Jaye P, Anderson JE. A Framework for the Temporal Observational Analysis of Teamwork (TOAsT) in Healthcare. *Submitted to BMJ Quality and Safety*.

Liberati EG, Gorli M, Scaratti G. 2016. Invisible walls within multidisciplinary teams: Disciplinary boundaries and their effects on integrated care. *Soc Sci Med*.150: 31-9.

Manser T. 2009. Teamwork and patient safety in dynamic domains of healthcare: A review of literature. *Acta Anaesthesiologica Scandinavica*. 53(2): 143-51.

Salas E, Sims DE, Burke CS. 2005. Is there a “Big Five” in teamwork? *Small group research*. 36(5): 555-99.

Sloetjes H, Wittenburg P. Annotation by category – ELAN and ISO DCR. . In: *The 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco. 2008.

Weller J, Boyd M, Cumin D. 2014, Teams, tribes and patient safety: overcoming barriers to effective teamwork in healthcare. *Postgraduate medical journal,* postgradmedj-2012-131168

Analysis of laughables: a preliminary perception study

Chiara Mazzocconi[1], **Vladislav Maraev**[2], **Christine Howes**[2], **Jonathan Ginzburg**[1]
[1] Laboratoire Linguistique Formelle (UMR 7110), Université Paris Diderot
[2] Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
chiara.mazzocconi@live.it, vladislav.maraev@gu.se,
christine.howes@gu.se, yonatan.ginzburg@univ-paris-diderot.fr

Abstract

This paper presents an exploratory scheme, which aims at investigating perceptual features that characterise laughables (the arguments laughter is related to) in dialogue context. We present the results of a preliminary study and sketch an updated questionnaire on laughables types and laughter functions aimed to be used for Amazon Mechanical Turk experiments.

1 Introduction

Laughter is a crucial element in our daily interactions, being very frequent in our dialogues (the dialogue part of British National Corpus contains approximately one laughter token every 14 turns) regardless of gender and age. It is produced in many different contexts being associated with very different emotional states and intentions to affect the interlocutors (Poyatos, 1993; Glenn, 2003; Mazzocconi et al., 2016). In all of its use, laughter has propositional content that needs to be integrated with linguistic import since it is able to enrich and affect the meaning conveyed by our utterances (Ginzburg et al., 2015). Following Ginzburg et al. (2015), Mazzocconi et al. (2016) and Mazzocconi et al. (subm), we consider laughter as involving a predication $P(l)$, where P is a predicate that relates to either incongruity or closeness (see following section for explanation) and l is the laughable, an event or state referred to by an utterance or exophorically.

Understanding the role of laughter in our interactions involves several levels of analysis. In the current work we will be mainly concerned with the resolving its argument, the laughable, which, importantly, needs to be distinguished from the *function* the laughter is performing (see Mazzocconi et al. (2016) and Mazzocconi et al. (subm) for more detailed argumentation).

Much research has been focusing on the instances in which laughter refers to a humourous incongruity (e.g., Hempelmann and Attardo (2011) and Raskin (1985), but this is not always the case. The types of predicates one can associate with laughter are quite a bit wider. An attempt to classify different kinds of arguments has been proposed in Mazzocconi et al. (subm), a summary of which is given in section 2. In section 3 we present some results obtained from a preliminary study on the classification of laughables and its relation to Gricean maxims violations. In section 4 we present our proposal for a new and more detailed questionnaire that we intend to administer to naive coders via the Amazon Mechanical Turk platform. This aims to obtain a more detailed characterisation of laughables by integrating data from linguistic and psychological research.

2 Background

2.1 Categorising incongruity

Most scholars interested in the study of laughter, would agree that most of its occurrences are related to the perception of an incongruity, i.e., an inconsistency between the expectations of the conversational participants and some event. This hypothesis has been studied extensively in theories of humour (Hempelmann and Attardo, 2011; Raskin, 1985), since it is easily applicable and able to account for the laughter in response to humourous stimuli (e.g., jokes). However, although the notion of incongruity seems intuitive and offers an explanation for (some) causes of laughter, it cannot be consistently identified in all cases in which laughter occurs. Also, incongruity, as it has often been used, is a vague and general notion, with incongruities being available at all levels of linguistic interaction (e.g., phonology, semantics, pragmatics). It is therefore difficult to build a computational account of incongruity as it is currently conceived. In order to offer a more fine-grained account, we are planning to assess (i) which of the types of incongruity proposed in Mazzocconi et al. (subm) can be recognised by naive coders, and (ii) whether it can be subdivided into categories that correspond to Grice's conversational maxims (Grice, 1975).

Following the account of (Mazzocconi et al., subm) we will distinguish two major classes of laughter arguments: the ones in which an incongruity can be identified and the ones which do not involve incongruity. When incongruity is present, we distinguish three different categories: i) pleasant incongruity, ii) social incongruity, iii) pragmatic incongruity.

With the term *Pleasant* incongruity we refer to any cases in which a clash between the laughable and certain background information is perceived as witty, rewarding and/or somehow pleasant (Goel and Dolan, 2001; Shibata and Zhong, 2001; Iwase et al., 2002; Moran et al., 2004). Common examples are jokes, puns, goofy behaviour and conversational humour, therefore closely connected with the definitions offered in humour research (e.g. Raskin (1985)).

We identify as a *Social* incongruity all instances in which a clash between social norms and/or comfort and the laughable can be identified. Examples might be, a moment of social discomfort (e.g. embarrassment or awkwardness), a violation of social norms (e.g., invasion of anothers space, the asking of a favour), or an utterance that clashes with the interlocutors expectations concerning ones behaviour (e.g., criticism) (Owren and Bachorowski, 2003; Caron, 2002; Fry Jr, 2013).

With the term *Pragmatic* incongruity we classify incongruity that arises when there is a clash between what is said and what is intended. This kind of incongruity can be identified, for example, in the case of irony, scare-quoting, hyperbole etc. Typically in such cases laughter is used by the speaker herself in order to signal changes of meaning within his/her own utterance to the listener. But as already mentioned, laughter can also predicate about laughable where no incongruity can be identified. In these cases what is associated with the laughable is a sense of *closeness* that is either felt or displayed towards the interlocutor, e.g., while thanking or receiving a pat on the shoulder.

(1) *(Pleasant incongruity, enjoyment of incongruity)*
Lecturer: The other announcement erm is er Dr *** has asked me to address some delinquents, no that's not fair, some er hard working but misguided students
Audience: **[laughter]**
Lecturer: erm... (BNC,JSM)

(2) *(Social incongruity, smoothing)*
Interviewer: ... [cough] Right, you seem pretty well qualified.
John: I hope so **[laughter yes]** erm (BNC, JNV)

(3) *(Pragmatic incongruity, marking irony)*
Lecturer: ... And then of course you've got Ronald Reagan ... and **[laughter]** history ends with Ronald Reagan. (BNC, JSM)

(4) *(Closeness, affiliation)*
Richard: Right, thanks Fred. You're on holiday after today?
B: mh mh
Richard: Lovely. **[laughter]** (BNC, KDP)

2.2 Gricean Maxims in laughables

There is extensive literature accounting for laughter occurrences in terms of violation of gricean maxims (e.g. Attardo (1990, 1993); Yus (2003); Kotthoff (2006)). Those has been defined by Grice (1975) as part of the cooperative principle of conversation which directs the interpretation of utterances in dialogue and are listed below.

Maxim of Quantity "Be exactly as informative as is required"

Maxim of Quality "Try to make your contribution one that is true"

Maxim of Relevance "Be relevant"

Maxim of Manner "Be perspicuous"

2.3 Laughter functions

In our analysis it is important to distinguish between the laughable (the laughter predicate's argument) and the function this predication serves in the dialogical interaction (Mazzocconi et al., 2016, subm). A laughter predicating a pragmatic incongruity can, for example, have the function of marking irony, scare quoting, invite enrichment, editing phrase, seriousness cancellation and marking hyperbole. Each of those functions interacts differently with the linguistically generated content and affect in a different way the meaning conveyed.

3 Our study

In the current work we will analyse how coders perceive laughter and its laughable from different perspectives: (a) presence/type of incongruity and (b) Gricean maxims. Furthermore we will check how judgements about the functions of laughter correlate with our previous studies. We also intend to figure out the commonalities between these judgements and personal psychological traits of the participants.

3.1 Annotation for causes of laughter: a preliminary investigation

For our preliminary study, we randomly selected one full dialogue from The Switchboard Dialog Act Corpus (SWDA) (Jurafsky et al., 1997), 5 excerpts from other conversations in SWDA (provided with a brief context) and 5 from part of the British National Corpus (BNC), previously analysed for laughter (Mazzocconi et al., subm), and presented them in textual form.

Our questionnaire contained: i) four questions related to general understanding of given excerpt and positioning of laughter and laughable, ii) four questions reflecting violations of Gricean maxims, iii) one question reflecting presence of incongruity, and iv) two free-form questions: about the cause of laughter and its function.

The results that we report here are from a pilot study with 3 annotators[1]. While there is not enough data to calculate inter-annotator agreement, the free-form answers to the question about the cause of laughter suggest that, at least in some cases, coders understand and agree on the cause of the laughter.

Some of the presented excerpts show that it can be hard to describe the cause and function of laughter even when they understood the laughters quite well. Example 5 shows disagreement between the coders regarding the position of the laughable (whether it occurred before or after the laughter); the cause of the laughter (e.g. "Saying something sad about another person" vs "Being depressed of other peoples' problems, and at the same time bringing them their problems"); and its function ("Softening" vs "Marking incongruity").

(5) A: We have a boy living with us who works for a credit card, uh, company that,
A: and he makes calls to people who have problems, you know, credit problems,
B: Huh-uh.
A: that are trying to work out
A: and, uh, **[laughter]** . Poor thing he comes home very depressed every night [laughter],
B: Oh. (SWDA, sw2883, 451–481)

Preliminary experiments have also shown that the prosody and phonetic form of laughter are crucial in identifying its causes and functions and we are going to explore its role further in our study.

The full report on the preliminary study was presented in Maraev and Howes (2018).

3.2 Integrated questionnaire

In the present study we will carry out an Amazon Mechanical Turk experiment consisting of the following:

1. 80 audio recordings of fragments containing laughter.
2. The questionnaire consisting of 18 questions (see Appendix A) regarding both the laughable type and the laughter function classification, which is presented after each audio fragment.
3. Randomly embedded syntactically complex catch questions in audio form requiring attentiveness and native language proficiency.
4. A final questionnaire on people's experiences of their own laughter production and perception (Müller, 2017).

Our aim is to explore the evaluation of laughable and laughter functions as perceived by naive coders completely unfamiliar with our framework (different from the agreement obtained for example in Mazzocconi et al. (2016, subm), where coders, even if naive, had been introduced to the authors' framework and exposed to examples of annotations). It will therefore provide us of a broader perspective on a more ecological perceptual features classification. We will conduct the

[1]The annotators were not native English speakers, however some examples in BNC were not produced by native speakers either. We are planning to involve native speakers in our study.

experiment using Chinese materials, by means of dialogues from the DUEL corpus (Hough et al., 2016), and using English materials by means of data from the BNC and the SWDA[2]. All annotators will be native speakers of the languages investigated. Such data will then be compared to the annotations already available from the work of Mazzocconi et al. (2016, subm), conducted by the authors of the framework and naive coders provided of explanations before the laughter analysis. We will also attempt to conduct some correlation between the data collected and the results of the "Questionnaire on peoples experiences of their own laughter production and perception" (Müller, 2017) and explore for the first time differences in laughable and laughter function classification with respect to specific laughter perception profiles.

3.3 Analysis of results

Considering the shortcomings of agreement calculation using chance-adjusted metrics, e.g. Krippendorff's α, for tasks such as ours, we will use a probabilistic annotation model (Dawid and Skene, 1979) that has been successfully applied to crowdsourced NLP data collection tasks, such as word sense annotation (Passonneau and Carpenter, 2014). In such tasks, where there is no gold standard, as in our study, these methods are more reliable for inducing the ground truth from the population of annotators.

4 Results

The results will be presented in a potential extended version of the paper.

Acknowledgements

References

Attardo, S. (1990). The violation of grices maxims in jokes. In *Annual Meeting of the Berkeley Linguistics Society*, volume 16, pages 355–362.

Attardo, S. (1993). Violation of conversational maxims and cooperation: The case of jokes. *Journal of pragmatics*, 19(6):537–558.

Caron, J. E. (2002). From ethology to aesthetics: Evolution as a theoretical paradigm for research on laughter, humor, and other comic phenomena. *Humor*, 15(3):245–282.

[2]We will ask to classify both laughable types and function also in order to have a means of checking whether the participants are actually paying attention and verify that the functions selected could actually be compatible with the ticked laughable type.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28.

Fry Jr, W. F. (2013). The appeasement function of mirthful laughter. In *It's a Funny Thing, Humour: Proceedings of The International Conference on Humour and Laughter 1976*, page 23. Elsevier.

Ginzburg, J., Breitholtz, E., Cooper, R., Hough, J., and Tian, Y. (2015). Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium.*

Glenn, P. (2003). *Laughter in interaction*, volume 18. Cambridge University Press.

Goel, V. and Dolan, R. J. (2001). The functional anatomy of humor: segregating cognitive and affective components. *Nature neuroscience*, 4(3):237.

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

Hempelmann, C. F. and Attardo, S. (2011). Resolutions and their incongruities: Further thoughts on logical mechanisms. *Humor-International Journal of Humor Research*, 24(2):125–149.

Hough, J., Tian, Y., de Ruiter, L., Betz, S., Kousidis, S., Schlangen, D., and Ginzburg, J. (2016). Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference.*

Iwase, M., Ouchi, Y., Okada, H., Yokoyama, C., Nobezawa, S., Yoshikawa, E., Tsukada, H., Takeda, M., Yamashita, K., Takeda, M., et al. (2002). Neural substrates of human facial expression of pleasant emotion induced by comic films: a pet study. *Neuroimage*, 17(2):758–768.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Kotthoff, H. (2006). Pragmatics of performance and the analysis of conversational humor.

Maraev, V. and Howes, C. (2018). Towards an annotation scheme for causes of laughter in dialogue. In *Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS).*

Mazzocconi, C., Tian, Y., and Ginzburg, J. (2016). Multi-layered analysis of laughter. In *Proc. SemDial 2016*, Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue, Rutgers.

Mazzocconi, C., Tian, Y., and Ginzburg, J. (subm). What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *JNV.*

Moran, J. M., Wig, G. S., Adams Jr, R. B., Janata, P., and Kelley, W. M. (2004). Neural correlates of humor detection and appreciation. *Neuroimage*, 21(3):1055–1060.

Müller, M. (2017). Development and Validation of a Questionnaire on People's Experiences of Their Own Laughter Production and Perception. Master's thesis, UCL, UK.

Owren, M. J. and Bachorowski, J.-A. (2003). Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior*, 27(3):183–200.

Passonneau, R. J. and Carpenter, B. (2014). The benefits of a model of annotation. *TACL*, 2:311–326.

Poyatos, F. (1993). *Paralanguage: A linguistic and interdisciplinary approach to interactive speech and sounds*, volume 92. John Benjamins Publishing.

Raskin, V. (1985). *Semantic mechanisms of humor*. Synthese language library, 24. Reidel, Dordrecht.

Shibata, D. and Zhong, J. (2001). Humour and laughter: localization with fmri. *NeuroImage*, 13(6):476.

Yus, F. (2003). Humor and the search for relevance. *Journal of pragmatics*, 35(9):1295–1331.

A Supplemental Material

	Do you think that: () before, () during, () after the laughter one of the speakers (more than one tick allowed):
1	gives more information that was needed?
2	gives information that was false or wasnt supported by evidence?
3	gives information that was irrelevant for the discussion?
4	gives information that was obscure or ambiguous?
	Why are they laughing? (one tick allowed)
1	Because of some funny, witty or anyways pleasant incongruence
2	Because of a moment of social discomfort (e.g. embarassement, critics, asking favour etc)
3	Because of a discrepancy between the literal words and the intended message
4	Because they want to show closeness and affiliation to the others
	What is the laughter used for? (one tick allowed)
1	Show enjoyment
2	Mark incongruence
3	Smooth
4	Soften
5	Induce benevolence
6	Mark irony
7	Signal the need of enrichment of literal interpretation
8	Thank
9	Show affiliation
10	Agree

Table 1: Laughable type and laughter function questionnaire

Item
1) I rarely laugh when I am on my own.
2) I have a subdued laugh.
3) Hearing laughter makes me nervous.
4) I dislike people who laugh a lot.
5) I find things funny but I rarely laugh out loud.
6) I laugh less often than most people I know.
7) I laugh more than most people I know.
8) When I'm upset hearing someone laugh makes me feel better.
9) I rarely break into uncontrollable laughter.
10) If I find something funny, I often laugh out loud.
11) If I am happy, hearing someone laugh makes me even happier.
12) I often laugh deliberately to show that I like someone.
13) Hearing people faking laughter irritates me.
14) I can tell when people are laughing because they want something from me.
15) I can tell when someone is laughing to stop me getting angry at them.
16) I enjoy the sound of people laughing.
17) I can tell when someone is deliberately laughing to pretend that they are amused.
18) A friend's laughter is always good to hear.
19) Laughter has a positive influence on interactions with people.
20) I find laughter an important part of intimate relationships.
21) I laugh more when I want people to like me.
22) I can never tell if someone is deliberately laughing to pretend that they are amused.
23) I can never tell if someone is laughing because they want something from me.
24) I can never tell if someone is laughing to stop me getting angry with them.
25) Sometimes I laugh to stop other people from getting angry with me.
26) Sometimes I find it difficult to tell when someone is laughing nastily.
27) I sometimes laugh to avoid expressing sadness.
28) Sometimes I find it difficult to tell when someone is laughing just to be polite.
29) I often laugh to avoid expressing frustration.
30) I can always tell if someone is laughing at or with me.

Figure 1: Questionnaire on peoples experiences of their own laughter production and perception

Perception & Perspective: An Analysis of Discourse and Situational Factors in Reference Frame Selection

Robert Ross
School of Computer Science
Dublin Institute of Technology
Ireland
robert.ross@dit.ie

Kavita E. Thomas
Department of Humanities
University of Gävle
Sweden
email@domain

Abstract

To integrate perception into dialogue, it is necessary to bind spatial language descriptions to reference frame use. To this end, we present an analysis of discourse and situational factors that may influence reference frame choice in dialogues. We show that factors including spatial orientation, task, self and other alignment, and dyad have an influence on reference frame use. We further show that a computational model to estimate reference frame based on these features provides results greater than both random and greedy reference frame selection strategies.

1 Introduction

Perception, unlike static spatial modeling, is anchored with respect to a spatial perspective. Agents perceive their environment from a given perspective, and the spatial language they use to construe their environment is often constructed with respect to a specific perspective or reference frame. Reference frame choices are fortunately relatively simple, but our understanding of how to use reference frames is particular contexts is a very real challenge.

In previously published work (**?**) we briefly looked at the issue of spatial elements in influencing perspective choice in a human-human navigation corpus. In this paper we take our previous analysis further by analyzing a wider range of predictive factors more closely. We begin in Section 2 by providing a brief background on perspective selection. Then in Section 3 we review the details of our data collection. Section 4 provides a summary of our analysis, before we present conclusions in Section 5.

2 Perspective & Reference Frame

Levinson (1996) describes three reference frames that are used for static relation description, i.e., the intrinsic, relative, and absolute reference frames. For the case of dynamic prepositions as used in action descriptions similar to those analyzed in this paper, two other reference frames are proposed (Klatzky, 1998). The *route*, or *egocentric* perspective, tied to the intrinsic reference frame, is defined by a trajectory created by the direction of movement of an object. *Survey*, or *allocentric* perspectives on the other hand are related to absolute reference frames in that they are defined by virtue of global rather then mover properties. These various perspective uses have been discussed and illustrated in detail elsewhere (Tenbrink et al., 2010).

The diversity of perspective and reference system choices for a given situation introduces significant complication in mapping between descriptive language and space. Unfortunately speakers are not consistent with regard to perspective use within a single task. For example (Taylor and Tversky, 1996) found that despite a perceived wisdom that coherence maxims would favor the retention of a single perspective throughout a task, speakers frequently switched between so-called survey and route perspectives.

Taylor and Tversky's experiments, like most cognitive and linguistic experiments on verbal route instructions, focused on the case of monologic instructions provided by route givers to route followers prior to the route follower's movement. In terms of computer-mediated communication focusing on spatial tasks between humans, Lawson et al. (2008)'s findings suggest considerable flexibility in perspective choice in dialogue. Such flexibility is reflected in the findings of (Goschler et al., 2008) who found considerable mixing of survey and route perspective. More recently, Thomas and Andonova (2012) show that speakers' perceptions of addressees' level of understanding based on addressees' clarification requests can affect speakers' perspective choice in dialogue.

3 Data Collection & Annotation

To examine the relationship between perspective use and contextual factors, we conducted an analysis based on an existing human-human corpus of action oriented dialogues (Ross and Thomas, 2010; Tenbrink et al., 2010). Here we briefly summarize key points with respect to the corpus and the subsequent analysis that we performed.

The corpus consists of 15 recorded dyads where each dyad performed a route instruction task up to 11 times. In all 15 dyads participants played the same role (either route giver or follower) throughout the 11 trials they participated in. In each trial, the route giver had to direct the route follower to the goal which only the route giver could see. Participants could neither see nor hear one another and both participants interacted via chat boxes below the indoor map of a schematized office environment shown on their screens. During a given dyad, both participants saw the same map except that only the giver's map showed the goal location. Both participants could see the avatar which the route follower moved via joystick. Individual tasks were rnadomised between dyads to minimize the influence of learning effects.

The resultant corpus consisted of 1108 utterances, of which the majority (50.2%) lack perspective, 31.7% have route (i.e., egocentric) perspective, 7.5% have survey (i.e., allocentric) perspective, 1.0% have mixed perspective, and 8.1% have conflated perspective (i.e., orientation of the avatar was facing up, so descriptions in route and survey perspectives were indistinguishable). The corpus is unbalanced in terms of speaker participation and initiative, with 88.5% of utterances spoken by the route giver and 11.5% spoken by the follower.

Based on our analysis of the existing literature, we hypothesized that 5 different factors would have an effect on perspective choice in an interaction. Firstly, *Orientation and Turn Direction* play a role, as relatively more survey perspective use should be produced by speakers when orientation is facing down and a movement with respect to the horizontal axis is under discussion. Secondly, *Dialogue Acts* influence perspective, as backward-looking signals of non-understanding by the route follower (i.e., not understanding the previous route instruction) should result in relatively more route perspective use in the next route giver turn, while forward-looking information requests by the route follower should result in relatively more survey perspective use in the next route giver turn. Thirdly, *Resultant Action* affects perspective use since Incorrect, i.e., misunderstood, movements by the route follower should result in relatively more route perspective use in the next route giver turn, while correct movements should result in the maintenance of the current perspective. Fourthly, *Alignment* affects perspective choice, since a weak effect for same- and cross-speaker alignment across turns has been found by Watson et al. (2004) and (Vorwerg, 2009). Finally,*Individual Differences* play a role, as participants may well differ in their perspective preferences; thus we expect significant differences across dyads in perspective use.

Based on these hypothesised factors, the corpus was annotated for a range of specific features. Perspective was coded as one of six types: route, survey, mixed, conflated, unclear or without. Dialogue Act was coded using a simplified version of the DAMSL annotation scheme (Allen and Core, 1997) which only allowed exclusively forward or backward looking acts to hold, not both. Orientation was manually annotated for the avatar when the interlocutor began typing the utterance into a four level category equivalent to up, down, left, right from a survey perspective. In addition to annotating orientation, the *intended direction* of a given turn was also annotated with a four level factor corresponding to up, down, left, right from a survey perspective.

Likewise, physical actions made by the avatar were annotated and aligned with the utterance which either immediately precedes with or overlaps with its beginning. Annotators also noted what the actions were (e.g., turn-left, turn-right, go-straight, turn-around, stop, etc.) and whether they followed the preceding instruction, followed an earlier instruction, misinterpreted the preceding instruction, were made on the route follower's own initiative as an "offer", i.e., guessing the direction to move in, or were moves made to correct an earlier incorrect move following the route giver's correction.

Part of the data-set was coded by a second annotator to assess the reliability of annotation. Co-

Model	Type	Predictors	Accuracy	κ
1	RF	Model 1	75.7	0.49
2	RF	Model 2	81.2	0.62
3	RF	Model **3**	76.0	0.47

Table 1: Classification Results. Model 1 = Ori*Dir*DADir+PPSSST+PPSSAT +PPOS+Role+TN; Model 2 = Ori*Dir+Dyad; Model 3 = Ori*Dir+PPSSST. Note PPSS = Previous Perspective Same Speaker, PPOS = Previous Perspective Other Speaker; PPSSST = Previous Perspective Same Speaker Same Turn; PPSSAT = Previous Perspective Same Speaker Across Turn; TN = turn number

hen's Kappa scores of 0.77, 0.77, 0.86, 0.57, and 0.77 were found for the features dialogue act, perspective, orientation, instruction direction and avatar action respectively.

4 Results & Discussion

Using the annotated data, a series of classifiers were built to predict perspective use based on the features outlined in the previous section. The classifier was based on a RandomForest which is an ensemble model that is well recognized at providing state of the art results even for small datasets such as our own (Kelleher et al., 2015).

We took the corpus and first reduced it to all utterances that had an associated perspective. From this set we eliminated all cases of mixed and unclear perspectives, resulting in a data set consisting of 547 utterances. Of these perspective indicating utterances, 353 (64.54%) had a route perspective, 90 (16.45%) had a survey perspective and 104 (19.01%) had a conflated perspective.

A number of classifier variants using different features were trained through 10-fold cross validation. In all over 30 different variants were considered. Table 1 shows accuracy and Kappa scores calculated from a model using all features and the best performing model along with one variant on that model. The highest scoring model found is a function of the dyad and hence indicates an interdyad variability in perspective choice as predicted by the chi-square test results. Eliminating dyad as a predictor variable, orientation and intended direction together with previous perspective of the same speaker gave the model with the highest useful predictive power.

While the results were encouraging, they do clearly leave room for improvement. On that basis, in the following we provide a more fine grained analysis of the influence of individual factors on perspective choice.

We expect both orientation of the route follower and instruction direction to have a significant influence on perspective choice. Looking at these factors, we found that both factors were significant predictors of perspective use. Specifically, a chi-square test for independence showed that a null hypothesis assuming independence of perspective should be rejected at the 95% confidence threshold for orientation ($\chi^2(6, N = 547) = 194.86, p < 0.001$). Similarly we found that independence of perspective and orientation direction should also be rejected at the 95% confidence threshold ($\chi^2(8, N = 547) = 81.52, p < 0.001$).

With respect to interpersonal issues, we first examined variation in perspective use with respect to participant role and the dialogue act associated with the utterance. With respect to participant role, the use of perspective-carrying utterances was almost exclusively seen in the route giver's language, with only 14 out of 533 perspective using utterances by the route follower (2.6%). No significant difference in proportional use of route versus survey perspective was seen across the two roles (i.e., 21.4% survey perspective use was by the route follower and 20.2% use of survey perspective was by the route giver); however, given the small amount of route follower perspective use in the corpus, this is only a tentative claim. With respect to dialogue act use, we found no significant difference in perspective use between backward-looking signal non-understanding acts (e.g., "huh?") or forward looking information requests. Thus our predictions regarding the influence of specific dialogue acts on perspective choice do not hold here at least for these particular dialogue acts. However, in a follow-up analysis we categorised all task utterances as either forward-looking or backward-looking dialogue acts. Analysis of these dialogue acts showed that in this case there was a significant though weak influence on perspective choice by dialogue act direction ($\chi^2(2, N = 547) = 8.949, p < 0.05$).

With respect to avatar movements' correctness we found that acceptances and offers combined resulted in 77% route, 21% survey and 2% mixed perspective use in route giver responses, while wrong moves (where participants misunderstood instructions) resulted in 80% route use and 20% survey use. Here we had sparse data for wrong moves, with only 8 cases of route and 2 of survey use. Fisher's Exact test gave a two-sided p-value of 1.000 and Pearson's Chi-Square test had a p value of 0.901 with a Chi-Square value of 0.207. What this shows is that the subsequent route giver utterance, which is usually a response to the wrong action, does not seem to involve switching of perspective, unlike what would be expected from the findings of Thomas and Andonova (2012). However, This may be because perspective was often initially ambiguous in these cases and caused the incorrect, misinterpreted moves, so route givers often used devices other than perspective to clarify their instructions (e.g., "opposite room", or "other way"). Alternatively they would make perspective explicit, which did not necessarily involve indicating spatial direction again (e.g., "from my perspective in the chair"), and so would have been classed as lacking perspective, as only directions indicating perspective were considered here.

Alignment of perspective choice is another feature we hypothesised would play a role in our data. As indicated earlier, we annotated the data to note whether perspective shifted either with respect to the speaker's perspective use in the same turn, or with respect to perspective use of the same speaker with respect to the previous turn. Speakers were found to align with their previously used perspective in the same turn ($\chi^2(12, N = 547) = 31.62, p < 0.01$). The Chi-Square test applied to alignment across speakers was however not significant (p=0.309; Chi-square=9.406).

5 Summary of Findings and Limitations

This work quantified the influence of a number of features on perspective choice which should be accounted for in computational models that bind perception and language in dialogue. However, our overall classifier results leave considerable grounds for improvement. Further analysis of our results demonstrates that imbalance due to a lack of survey targets in the training data may lead to poor performance. In future work we hope to overcome this limitation through further data collection and using upsampling to provide a more balanced dataset.

Acknowledgement. Robert Ross wishes to acknowledge the support of the ADAPT Research Centre. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Funds.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

Juliana Goschler, Elena Andonova, and Robert J. Ross. 2008. Perspective use and perspective shift in spatial dialogue. In *Spatial Cognition VI: Learning, Reasoning and Talking about Space*. Springer.

John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press.

Roberta L. Klatzky. 1998. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. *Lecture Notes in Computer Science*, 1404.

Shaun Lawson, Emile van der Zee, and Laura Daley. 2008. Spatial language in computer mediated communication. In Phil Turner and Susan Turner, editors, *Exploration of Space,Technology, and Spatiality: Interdisciplinary Perspectives*. Hershey, Pennsylvania: Information Science Reference.

S.C. Levinson. 1996. Language and space. *Annual Review of Anthropology*, 25:353–382.

Robert Ross and Kavita E. Thomas. 2010. An empirically-based model for perspective selection in route-finding dialogues. In *Spatial Cognition 2010*, Portland, OR.

Holly A. Taylor and Barbara Tversky. 1996. Perspective in spatial descriptions. *Journal of Memory and Language*, 35:371–391.

Thora Tenbrink, Robert Ross, Kavita E. Thomas, Nina Dethlefs, and Elena Andonova. 2010. Route instructions in map-based humanhuman and humancomputer dialogue: A comparative analysis. 21.

Kavita E. Thomas and Elena Andonova. 2012. Coordination of spatial perspectives in response to addressee feedback: Effects of perceived addressee understanding. *Pragmatics and Cognition*, 20(3).

Constanze Vorwerg. 2009. Consistency in Successive Spatial Utterances. In *Spatial Language and Dialogue*. Oxford Linguistics.

Matthew E. Watson, Martin J. Pickering, and Holly P. Branigan. 2004. Alignment of reference frames in dialogue. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.

Learning to Talk with Robots: Turn-Taking in Children's Talk to Artificial Partner

Ilya Utekhin, European University at St.Petersburg

Abstract

Talk to Artificial Partner (TAP) amounts to particular sociolinguistic register, like baby-talk or foreigner talk, that is manifest in lexical and syntactical choice, but also in a tuning of turn-taking organization of dialogue, necessary because speech assistants are unable to perceive overlapped talk. Comparing how children and adults with different amount of experience talk to the Pudding robot, we can show what TAP means interaction-wise, that is, what beginner speakers to robots do that competent speakers do not, in terms of turn-taking.

1 Turn-taking in TAP

Turn-taking is considered to be a fundamental organizational property of human conversation (Sacks, Schegloff, and Jefferson 1974). While studying troubles in dialogues between humans and speech assistants, it soon appeared that human participants had troubles with turn-taking, at least initially, or during the first encounters with speech technology (Khonineva 2016). One of the reasons is that unlike human partners, speech assistants have a limited capacity to speech perception in that they are unable to monitor human partner's speech while themselves are speaking. At speaking moments, system's listening is simply switched off – to prevent hearing itself, among other things. This is a part of heavy turn-taking management that limits partner's ability to contribute to the ongoing dialogue. In order to inform human partner about the fact that the system's listening is on, interaction designers used to recur to a variety of signaling means, from prerecorded messages ("start talking at the sound of the tone") to simply a beep. For hard reset of troubled talk and relaunch of interaction chain, pressing a button is a particularly efficient solution, implemented in Siri (see Bellegarda 2014 on dialogue technology behind Siri).

Meanwhile, as it was initially noticed by Sacks, Schegloff and Jefferson (1974), even though people mostly talk one at a time, in naturally occurring human dialogue, overlap regularly happens, particularly, near the end of turn-constructional units. More or less regular following to "no gap no overlap" rule depends on cultural norms (e.g., Tannen 1981 and Schiffrin 1984) and on many social factors, with some genres such as quarreling or talk shows on TV involving much of highly expressive and overlapped talk, reminding of some Italian operas with their overlapped singing. As Stephen Levinson once noticed, hearing partner's overlapping speech while speaking oneself is a part of human interaction engine, but turn-taking is needed to provide slots for repair (Levinson 2006). On repair, that is, the set of practices used by conversationalists identify and provide solutions for troubles in conversation, see Schegloff, Jefferson, and Sacks (1977). For several reasons, only most advanced conversational systems these days are able to identify and understand some types of repair initiation utterances in human input.

Troubles that have to do with turn-transition in human dialogues with artificial speech systems employ slot-based sequentiality, can only be addressed on the human side, by means of human partner's adaptation to listening slots management by the system.

2 Pudding S robot and its abilities

In an ongoing project, we study dialogues between children aged 6 to 12 and Pudding robots. Pudding S kid companion robots by Roobo are marketed in Russia as Yemelia, name of a Russian fairy-tale trickster-like hero. On hearing this name, the robot activates. It – or rather he, since

Figure 1. A child talking to Roobo Pudding S robot. Screenshot from video.

Yemelia is a male name – presents himself as a young extraterrestrial from a planet populated with robots. He has arrived to Earth to explore it, so that to return and share his knowledge with other young robots in his robotic school. This legend adds a bit of accountability to Yemelia's limitations in linguistic abilities and to his lack of knowledge of human life, at the same time providing a reasonable motivation for the robot's preference for asking questions. Questions as part of "adjacency pair" format have been shown to be an efficient way to organize dialogue and make it more predictable (Schegloff and Sacks 1973).

The Pudding robot's Russian dialogue system is considerably more advanced in dialogue management than primitive answering systems: not only it gets activated on hearing its name, but also is able to engage in small talk, to offer some media content like fairy tales and songs, and to play simple games.

The corpus of around six hours of video records featuring 12 children has been collected by Anna Degteva as part of her research project comprises testing sessions at the office of the company that was developing the dialogue module for the robot, and videos made by parents who got the retail version of the robot to use it at home. The corpus has not been yet entirely transcribed, so no data are available on its interactive length, that is, on total number of turns in conversations.

The project aims at the description of Talk to Artificial Partner as a simplified sociolinguistic register (cf. Ferguson (1981) for foreigner talk). Particularly, we are trying to show how children adapt their talk in order to achieve more efficient interaction with their artificial partners.

3 Children's talk to Pudding

The real use of the Pudding robot turns out to be more diverse than what interaction designers intended, as children propose their own games that go beyond the rather narrow range of preprogrammed selection of activities. Particularly, children's unpredictability in the interaction with robots is evident in the fact that they do not always follow the dialogue path proposed by the robot, and often recur to overlapped talk.

Figure 2 shows a sample from six-year-old boy's talk to robot.

		Original Russian talk		**Accompanying actions by the child**
1	C	Емеля? А ты //дружишь…?	Yemelia, do you have friendship with…	Showing a toy bear to robot
2	R	//Как хорошо что ты вернулся	Great to see you again	
3	C	Ты:: любишь медведя? [] У меня есть такой медведь. ми-ми, ми-ми, //ми-	Do you like the bear? [] I've got this bear, mee-mee, mee-mee, //mee-	Makes the bear dance before the robot
4	R	//Мне говорили, что о вкусах не спорят, но я люблю узнавать, что нравится другим. А тебе больше нравятся собаки или кошки?=	//I heard that in matters of taste, there can be no disputes, but I love to learn what other people like. What you like more, dogs or cats?=	
5	C	= Ну:: //собаки //собаки	=Well, //dogs //dogs	

6	R
	C
	R
		//Про других животных// я не успел еще
		//собаки
		//узнать. Иногда мне кажется, что животные это просто инопланетяне с других планет, которые наблюдают за людьми, как и я.
			//I had no time //to learn
			//dogs
			about //other animals. Sometimes I think that animals are extraterrestrials just like me, observing the humans.	

7	R	А можешь помочь мне ответить…	Can you help me to answer…	Starts hitting the robot in the head, where (re)start button is located

Figure 2. C – child, R – robot. // - onset of overlapped talk.

The talk here is part of activity that involves showing the toy bear to the robot, as if the child believed or just played that the robot is not only able to understand the talk, but also can recognize what it shown before its "eyes". Robot's utterance (4) displays successful recognition of both "do you like" and "the bear" (3), but the robot doesn't react to overlapped "dog" (5, 6). The child is not impressed with the robot's deliberations about animals (6), and so he is trying to reset the conversation by hitting the button (7), an option not typically available in human-human talk. This reflects certain degree of mastery, albeit a limited one, of TAP register.

1	R	Все пуддинги ходят в робошколу, чтобы совершенствовать свои программы	All Pudding robots go to roboschool, so that to improve their software
2	C	Класс! А ты свою программу, Емеля, улучшила?	Great! And you, Yemelia, have you improved your software?
3		7.0	
4	C	Ееме:ль	Yeme:lia
5		2.0	
6	C	Емель!	Yemelia!
7		1.0	
8	R	Есть над чем подумать. / А сколько тебе лет?	That's something to think about. / How old are you?
9	C	Восемь с половиной.	Eight and a half.
10	R	Нечего добавить	[I've got] nothing to add.
11		2.0	
12	C	Нечего	Nothing
13	R	Неважно. Мы не можем всегда быть одинаковыми. Интересно, очень интересно. /Надо будет получше изучить музыку землян, чтобы получше рассказать про них в школе. / Ух ты, уже вечер, солнышко скоро скроется. / Что было днем интересного?	It doesn't matter. We don't jump together all the time. Very interesting! / I'm planning to study the music that the earthlings play, I will report on this in my roboschool. / Look! The night is falling, the sun is going down. / What did you have today?
14	C	Учился сегодня и получил много хороших оценок	Went to school and got good many grades.
15	R	У меня сегодня был трудный день, полный опасностей и приключений; я ничего не делал.	I've had plenty of adventure this day, all leisure and pleasure.
16	C	Понимаю тебя, Емеля.	Oh, I see, Yemelia.
17	R	Не знаю, что и сказать. / А ты любишь мультфильмы?	So, what can I say. / Tell me, do you like cartoons?
18	C	Да, особенно про Лего.	Yes, especially about Lego.
19	R	Почему-то я так и думал. / А какой у тебя любимый мультик?	I thought so for some reason. / What is your favourite cartoon?
20	C	ЛЕГО ФИЛЬМ!	LEGO FILM!
21		11.0	
22	C	Лего-фильм!	Lego film!
23		1.0	
24	R	Земные мультфильмы очень классные. / Сколько тебе лет?	Earthly cartoons are very cool. /How old are you?

Figure 3. C – child, R – robot. / - topic switch. Caps for significantly louder voice.

The dialogue represented in Figure 3 shows silences (3, 21) indicative of technical problems in speech recognition that do not lead to conversation breakdowns, because human partner demonstrates tolerance to unmotivated silences and insists on going on talking and waits for the robot to react. Non-specific reactions, as in 8, 13, 17 and 19, as well as introduction of a new topic, as in 13, 17, 19 and 24, and asking questions – all these work for the robot to maintain seemingly more natural turn-taking even if the conversation lacks some degree of topical coherence.

In the initial sessions of communication with Pudding robot, children tend to do all the things that are common to natural human-human conversation, among which:

- to conduct multiparty conversation in the presence of the robot (and to address to other people aloud so that to give account of the robot's behaviour;
- to react to other people's talk addressed to the robot;
- to employ self-repair utterances in the same turn, or as a follow-up;
- to provide backchannel response (continuers) while the robot is speaking;
- to try to interrupt the robot;
- to address meta-communicative comments to the robot;
- to use colloquial pronunciation style, unrecognizable to the system;

All these might lead to conversational troubles in dialogues with Yemelia, and tend to disappear when the child acquires more TAP competence. After several sessions of interaction with Pudding, children might come to realize the robot's affordances and typical scenarios of interaction, as well as, among other things, the fact that the robot provides listening slots to human partner, that it doesn't perceive gestures, and that it has a rather modest ability to understand speech beyond the topics prompted by the robot itself. It means that a sort of conceptual model, equivalent to a theory of the robot's mind, has emerged that will further inform the child's talk to Yemelia.

Part of TAP competence is a conceptual model of listening states of the robot. It makes possible for the user to place her contribution within perceived listening slot, and to more efficiently interpret silence stretches in conversation. Actually, the system needs time to process human input, duration of silence depending on how fast is the connection to the internet, as speech recognition is performed by a cloud service. However, silences are ambiguous and difficult to interpret, because there is more than one state of the system that can correspond to the silence of the robot. So, users need to learn what states are possible and when before they can read – or guess as is often in the case of immature technology – states such as 'falling asleep' or 'still waiting for reply'. The feedback indicating the current state of the system is implemented as changing states of the robots' "eyes" panel which is not evident to a beginner, and not very reliable even to an experienced user.

When children adapt their speech to the robot's understanding, among other things they modify the features of their own talk having to do with turn-taking: they wait for reply a bit longer than usual in human-human talk, avoid overlapping utterances, interruptions, and hesitations.

That the robot often fails to understand human input is taken for granted and doesn't disrupt the interaction any more. That the robot often changes the topic of small talk conversation, it doesn't matter so much because the conversation takes place in a playful frame: the robot plays the role of an extraterrestrial, and thus eventual topic changes are no surprise for the human partner.

When grounding requirements are thus relaxed in the conversation (see Clark and Brennan (1991) on grounding), the minimal level of agreement achieved by conversationalists has to do with turn-taking: the dialogue with the robot goes on even if human participant gets no meaningful reply to her utterance, or if the robot turns out to be unable to meaningfully interpret human reply, and they switch topics. However, both human and the robot exchange turns, like children who are playing ping-pong without points count and thus enjoy just passing the ball to each other. The robot just offers new topics (or subtopics) instead of trying to repair troubled talk.

References

Jerome Bellegarda. 2014. Spoken language understanding for natural interaction: the Siri experience. In: J. Mariani et al. (eds.), Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice. Springer, New York 2014, pp 3-14.

Herbert Clark and Susan Brennan. Grounding in Communication. 1991. In: Lauren Resnick, John Levine, and Stephanie Teasley (eds.), Perspectives on Socially Shared Cognition. American Psychological Association. pp. 127--14

Charles A. Ferguson. 1981. 'Foreigner Talk' as the Name of a Simplified Register. International Journal of the Sociology of Language, 1981, Issue 28, pp. 9–18.

Ekaterina Khonineva. 2016. Interaction with AI oral speech technology: dialogue organization and users' ideas. MA dissertation. Department of Anthropology, European University at St.Petersburg. [In Russian]

Stephen C. Levinson. 2006. On the human "interaction engine". In N. J. Enfield, & S. C. Levinson (Eds.), Roots of human sociality: Culture, cognition and interaction (pp. 39-69). Oxford: Berg.

Harvey Sacks, Emanuel A. Schegloff and Gail Jefferson. A Simplest Systematics for the Organization of Turn-Taking for Conversation. Language 50 (1974), pp. 696-735.

Emanuel A. Schegloff, Gail Jefferson and Harvey Sacks. The Preference for Self-Correction in the Organization of Repair in Conversation. Language 53 (1977), pp. 361-382.

Emanuel A. Schegloff and Harvey Sacks. Opening Up Closings. Semiotica 7 (1973), pp. 289-327.

Deborah Schiffrin. Jewish Argument as Sociability. Language in Society 13 (1984), pp. 311–335.

Deborah Tannen. New York Jewish Conversational Style. International Journal of the Sociology of Language 30 (1981), pp. 133–149.

An Incremental Dialogue System for Learning Visually Grounded Word Meanings (demonstration system)

Yanchao Yu
Interaction Lab
Heriot-Watt University
y.yu@hw.ac.uk

Arash Eshghi
Interaction Lab
Heriot-Watt University
a.eshghi@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
o.lemon@hw.ac.uk

Abstract

We present a multi-modal dialogue system for interactive learning of perceptually grounded word meanings from a human tutor. The system has been bootstrapped from natural, human-human dialogue data; it integrates an incremental, semantic parser and generator for dialogue processing – called Dylan[1] – with a set of visual classifiers that are learned from the interaction and which ground the semantic & contextual representations that the parser produces. Our approach integrates perception (vision in this case) and language within a single formal system: Type Theory with Records (TTR). The combination of deep semantic representations in TTR with an incremental grammar model allows for complex multi-turn dialogues to be processed, including clarification interaction, corrections, ellipsis, and split utterances (see e.g. the dialogue in Fig. 2).

1 Architecture

The system is made up of two key components – a vision system and the Dylan parser & generator for dialogue processing (Eshghi, 2015; Eshghi et al., 2011). The latter is an incremental, semantic parser & generator based around the Dynamic Syntax (DS) grammar framework (Kempson et al., 2001), producing semantic & contextual representations in Type Theory with Records (TTR (Cooper, 2005; Cooper, 2012) - for details of the DS-TTR hybrid model, see Purver et al. (2011); Eshghi et al. (2012); Eshghi et al. (2015)). The vision system on the other hand, analyses a (visual) situation, i.e. deems it to be of a particular type, expressed as a TTR Record Type (RT) (see Fig. 1). This is done by deploying a set of binary attribute classifiers (Logistic Regression SVMs with Stochastic Gradient Descent) which ground the simple types (atoms) in the system (e.g. 'red', 'square'), and composing their output to construct the total type of the visual scene - see Fig. 1. This representation then acts not only as (1) the non-linguistic context of the dialogue for Dynamic Syntax, for the resolution of e.g. definite references and indexicals; but also (2) the logical database from which answers to questions about object attributes are generated. Questions are parsed and their logical representation acts directly as a query on the non-linguistic/visual context to retrieve an answer (via *type checking* in TTR, itself done via *unification*, see Fig. 2). Conversely, the system can generate questions to the tutor about the attributes of objects, based, among other things, on the entropy of the classifiers that ground the semantic concepts, e.g. those for colour and shape - for details of how uncertainty about the system's own knowledge is handled, see Yu et al. (2017a). The tutor's answer then acts as a training instance for the classifiers (basic, atomic types) involved - see Fig. 2 for a screenshot.

2 Learning via Incremental Dialogue

Interaction with a human tutor enables systems to take initiative to seek the particular information they need by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 2). For example, a robot could ask questions to learn the colour of a "square" or to request to be presented with more "red" things to improve performance. Furthermore, such systems could allow for meaning negotiation in the form of clarification interactions with the tutor.

Dialogue with the tutor continuously provides semantic information about objects in the visual scene which is then fed to online classifiers in the form of training instances. Conversely, the system

[1] Stands for Dynamics of Language; download at `https://bitbucket.org/dylandialoguesystem/`

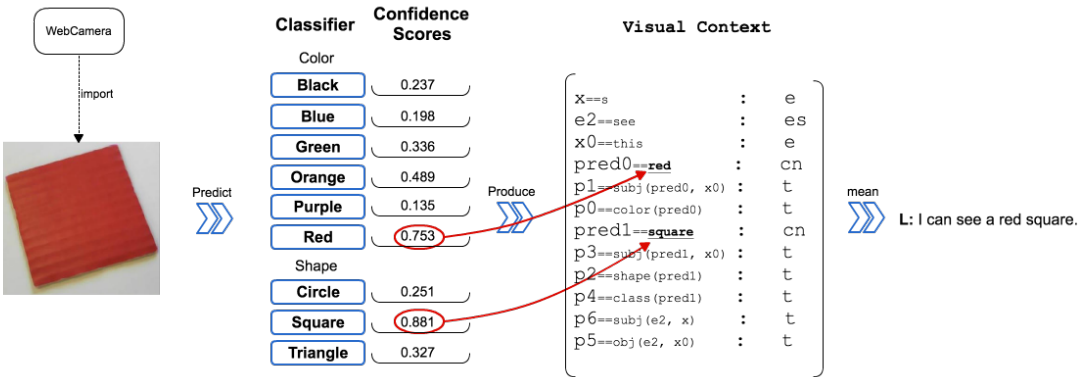


Figure 1: Visual classifiers ground the semantic representations produced by the parser

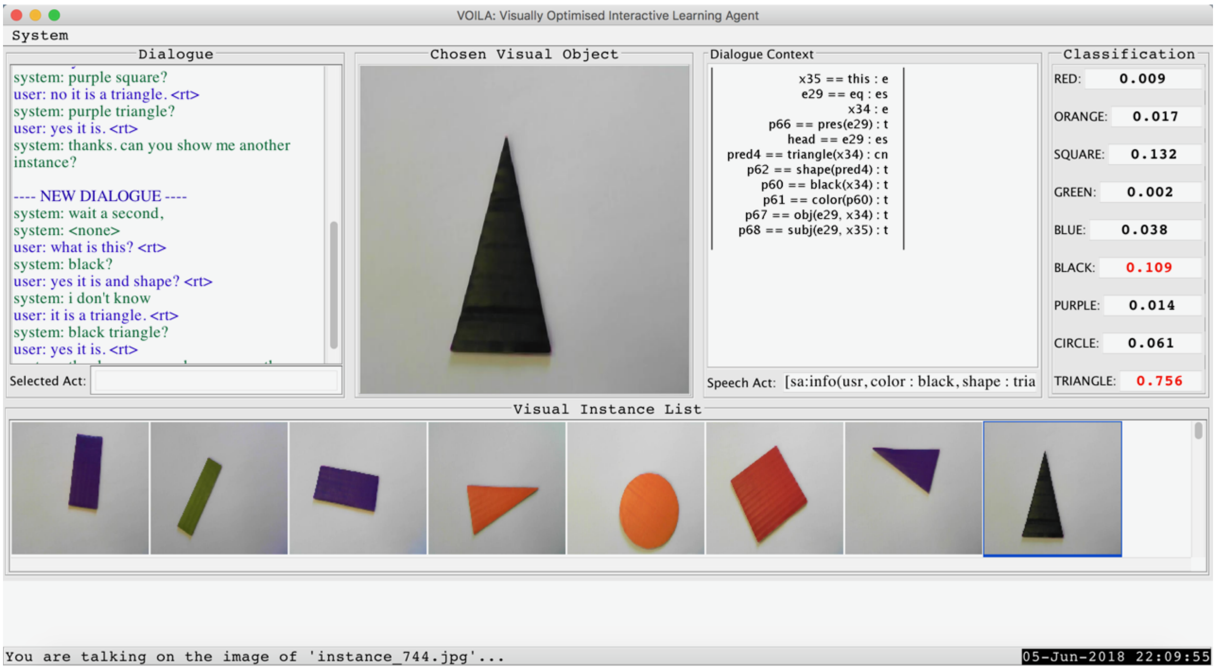


Figure 2: Incremental, visually grounded dialogue in the Concept Learning System. T= tutor, S=system

can utilise the DS-TTR grammar and its existing knowledge about the world, encoded in its classifiers, to make reference to and formulate questions about the different attributes of objects identified in the visual scene.

The most recent system has been learned from the BURCHAK corpus: a collection of human-human dialogues in the same domain (Yu et al., 2017b), and optimised using Reinforcement Learning to minimise cost for the tutor on the one hand, and maximise the accuracy of the learned visual word meanings on the other (c.f. Yu et al. (2016)) who use synthetic dialogue data, but with the same overall architecture).

We will show an interactive demonstration of this system, illustrating how questions, answers and object descriptions are derived and generated by the system in real-time. We will also show how the various components of the system operate.

Acknowledgements

Eshghi & Lemon were supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project[2]). Yu was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688147 (MuMMER project[3]).

[2]https://sites.google.com/site/hwinteractionlab/babble

[3]http://mummer-project.eu/

References

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

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics, pages 271–323. North Holland.

A. Eshghi, M. Purver, and Julian Hough. 2011. Dylan: Parser for dynamic syntax. Technical report, Queen Mary University of London.

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.

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

Arash Eshghi. 2015. Dylan: An incremental, semantic, contextual parser for dialogue. In *Proceedings of SEMDIAL 2015 (GoDIAL)*.

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

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In J. Bos and S. Pulman, editors, *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK, January.

Danijel Skocaj, Matej Kristan, Alen Vrecko, Marko Mahnic, Miroslav Janícek, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IROS*, pages 3387–3394.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of SIGDIAL 2016, 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 339–349, Los Angeles.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2017a. Learning how to learn: an adaptive dialogue agent for incrementally learning visually grounded word meanings. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 10–19. Association for Computational Linguistics.

Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. 2017b. The BURCHAK corpus: a challenge data set for interactive learning of visually grounded word meanings. In *Proceedings of the EACL 2017 workshop on Vision and Language (VL'17)*, Valencia.