SemDial 2022

DubDial

Proceedings of the 26th Workshop
On the Semantics and Pragmatics of Dialogue

Held at
Technological University Dublin, Ireland / The Internet
August 22-24 2022

Eleni Gregoromichelaki, Julian Hough & John D. Kelleher (eds.)

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL
UNIVERSITY DUBLIN

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.semdial.org/

Co-presidents: Ellen Breitholtz and Julian Hough

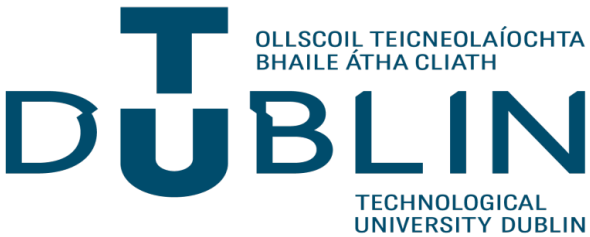
Anthologists: Christine Howes, Casey Kennington and Brielen Madureira

Webmasters: Janosch Haber, Julian Hough

DubDial Website

https://semdial2022.github.io/

DubDial Sponsors



ADAPT Research Centre
Technological University Dublin

DubDial Endorsements

Preface

Technological University Dublin (or *as Gaeilge*, in the Irish language, *Ollscoil Teicneolaíochta Bhaile Átha Cliath*) is delighted to bring the SemDial Workshop on the Semantics and Pragmatics of Dialogue to Dublin. TU Dublin is a very new university, legally established on the 1^{st} of January 2019. Although the university and much of its Grangegorman campus are brand new, the venue for DubDial is the beautiful Saint Laurence's Church which was built in 1850. Consequently, DubDial brings the old and new together and also, through its hybrid format, the in-person and virtual. It is wonderful to be able to meet each other again, and we wish everyone, in-person and online, an enjoyable, enlightening, and safe workshop.

This year, focusing on the general theme of interactivism, more familiar in Cognitive Science, Psychology, and Philosophy, we were fortunate to be able to include keynote speakers from that audience, to interact with the more traditional dialogue research themes including experimental studies, corpus studies, and computational and formal models. We also drew in a wide range of submissions from both the traditional SemDial authors and from those working in the interactivist tradition.

This year we received 27 full paper submissions, 19 of which were accepted as full papers after a peer-review process, during which each submission was reviewed by a panel of at least two experts. The poster abstracts had 23 submissions from a combination of recommended pre-accepted re-submissions of long papers and a further call for research in progress and short papers- 21 of these poster abstracts were presented. All accepted full papers and poster abstracts are included in this volume.

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

DubDial features three keynote presentations by Mark Bickhard, the Henry R. Luce Professor of Cognitive Robotics and the Philosophy of Knowledge at Lehigh University; Prof Joanna Rączaszek-Leonardi, Professor at the Faculty of Psychology, University of Warsaw and head of the Human Interactivity and Language Lab; Yvette Graham, Assistant Professor in Artificial Intelligence at the School of Computer Science and Statistics, Trinity College Dublin. We are honoured to have them in this year's SemDial and we thank them for their participation. Abstracts of their contributions are also included in this volume.

This year we also include a special session, the 2nd edition of SummDial, on the Summarization of Dialogues and Multi-Party Meetings, organized by Tirthankar Ghosal (Charles University), Xinnuo Xu, (University of Edinburgh), Muskaan Singh (IDIAP Research Institute, Switzerland) and Ondřej Bojar (Charles University) with a Keynote by Verena Rieser (Heriot-Watt University). The SummDial proceedings are appended as a final volume to the proceedings.

DubDial has received generous financial and in-kind support from Technological University Dublin. We are grateful to the wonderful Saint Laurence's Church venue team. We have also been given an endorsement by the ACL Special Interest Group SigDial.

We would also like to thank our local organizers at DubDial University, particularly John Kelleher for chairing and bringing SemDial to such a special setting. We also thank Julian Hough and Eleni Gregoromichelaki for core administrative support building up to the event, and to Chris Howes, Casey Kennington and Brielen Madureira for their support on converting the proceedings to the SemDial anthology format. We thank our local Technological University Dublin support team of Katryna Cisek, Hussain Ghulam, Elizabeth Hunter, Filip Klubička, Vasudevan Nedumpozhimani, Thi Nguyet Que Nguyen (Que), Michael O'Mahony, and Anh Duong Trinh (Senja). Thanks to everyone who helped with all aspects of

the organization.

And last, but not least, a special thank you to the authors, whose contributions make this an exciting SemDial with the potential for future work in tackling the ever enigmatic object of study that is dialogue.

Eleni Gregoromichelaki, Julian Hough and John D. Kelleher

Dublin

August 2022

Programme Committee

Eleni Gregoromichelaki (chair)	University of Gothenburg
Julian Hough (chair)	Queen Mary University of London
John D. Kelleher (chair)	Technological College Dublin
Jedediah Allen	Bilkent University
Maxime Amblard	University of Lorraine, LORIA
Ron Artstein	University of Southern California
Alex Berman	University of Gothenburg
Maria Boritchev	Polish Academy of Sciences
Ellen Breitholtz	University of Gothenburg
Harry Bunt	Tilburg University
Robin Cooper	University of Gothenburg
Valeria De Paiva	Topos Institute
Emilie Destruel	University of Iowa
Simon Dobnik	University of Gothenburg
Kerstin Fischer	University of Southern Denmark
Kallirroi Georgila	ICT, University of Southern California
Emer Gilmartin	Trinity College Dublin
Jonathan Ginzburg	Université Paris Cité
Christine Howes	University of Gothenburg
Julie Hunter	Lingora
Ruth Kempson	King's College London
Staffan Larsson	University of Gothenburg
Alex Lascarides	University of Edinburgh
Andy Lücking	University of Frankfurt / Université Paris Cité
Chiara Mazzocconi	Aix-Marseille Université
Gregory Mills	University of Groningen
Robert Mirski	John Paul II Catholic University of Lublin
Bill Noble	University of Gothenburg
Massimo Poesio	Queen Mary University of London
Laurent Prévot	Aix-Marseille Université
Matthew Purver	Queen Mary University of London / Jožef Stefan Institute
Hannes Rieser	Bielefeld University
Robert Ross	Technological University Dublin
David Schlangen	University of Potsdam
Matthew Stone	Rutgers, State University of New Jersey
Peter Sutton	Universitat Pompeu Fabra, Barcelona
Lucas Thorpe	Bogazici University
Ye Tian	Wluper

Local Organizing Committee

John D. Kelleher (chair)	Technological University Dublin
Julian Hough (webmaster and admin)	Queen Mary University of London
Eleni Gregoromichelaki (admin)	University of Gothenburg
Katryna Cisek	Technological University Dublin
Hussain Ghulam	Technological University Dublin
Elizabeth Hunter	Technological University Dublin
Filip Klubička	Technological University Dublin
Vasudevan Nedumpozhimani	Technological University Dublin
Thi Nguyet Que Nguyen (Que)	Technological University Dublin
Michael O'Mahony	Technological University Dublin
Anh Duong Trinh (Senja)	Technological University Dublin

Table of Contents

Invited Talks

Full Papers

Poster Abstracts

Invited Talks

Early interaction: language ungrounding and grounding in dialogue

Joanna Rączaszek-Leonardi
Human Interactivity and Language Lab
Faculty of Psychology
University of Warsaw
raczasze@psych.uw.edu.pl

Dialogue, as a naturally collaborative phenomenon, does not yield easily to individual-based explanations. Attempts to model it as series contributions of encoded contents and responses to them often miss the readiness to take-up the turn and the participatory sense-making in the co-construction of utterances. Approaches that focus on processes and that allow emergent dialogical structures as a level of organization, should be more relevant. Early interactions are particularly vivid examples that this level of organization is present from the start. The division of sense-making labour is evident as infants are treated as agents and contributors to routines. Language accompanies these early co-actions with its particular rhythm and placement within interactive events. Recognizing that linguistic dialogues emerge amidst coaction in development facilitates tracing the processes that lead to individuation of linguistic layer and its certain freedom from immediate co-action on the one hand, and understanding its continuous power to control interactive dynamics on the other.

Language as an (Inter-)Action System

Mark H. Bickhard
Lehigh University
mhb0@lehigh.edu

The classic framework of sensory transduction, cognitive processes, and encoding of mental contents into utterances is seriously problematic, and many researchers reject it. But it is not necessarily easy to diagnose, uncover, and correct the theoretical and presuppositional problems that support this framework. I will argue that one basic problematic assumption is that all representation is constituted as encodings, and that this assumption, in turn, is based on an underlying 'substance' or 'entity' metaphysics. These assumptions have dominated Western thought for some time, so, if they are not only false but also impossible (as will be argued), then it is understandable that modeling cognitive processes, including those of language, have been obstructed. I will briefly outline an alternative process metaphysics, and develop an outline of a model of language processes within that framework that construes languaging as a joint activity — an interactive activity — among participants in social situations.

Towards better dialogue system evaluation

Yvette Graham
School of Computer Science and Statistics
Trinity College Dublin
ygraham@tcd.ie

Evaluation of open-domain dialogue systems is highly challenging, and development of better techniques is highlighted time and again as desperately needed. Despite substantial efforts to carry out reliable live evaluation of systems in recent competitions, annotations have been abandoned and reported as too unreliable to yield sensible results. This is a serious problem since automatic metrics are not known to provide a good indication of what may or may not be a high-quality conversation. Answering the distress call of competitions that have emphasized the urgent need for better evaluation techniques in dialogue, this talk presents the successful development of human evaluation that is highly reliable while still remaining feasible and low cost. Self-replication experiments reveal almost perfectly repeatable results with a correlation of $r = 0.969$.

Due to the lack of appropriate methods of statistical significance testing, the likelihood of potential improvements to systems occurring due to chance is rarely taken into account in dialogue evaluation, and the evaluation presented facilitates application of standard tests. Highly reliable evaluation methods then provide new insight into system performance and this talk includes a comparison of state-of-the-art models (i) with and without personas, to measure the contribution of personas to conversation quality, as well as (ii) prescribed versus freely chosen topics. Interestingly with respect to personas, results indicate that personas do not positively contribute to conversation quality as expected, a surprising result that will hopefully inspire discussion within the dialogue community.

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

Full Papers

The Integrated Model of Memory: a dialogical perspective

Jonathan Ginzburg[1] **and Andy Lücking**[1,2]
[1]Université Paris Cité, CNRS,
Laboratoire de Linguistique Formelle (UMR 7110)
[2]Goethe University Frankfurt
yonatan.ginzburg@u-paris.fr, luecking@em.uni-frankfurt.de

Abstract

The increasing complexity of dialogue information states raises the question of their ontological status. To this foundational question one can add a more concrete concern: all existing semantic frameworks for dialogue while designed to explain how meaning emerges from the 'accumulation of information', have no corresponding means of *eliminating* information. Our claim, which we exemplify, is that *memory boundedness impacts dialogue coherence*. This paper aims to offer an initial sketch of an approach that both resolves the foundational issue raised above and the issue of memory fragility. We propose to construe dialogue information states as properties of brain networks. This follows in the programme of brain-grounded semantics (Hagoort, 2020). Our strategy involves taking a recent framework for describing the dynamics of memory (Bastin et al., 2019) as a basis for developing a suitable notion of cognitive states and their dynamics for dialogue interaction. We sketch a semantic description of this system, suggesting that this imposes strict conditions on potential semantic frameworks.

1 Introduction

All contemporary semantics for dialogue are *dynamic*: they view many aspects of meaning as emerging from context change. But whereas 'context' was an inert, abstract notion in early Montague semantics (Montague, 1974) and an eventuality in situation semantics (Barwise and Perry, 1983), dynamic semantics starting with Discourse Representation Theory (DRT) (Kamp, 1981) identified contexts with *information states*. Whereas originally such information states tracked discourse referents and presuppositions, in recent work on dialogue information states have become complex as a wide range of phenomena have been analyzed, including the visual field (Lücking, 2016) (for analyzing manual gesture), emotional structure (Ginzburg et al., 2020) (for analyzing laughter), and defeasible common sense knowledge (*topoi/enthymemes* (Breitholtz, 2020) (for analyzing rhetorical relations). While there seems little doubt that this range of information is used in dialogue interaction, it does raise the question what kind of entity encompasses all these diverse types of information. What *is* the *dialogue gameboard* (DGB) posited in frameworks like KoS (Ginzburg, 2012)?

One is free to adopt a Cartesian perspective, as has often been the case in Chomskyan theoretical linguistics, though this is arguably an avenue that leads to untestable modelling (Poeppel and Embick, 2005). To this foundational question one can add a more concrete concern: all existing semantic frameworks for dialogue while designed to explain how meaning emerges from the 'accumulation of information', have no corresponding means of *eliminating* information—there are operations in DRT that make discourse referents inaccessible and KoS has notions of downdating questions, but long-term information established as accepted, is locked in for ever more. This means that, as Ginzburg and Lücking (2020) put it, 'forgetting is forgotten'—there is no natural way to deal with the fragility of memory, an intrinsic and concrete feature of human interaction, both involving neurotypicals and non-neurotypicals like dementia sufferers. Our claim, exemplified below in section 2, is that *memory boundedness impacts dialogue coherence*.

This paper aims to offer an initial sketch of an approach that both resolves the foundational issue raised above and the issue of memory fragility. The basic idea is straightforward, namely to construe dialogue information states as properties of brain networks (Bressler and Menon, 2010). This follows in the programme of brain-grounded semantics (Hagoort, 2020). This emphasizes the need to ground semantics in brain–internal processes, while ensuring that top-down causation (coming from the computational level, in this case, say the DGB) is given its due (Campbell, 1974). Thus, in Marrian

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

terms, this does not mean in any way downgrading the computational level of explanation, as provided by semantic theories of dialogue, but ensuring that this is commensurate with the algorithmic (and ultimately) implementational levels.[1]

Our strategy will be to take a recent framework for describing the dynamics of memory (Bastin et al., 2019), which we survey in section 3, as a basis for developing a suitable notion of cognitive state for dialogue interaction. In section 4 we sketch a semantic description of this system, suggesting that this imposes strict conditions on potential semantic frameworks—requiring probabilistic judgements and operations adding and removing structure from representations; in contrast to the passive view implicit in the lab encloistered memory literature, recollection processes are constituents of interactions, giving rise to clarification interaction, laughter, and crying. We exemplify the framework with reference to our earlier examples in section 5. This does not mean a behaviorist account eschewing unobservables, but an attempt to formulate theory in a way that is *ceteris paribus* consistent with current observations about brain geography and dynamics.

We build on an earlier work (Ginzburg and Lücking, 2020) that tried to forge a link between dialogue semantics and theories of memory. In particular, the assumption that DGBs are constituents of *episodic memory*. The emphasis in the earlier paper was on short-term memory aspects of dialogue, which are indeed the most salient aspects needed for dialogue processing (resolution of indexicals, non-sentential utterances, disfluencies etc), though the paper also addressed long-term aspects. We will concentrate on the latter here while offering some significant modifications to the earlier account. That account was primarily a formalization of a Baddeley style architecture (Baddeley, 2012), which is highly motivated empirically, but has no pretensions to direct brain realization (Hasson et al., 2015). We will not assume a dichotomous short/long-term distinction, but follow, e.g., Hasson et al. (2015) by assuming that such differences can be captured in terms of short/long temporal receptive windows (Kiebel et al., 2008; Golesorkhi et al., 2021), a view which is also consistent with recent work that suggests that time-dependent forgetting across both short and long terms is related to degradation of hippocampal-dependent relational information (Sadeh and Pertzov, 2020).

[1] We hope this provides at least a partial answer to a worry expressed by an anonymous SemDial reviewer 'I don't see why we can't leave the modeling of when something is accessible through memory, and for how long, to the cognitive scientists and then on the linguistic side pick up the ball once it has been determined that there is or is not a referent.'

2 Memory and Dialogue Coherence: some data

Consider first (1). The initial laughs by A and B, as suggested by Ginzburg and Lücking (2020), arise as a consequence of the clash between the observed visual scene and the topos *presidents wear formal suits*. Now consider B's second laugh a year later: this is ambiguous between a laugh about the incongruity of the recollected event of viewing Putin or a pleasure laugh about the autobiographical event a year before. This can only be explained by appeal to episodic memory (and semantic memory for the topos), distinctions unavailable in standard dynamic semantic treatments of context.

(1) A and B observe Putin wearing a hazmat suit on tv:[2]

A: laughs
B: laughs
[*A year later:*]
A: Do you remember that bizarre situation with Putin during Covid?
B: laughs

(2) is an apocryphal story about the mathematician Paul Erdös. This illustrates a basic feature of conversational interaction, namely that this involves an initial check whether the interlocutor is familiar or not; familiarity requires an initial intimacy interaction, whereas lack of familiarity (as here) an establishment of the interlocutor's identity:

[2] Kremlin.ru, CC BY 4.0, https://commons.wikimedia.org/wiki/File:Vladimir_Putin_in_Kommunarka_hospital1.jpg.

(2) ERDÖS: Where are you from?
MATHEMATICIAN: Vancouver
ERDÖS: Really? Then you must know my friend Elliot Mendelson.
MATHEMATICIAN: (pause) I am your friend Elliot Mendelson.

(3) illustrates forgetting on a short time scale so requires a means of explaining short-term lack of recollection for an event, along with the attendant potential for repair:

(3) CAROL: Suddenly this means a lot to them. Yes? / Critical illness cover, that's great. Excuse me a minute. (Knocking at the door)
UNKNOWN1: Sorry to interrupt, I've come to collect the packet. /
CAROL: Oh right, it's the bag, sorry there isn't one tonight. /
UNKNOWN1: See you then /
CAROL: Thanks for coming then, yes, bye. That's good, I forgot the post. **Erm, where was I? What was I talking about?** /
UNKNOWN2: Single people. (BNC)

(4) and (5) both involve dementia sufferers (participants (PAR) interacting with an investigator (INV)).[3] In (4) there is explicit reference to a failed word recollection; in (5) the speaker makes reference to a descriptively potent but name–lacking cognitive state. In both cases the speakers' cognitive states maintain social norms relating to embarrassment which underwrite the laughter, but (5) in particular exhibits depression, characteristic of dementia sufferers, in part caused by repeated memory failure.[4]

(4) Becker et al. (1994), Pitt corpus, fluency 043-0, 04–10
INV: I want you to tell me as many animals as you can think of in one minute „ okay?
INV: they can be farm animals or zoo animals or pets.
INV: they can't be birds or fish or insects „ okay?
INV: can you begin?
PAR: &=laughs no. [+ exc]
INV: +< no?
PAR: +< (be)cause I forgot. [+ exc]

(5) DePaul (2017), depaul2a, 12–15
PAR: I can picture &=points:forehead whatever things that I'm still seeing or whatever.
PAR: but I don't know what to call it.
PAR: that's [/] that's what's whatever.
PAR: when I go to heaven it's gonna be &=looks:down &=head:shakes fine &=laughs.

(6) illustrates that a successfully recalled event involves reappraisal:

(6) (Interview with Pete Doherty about his relationship with Barât—they had one of the most fractious relationships in rock music...)
JOURNALIST: And yet the intensity of your bond was palpable.
DOHERTY: Absolutely. You're making me quite emotional my eyes are filling with tears. (The Guardian, June 2022).

3 The Integrated Model of Memory

In this section we summarize the Integrative Model of Memory (IMM) (Bastin et al., 2019), a synthetic effort to incorporate recent neuropsychological models of memory. We use this framework as a basic description of relevant brain networks.

3.1 Basic phenomena

The two main phenomena the theory attempts to explain are the brain processes which give rise to (event) *recollection* and (entity) *familiarity*, examples of which we saw in section 2.

[3] Transcription follows the CHAT format (MacWhinney, 2000). The symbol "[/]" indicates a repeated attempt to produce a word, double comma „ is an interactional marker for an intonational group, the symbol "&=" prefixes speaker actions, "+<" indicates a slight overlap of utterances, parentheses enframe omitted material, "[+ exc]" is a user-defined postcode which marks utterances that are excluded from analyses.

[4] Although the relationships between dementia and depression are complex, e.g., Bennett and Thomas, 2014.

3.2 Basic explanatory mechanisms

IMM relies on a combination of distinct types of representations and processes in its explanation of recollection and familiarity. As far as representations go, it distinguishes between the following kinds of representations:

- Event representations: representations of relations between two or more entities. These representations are associated with the hippocampus. Following indexing theory (Teyler and Rudy, 2007), such representations do not actually store a detailed representation of the event, but an index that enables to retrieve from the neocortex the original modes in which the event was perceived (visual, aural etc).

- Entity representations. These are representations often arising in a one shot manner (Kent et al., 2016) in the perirhinal/anterolateral entorhinal cortex that allow the discrimination of objects with overlapping features such as faces in a viewpoint-invariant manner (Erez et al., 2016). This enables quick recognition of familiar objects in the stream of objects perceived in the environment. This location is also one where conceptual features may get bound to entities via interaction with the anterior temporal area (Martin et al., 2018). The entity-level representations in the anterolateral entorhinal/perirhinal cortex correspond to a higher level of representation of the object, representing the individual object in a way abstracted from its presentation characteristics (viewpoint, perceptual conditions of presentation etc.)

- Background representations:[5] a network linking the parahippocampal/posteromedial entorhinal cortex and the occipitoparietal cortex and retrosplenial cortex. This system provides the setting into which entities fit in within events, binding the two enables entities to gain distinct significance based on diverse settings.

Key processes in the IMM are:

- Pattern separation (Rolls, 2016; Ngo et al., 2021): a hipppocampal process in which similar inputs are given separate representations based on specific conjunctions of features.

- Pattern completion (Rolls, 2016; Ngo et al., 2021): a hipppocampal process by means of which a partial information cue triggers the reactivation of the complete pattern.

- Attribution mechanisms (Whittlesea and Williams, 2000): recollection and familiarity are not merely determined by the accuracy of representations, but by task-dependent confidence thresholds. In the highest band are commonly encountered entities whose familiarity is automatic, in the lowest band unknowns; the middle band consists of entities whose recognition triggers incongruity—this incongruity is the subjective feeling of fluency. Seeing a person resembling a work colleague will lead to different judgements and actions depending on whether I need to decide if to greet him or merely to report seeing him.

The representational structure of the IMM is summarized in Fig. 1.

3.3 Unofficial Extensions: Semantic Memory and Emotion

The IMM is an ambitious programme, but in its initial formulation at least (Bastin et al., 2019), it makes some understandable simplifying assumptions. We mention here two, which we think need to be eliminated for the viability of a linguistically oriented theory, using suggestions in Bastin et al. (2019) and in responses to the paper.

The IMM considers only episodic memory. But as argued in Greenberg and Verfaellie (2010) and, building on this, by Gainotti (2019), there is an intrinsic dependence between this system and what has been called *semantic memory*—"the memory necessary for the use of language" (Tulving, 1972, p. 386). There is ample evidence of disassociation between the two—medio temporal lobe (MTL) damage can severely hinder the subsequent formation of episodic memories without affecting semantic memory (Scoville and Milner, 1957), whereas semantic dementia, which leads to loss of naming ability, can have minor effects on episodic memories (Chan et al., 2001). Nonetheless, there is evidence that semantic memory facilitates the acquisition of new episodic memories and vice

[5] In the IMM these are called 'context representations' given that much of the experimental data derives from manipulation of cards with visual images. Since 'context' plays a major role in formal semantics, we have changed the terminology.

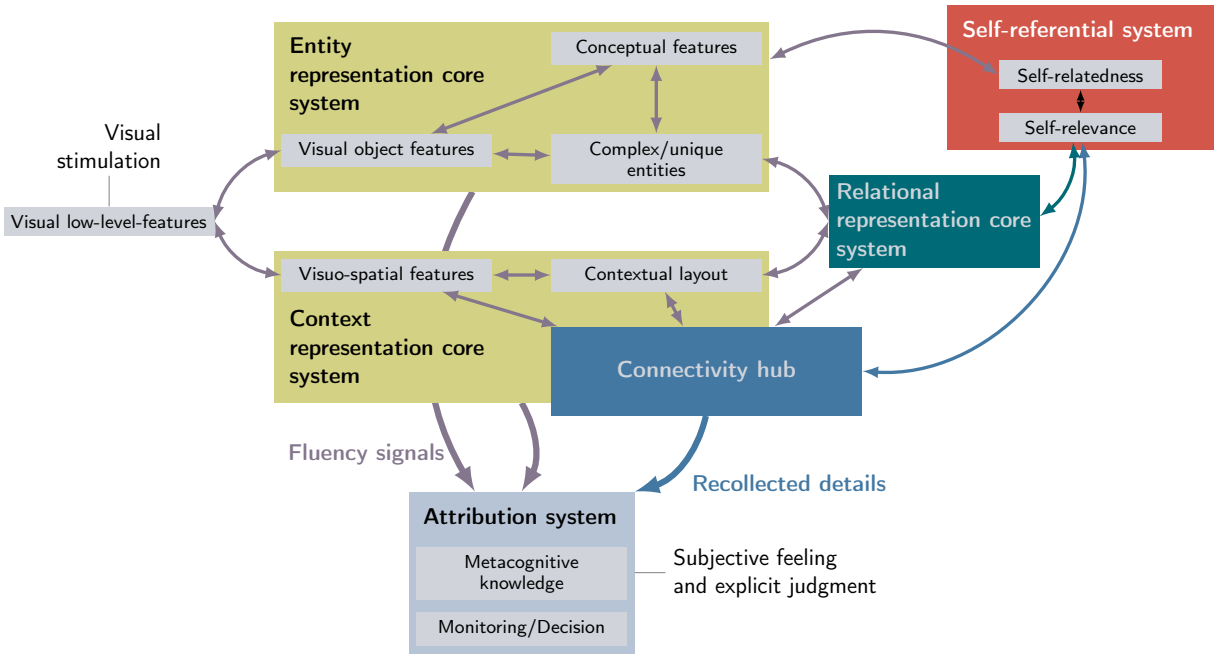


Figure 1: Slightly shortened illustration of the IMM, adopted from Bastin et al. (2019, p. 4).

versa (Greenberg and Verfaellie, 2010). Conversely, episodic memories facilitate the retrieval of information from semantic memory, and semantic memories constitute an important base from which complex and detailed episodic memories are constructed. The distinction between episodic and semantic memory is not straightforward and goes against a dichotomous explication (somewhat reminiscent of the stage level/individual level distinction in semantics (Carlson, 1977)). Tulving (1972) suggested a serial encoding hypothesis (perceptual → semantic → episodic), but the finding that episodic memory can facilitate new semantic learning is harder to reconcile with this, while it is consistent with the view of semantic memory as decontextualized episodic memory (Baddeley, 1988). However, as Greenberg and Verfaellie (2010) argue, the fact that degeneration of semantic memory is correlated with a severely weakened and vague episodic memory does not cohere well with a notion of parallel storage; a more attractive view is that episodic memory effectuates a binding between contextual information and material found in semantic memory.

The second simplification inherent in the current version of the IMM concerns its abstracting away from emotion. Already in the 1970s there was evidence that positively valenced events are remembered at a higher rate (Kintsch and Bates, 1977); there is much more recent evidence that emotional memories are forgotten at a slower rate than neutral memories over long timescales (e.g., a day vs. 5 min; Sharot and Yonelinas, 2008). Yonelinas and Ritchey (2015) argue that the slower forgetting of emotional memories can be linked to a dependence on the amygdala and its interaction with nonhippocampal MTL structures, rather than on the hippocampus. This hypothesis aligns well with the notion, promoted in Sadeh and Pertzov (2020), that hippocampal representations are more prone to temporal degradation than nonhippocampal representations.

3.4 Applications to memory deterioration

3.4.1 Neurotypical forgetting

One account of forgetting links it to *contextual drift* (Yonelinas et al., 2019). On this view forgetting as evinced in lab settings arises from a change or drift in context between study and test. Furthermore, on this approach, forgetting may be further promoted by *contextual interference*, such as intervening activities or physical changes. Manohar et al. (2019) suggest that memory encoding depends on rapid plasticity in flexibly coding neurons that may reside in the hippocampus. Such plasticity allows distinct representations that give rise to binding which results in a coherent memory representation. Time-dependent volatility of the synaptic weights is expected to lead to forgetting of relational or conjunctive information over time. Such forget-

ting does not occur "because of any specific decay rule, but rather because the plasticity rule operates continuously to alter all synaptic weights, and this 'erodes' the representations that are not currently active" (Manohar et al., 2019).

On the neuronal level, stored activation patterns (i.e. memory) are subject to three kinds of persistence-affecting processes, namely (i) neurogenesis (that is the creation of new neurons in, e.g., the hippocampus), (ii) synaptic weight decay, and (iii) synapse elimination (Richards and Frankland, 2017, p. 1072). As a consequence, memory even of neurotypical beings is a "transient" affair.

3.4.2 Neuroatypical memory failure

Neuroatypical characteristics may reinforce the afore-mentioned neuronal processes of synaptic (in-)stability. According to the IMM, the dissociation of recollection and familiarity in patients with lesions selective to the hippocampus or perirhinal/entorhinal cortex (e.g., Barbeau et al. (2011)) arises because the core representations are damaged. Clinical evidence validating these predictions is discussed in Bastin et al. (2019), in particular with respect to Alzheimer Disease.

4 Integrating Dialogue Semantics and Memory

In this section we introduce basic notions of KoS, which exemplifies a theory of dialogue states and their dynamics (at a computational level). We then sketch how this theory can be construed in terms of memory structures (at an 'algorithmic level').

4.1 Dialogical Cognitive States

KoS (Ginzburg, 2012; Ginzburg et al., 2020)—formulated using the logical framework TTR (Cooper and Ginzburg, 2015; Cooper, 2022)—is a theory of dialogue that offers an account of how speech events and other multimodal meaning bearing events change an individual's cognitive state. Instead of assuming a single context to be operative, a collective notion is emergent from individual *Total Cognitive States* (TCS), one per participant. A TCS has two partitions, namely a *private*—about which we will not elaborate here—for details see (Larsson, 2002), and a *public* one, the DGB.

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

Dialogue gameboards (see 8a for the basic structure) track various aspects of the emerging context The parameters *spkr* and *addr* together with the addressing condition (at a given time) track verbal turns and mutual engagement; *Vis-sit* represents the visual situation of an agent, including his or her focus of attention (*foa*), which can be an object (*Ind*), or a situation or event (*Sit*), relevant *inter alia* for processing gestural answers; *facts* represents the shared assumptions of the interlocutors; uncertainty about mutual understanding that remain to be resolved across participants—*questions under discussion*—are a key notion in explaining coherence and various anaphoric processes (Ginzburg, 2012; Roberts, 1996) and is tracked by the parameter *qud*; dialogue moves that are in the process of being grounded or under clarification are the elements of the *pending* list; already grounded moves are moved to the *moves* list; finally, *mood* represents the publicly accessible emotional aspect of an agent that arises by publicly visible actions (such as non-verbal social signals, as well as by verbal exclamations), which can but need not diverge from the private emotional state; the result of appraisals is given in terms of structures like (8b) (Russell, 2003).

(8)

a. $\mathit{DGBType} =_{def}$

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

b. $\mathit{Appraisal} =_{def}$

$$\begin{bmatrix} \text{pleasant} : \begin{bmatrix} \text{Pred} = \mathit{Pleasant} : \mathit{EmotivePred} \\ \text{affect} : \begin{bmatrix} \text{pve} : \mathbb{N} \\ \text{nve} : \mathbb{N} \end{bmatrix} \end{bmatrix} \\ \text{responsible} : \mathit{RecType} \\ \text{power} : \begin{bmatrix} \text{Pred} = \mathit{Powerful} : \mathit{EmotivePred} \\ \text{control} : \mathbb{N} \end{bmatrix} \end{bmatrix}$$

Conversational rules are the means for specifying how DGBs evolve. The types specifying its domain and its range we dub, respectively, the *pre(conditions)* and the *effects*, both of which are

subtypes of DGBType: they apply to a subclass of records that constitute possible DGBs and modify them to records that constitute possible DGBs. Conversational rules are written here in a form where the preconditions represent information specific to the preconditions of this particular interaction type and the effects represent those aspects of the preconditions that have changed.

KoS can represent locutionary, (9a,b), illocutionary updates, as in (9c,d), and emotion-based updates, such as (9e):

(9) a. Utterance integration: an utterance is perceived, updates Pending as a *locutionary proposition* (a record consisting of a representation of the utterance u and a grammatical type T_u calculated to classify it); there is then an attempted instantiation of the contextual parameters of T_u; if successful, the locutionary proposition is updated with the contextual instantiation and an attempt is made to find an appropriate Move update rule; if successful, Moves gets updated; otherwise repair ensues: the utterance remains in Pending and a clarification question is calculated and posed.

Clarification question: if A's utterance u is in Pending, QUD can be updated with the question *What did A mean by u*.

b. Ask/Assert QUD-incrementation: given a question q and ASK(A,B,q)/Assert(A,B,p) being the LatestMove, one can update QUD with q/p? as MaxQUD.

c. QSPEC: this rule characterizes the contextual background of reactive queries and assertions—if q is MaxQUD, then subsequent to this either conversational participant may make a move constrained to be q-specific (i.e., either a direct answer or a sub–question of q).

d. Positive affect incrementation of Mood: given the LatestMove being an incongruity proposition by the speaker, the speaker increments the positive pleasantness recorded in Mood to an extent determined by the laughter's arousal value.

The latter rule, which will play some role below, can be formalized as in (10)—updates are weighted between new and old values using the weight ε:[6]

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

$$\begin{bmatrix} \text{preconditions: } \left[\text{LatestMove.cont} : \mathit{IllocProp}\right] \\ \text{effect} : \begin{bmatrix} \text{Mood.pleasant.affect.pve} = \\ \varepsilon(\text{preconds.Mood.pleasant.affect.pve}) + \\ \quad (1-\varepsilon)\delta : \mathit{Real} \\ \text{Mood.pleasant.affect.nve} = \\ \varepsilon(\text{preconds.Mood.pleasant.affect.nve}) : \\ \quad \mathit{Real} \end{bmatrix} \end{bmatrix}$$

4.2 Dialogical Cognitive States and Memory Dynamics

Our starting point towards integrating dialogical cognitive states in memory is the idea from Ginzburg and Lücking (2020) that conversations are elements of episodic memory, which for concreteness we will assume are structured by DGBs. Whereas Ginzburg and Lücking (2020) considered short-term memory, within a Baddeley-style WM approach, we will not consider such aspects here, hence short-term elements relating to perception such as *Pending* (corresponding to the phonological loop) and *VisualSituation* (corresponding to the visuo-spatial sketchpad) are not included. What remain is specified by the type L(ong-term)DGBType, given in (11a).[7] Hence, we assume episodic memory track such episodes, as in (11b):

(11)

a. $\text{LDGBType} =_{def}$

$$\begin{bmatrix} \text{participants} = \left\{\text{x,y}\right\}\text{: } \mathit{Set}(\mathit{Ind}) \\ \text{Moves} : \mathit{List}(\mathit{LocProp}) \\ \text{QUD} : \mathit{Poset}(\mathit{Question}) \\ \text{Mood} : \mathit{Appraisal} \end{bmatrix}$$

b. $\text{Episodic} =_{def}$

$$\left[\text{Conversational} : \text{list(LDGBType)}\right]$$

We distinguish several distinct types of memory representations. Events are perceived visually or aurally or often multimodally. We assume such

[6] *NegativePleasantnessIncr* is the analogous operation incrementing the .nve and .pve values of pleasantness *mutatis mutandis*.

[7] Eliminating *Pending* and *VisualSituation* from LDGBType is a simplifying assumption. There clearly has to be some representation of the perceptible visual scene during a conversation as part of its recollection. This issue relates to the fundamental issue of how short-term memory structure relates/maps onto long-term memory structure which we plan to address in an expanded version of this paper.

events are represented by structured, relational representations—formally via TTR record types (Cooper, 2022); the tokens are the external, real world manifestations of the internal types.[8] Events undergo appraisal which leads to both updates in the current emotional makeup of the cognitive state (see the type *Appraisal* above) and to creating episodic indices in the hippocampus, which are in effect vertices in a network connecting to percepts of events stored neocortically. We assume that such indices are created for events with positive pleasantness above a threshold or negative pleasantness above a larger threshold—which yields a bias for long-term memory of enjoyable events or of highly unpleasant ones. The rule in (12) creates a fresh index and associates it with the current pending event ("HC" abbreviates *hippocampus*):[9]

(12) HC index creation

$$\begin{bmatrix} \text{preconds} : \begin{bmatrix} \text{Pending} : \textit{RecType} \\ \text{c1} : \text{Pending.Mood.pleasant.affect.pve} \\ \geq \theta_1 \\ \quad \vee \text{Pending.Mood.pleasant.affect.nve} \\ \geq \theta_2 \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{n} = \text{card(HC-Indices)+1} : \mathbb{N} \\ \text{HC-indices} := \text{HC-Indices} \cup \left\langle \text{n,Pending} \right\rangle \end{bmatrix} \end{bmatrix}$$

Both entity and semantic memory representations[10] are modelled as record types whose external witnesses correspond to real world individuals and (spatio-temporally unlocated) facts about these. We assume these arise from event percept (representations) by record type projection. We do not offer here general definitions, merely exemplify for the entity case:

(13) Entity representation creation:

a. Input: $\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{C} : \textit{faceshape} \\ \text{c1} : \text{C(x)} \\ \text{c}_{\textit{name}} : \text{Name(Emmo,x)} \\ \text{y} : \textit{Ind} \\ \text{c2} : \text{Hammer(y)} \\ \text{t} : \textit{Time} \\ \text{c3} : \text{Hold(x,y,t)} \end{bmatrix}$

b. Output: $\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{C}: \textit{faceshape} \\ \text{c1} : \text{C(x)} \\ \text{c}_{\textit{name}} : \text{Name(Emmo,x)} \end{bmatrix}$

Building on the discussion in section 3, we can describe the process for testing whether an entity is familiar. For simplicity we assume that the parameter used by the attribution system is relativized by the maximal element of QUD, though clearly this is a more intricate, domain sensitive (range of) parameter(s):[11]

1. Given an entity of type T_{source}, one searches in *Entities* for a match, a type T_{target} such that $T_{source} \sqsubset T_{target}$.

2. If one finds T_{target} such that $prob(match(T_{source}, T_{target}, \text{MaxQUD})) \geq \theta_{high}$, then $known(T_{source}.x)$.

3. If one finds T_{target} such that $\theta_{high} \geq prob(match(T_{source}, T_{target}, \text{MaxQUD})) \geq \theta_{low}$, then $familiar(T_{source}.x)$.

4. If all potential matches are evaluated as $\theta_{low} \geq prob(match(T_{source}, T_{target}, \text{MaxQUD}))$, then $\neg familiar(T_{source}.x)$

Given this notion of familiarity, we can sketch the process of *familiarity testing* that occurs as an interaction is initiated, resulting either in the latest-move (l-m) being an initial pleasantry or identity clarification:[12]

(14) Familiarity witnessing

a. $\begin{bmatrix} \text{preconds} : \begin{bmatrix} \text{moves} = \left\langle \right\rangle \\ \text{addr} : \textit{Ind} \\ \text{c1} : \text{familiar(addr)} \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{l-m.cont} : \textit{IllocProp} \\ \text{q} : \textit{Question} \\ \text{c2} : \textit{Recent-common-experience}(\text{q}) \\ \text{c3}: \textit{Co-Propositional}(\text{l-m.content,q}) \end{bmatrix} \end{bmatrix}$

[8] Though of course misperception/delusion can lead to representations without external counterparts.

[9] Although we do not spell this out here, we could postulate additional binding to the amygdala in case of strong emotional arousal, both negative and positive (Maren and Quirk, 2004; Phelps, 2004). This would capture the fact that it is much more difficult to forget highly emotional events since the amygdala is more stable than the HC.

[10] It is not impossible to have episodic metalinguistic memories, but not the norm.

[11] We thank an anonymous SemDial reviewer for a subtle but important correction of stage 2 of the process.

[12] Here two utterances are CoPropositional if the questions (construed as propositional functions) they update QUD with (see rule 9b) have overlapping ranges (answers); for instance 'Whether Bo left', 'Who left', and 'Which student left' (assuming Bo is a student.) are all co-propositional.

b.
$$\begin{bmatrix} \text{preconds} : \begin{bmatrix} \text{moves} = \langle\rangle \\ \text{addr} : \textit{Ind} \\ \text{c1} : \neg\text{familiar(addr)} \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{l-m.cont} : \textit{IllocProp} \\ \text{q} = ?\text{Identity(addr)} : \textit{Question} \\ \text{c3: } \textit{Co-Propositional}(\text{l-m.content,q}) \end{bmatrix} \end{bmatrix}$$

Finally, we sketch event recollection.

1. Given an event of type T_{source}, one searches in the neocortex for a match accessible via an index in the hippocampus, a type T_{target} such that $T_{source} \sqsubset T_{target}$.

2. If one finds T_{target} such that $prob(match(T_{source}, T_{target}, \text{MaxQUD})) \geq \theta_{high}$, then $recall(T_{source})$ and $appraise(T_{target})$.

3. If all potential matches are evaluated as $\theta_{low} \geq prob(match(T_{source}, T_{target}, \text{MaxQUD}))$, then $\neg recall(T_{source})$.

Negative event recall has two consequences, an incrementation of *negative pleasantness* in Mood and the potential for clarification interaction (if to a co-present interlocutor or as a self-addressed question):

(15) a.
$$\begin{bmatrix} \text{preconds} : \begin{bmatrix} \text{e} : \textit{RecType} \\ \text{c1} : \neg\text{Recall(spkr,e)} \end{bmatrix} \\ \text{effects} : \left[\textit{NegativePleasantnessIncr}(\delta, \varepsilon)\right] \end{bmatrix}$$

b.
$$\begin{bmatrix} \text{preconds} : \begin{bmatrix} \text{e} : \textit{RecType} \\ \text{c1} : \neg\text{Recall(spkr,e)} \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{l-m.cont} : \textit{IllocProp} \\ \text{q} = \lambda P.P(\text{preconds.e}) : \textit{Question} \\ \text{c2: } \textit{Co-Propositional}(\text{l-m.content,q}) \end{bmatrix} \end{bmatrix}$$

On the model of the memory system sketched here, damage to the memory system can occur as follows:

- damage to the hippocampus: loss of event indices—some past experiences inaccessible, no way to create new event memories;
- damage to the perirhinal/anterolateral entorhinal cortex: fewer familiar individuals;
- damage to semantic memory: fewer means to talk about familiar individuals.

We summarize the basic structure of memory sketched here:

(16) Memory =
$$\begin{bmatrix} \text{Episodic} : \left[\text{Conversational} : \textit{list}(\text{LDGBType})\right] \\ \text{HC-indices} : \textit{set}(\langle \text{n} : \mathbb{N}, \text{e} : \textit{RecType} \rangle) \\ \text{Entities} : \textit{set}(\textit{RecType}) \\ \text{Sem-mem} : \textit{set}(\textit{RecType}) \end{bmatrix}$$

5 Discussion of Initial Examples

We can now return to reconsider the data from section 2.

Example (1) Initially we have a visual percept that includes several individuals; (in a tv size version of this scene) Putin is retrieved from entity memory, and retrieved from semantic memory is the fact that Putin is a leader and the topos 'leaders should wear formal clothes'.[13] The incongruity between the visual scene and the topos triggers the initial laugh. This leads to a pleasantness increment and the creation of a hippocampal index for the interaction and for the perceived visual scene. The interaction a year later involves successful recollection which can unify either on the index for the visual scene or for the conversational interaction. Whichever event is recalled is reappraised, so new potential for laughter.

Example (2) Originally Erdös had met Eliott Mendelson, who told him where he was from. This made EM and Vancouver familiar entities for Erdös, as well as updating his semantic memory in this respect. Due to Erdös's facial agnosia, when he encountered Eliott Mendelson, he was not (visually) familiar, which triggers the initial identity question. The answer to this question reveals the conceptually familiar entity Vancouver, which pattern completes to Elliot Mendelson, hence his deduction.

Example (3) In this case Carol's initial interaction is interrupted, which leads to the initial interaction being imperfectly recalled, perhaps via the mechanism proposed by Manohar et al. (2019) (viz. plasticity of synaptic weights; cf. section 3.4.1) and licensing the clarification interaction.

Examples (4) and (5) In both cases we have damaged semantic memories; the failed recollection

[13] Whether topoi live in semantic memory or in some more procedural section of memory we will not consider now.

licenses laughter in both cases, triggered by social incongruity the dementia sufferers are still aware of; the repeated recall failures take their toll in the depression exhibited in (5).

Example (6) This simply illustrates that successful recall triggers appraisal of the recalled event, with the consequent signals (laughter/crying) this can give rise to.

6 Conclusions and Future Work

In this paper we have sketched in rough outline a potential construal of certain aspects of dialogue context in terms of brain networks. We have suggested that this is the most parsimonious answer to the question of how to construe what dialogue contexts are in a way that directly captures memory fragility. This is, in turn, we have argued, pervasively present in interaction and needs to be integrated in accounts of dialogue coherence. At the same time we emphasize that the aim is not to replace computational theories of dialogue, which need to specify interaction in high level terms; the aim is to ensure bi-directional communication between such theories and theories formulated at the algorithmic and impementational levels of brain structures. While the roughness of our sketch is in no doubt, we believe that providing a dialogue–oriented semantics to models coming from neuropsychological research into memory has the potential of pushing such research to address spontaneous dialogue, which is an important aim.

References

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

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

Emmanuel J. Barbeau, Jérémie Pariente, Olivier Felician, and Michele Puel. 2011. Visual recognition memory: A double anatomo-functional dissociation. *Hippocampus*, 21(9):929–934.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. Bradford Books. MIT Press, Cambridge.

Christine Bastin, Gabriel Besson, Jessica Simon, Emma Delhaye, Marie Geurten, Sylvie Willems, and Eric Salmon. 2019. An integrative memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences*, 42.

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Sophia Bennett and Alan J. Thomas. 2014. Depression and dementia: cause, consequence or coincidence? *Maturitas*, 79(2):184–190. PMID: 24931304.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue*. Number 41 in Current Research in the Semantics/Pragmatics Interface. Brill, Leiden and Boston.

Steven L. Bressler and Vinod Menon. 2010. Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences*, 14(6):277–290.

Donald T Campbell. 1974. 'downward causation'in hierarchically organised biological systems. In *Studies in the Philosophy of Biology*, pages 179–186. Springer.

Greg N. Carlson. 1977. *Reference to kinds in English*. Ph.D. thesis, University of Massachusetts.

Dennis Chan, Nick C. Fox, Rachael I. Scahill, William R. Crum, Jennifer L. Whitwell, Guy Leschziner, Alex M. Rossor, John M. Stevens, Lisa Cipolotti, and Martin N. Rossor. 2001. Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Annals of neurology*, 49(4):433–442.

Robin Cooper. 2022. *From perception to communication: An analysis of meaning and action using a theory of types with records (TTR)*. Oxford University Press.

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

Roxanne DePaul. 2017. DementiaBank English PPA DePaul Corpus. doi:10.21415/T5ZH5T.

Jonathan Erez, Rhodri Cusack, William Kendall, and Morgan D. Barense. 2016. Conjunctive Coding of Complex Object Features. *Cerebral Cortex*, 26(5):2271–2282.

Guido Gainotti. 2019. What face familiarity feelings say about the lateralization of specific entities within the core system. *Behavioral and Brain Sciences*, 42.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, UK.

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

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

Mehrshad Golesorkhi, Javier Gomez-Pilar, Federico Zilio, Nareg Berberian, Annemarie Wolff, Mustapha C. E. Yagoub, and Georg Northoff. 2021. The brain and its time: intrinsic neural timescales are key for input processing. *Communications Biology*, 4.

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

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

Uri Hasson, Janice Chen, and Christopher J. Honey. 2015. Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, 19(6):304–313.

Hans Kamp. 1981. A theory of truth and semantic representation. In Jeroen Groenendijk, editor, *Formal Methods in Semantics*. Amsterdam Centre for Mathematics.

B. A. Kent, M. Hvoslef-Eide, L. M. Saksida, and T. J. Bussey. 2016. The representational–hierarchical view of pattern separation: Not just hippocampus, not just space, not just memory? *Neurobiology of Learning and Memory*, 129:99–106. Pattern Separation and Pattern Completion in the Hippocampal System.

Stefan J. Kiebel, Jean Daunizeau, and Karl J. Friston. 2008. A hierarchy of time-scales and the brain. *PLOS Computational Biology*, 4(11):1–12.

Walter Kintsch and Elizabeth Bates. 1977. Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory*, 3(2):150.

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

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

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Sanjay G. Manohar, Nahid Zokaei, Sean J. Fallon, Tim P. Vogels, and Masud Husain. 2019. Neural mechanisms of attending to items in working memory. *Neuroscience and Biobehavioral Reviews*, 101:1–12.

Stephen Maren and Gregory J. Quirk. 2004. Neuronal signalling of fear memory. *Nature Reviews Neuroscience*, 5:844–852.

Chris B. Martin, Danielle Douglas, Rachel N. Newsome, Louisa L. Y. Man, and Morgan D. Barense. 2018. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife*, 7:e31873.

Richard Montague. 1974. Pragmatics. In Richmond Thomason, editor, *Formal Philosophy*. Yale UP, New Haven.

Chi T. Ngo, Sebastian Michelmann, Ingrid R. Olson, and Nora S. Newcombe. 2021. Pattern separation and pattern completion: Behaviorally separable processes? *Mem Cognit*, 49(1):193–205.

Elizabeth A. Phelps. 2004. Human emotion and memory: interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology*, 14(2):198–202.

David Poeppel and David Embick. 2005. Defining the relation between linguistics and neuroscience. In A. Cutler, editor, *Twenty-first century psycholinguistics: Four cornerstones*. Lawrence Erlbaum.

Blake A. Richards and Paul W. Frankland. 2017. The persistence and transience of memory. *Neuron*, 94(6):1071–1084.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136. Reprinted in Semantics and Pragmatics, 2012.

Edmund T. Rolls. 2016. Pattern separation, completion, and categorisation in the hippocampus and neocortex. *Neurobiology of Learning and Memory*, 129:4–28.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Talya Sadeh and Yoni Pertzov. 2020. Scale-invariant characteristics of forgetting: Toward a unifying account of hippocampal forgetting across short and long timescales. *Journal of Cognitive Neuroscience*, 32(3):386–402.

William Beecher Scoville and Brenda Milner. 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, 20(1):11–21.

Tali Sharot and Andrew P. Yonelinas. 2008. Differential time-dependent effects of emotion on recollective experience and memory for contextual information. *Cognition*, 106(1):538–547.

Timothy J. Teyler and Jerry W. Rudy. 2007. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus*, 17(12):1158–1169.

Endel Tulving. 1972. Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of memory*. Academic Press, New York.

Bruce W. A. Whittlesea and Lisa D. Williams. 2000. The source of feelings of familiarity: the discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3):547.

Andrew P. Yonelinas, Charan Ranganath, Arne D. Ekstrom, and Brian J. Wiltgen. 2019. A contextual binding theory of episodic memory: systems consolidation reconsidered. *Nature Reviews Neuroscience*, 20(6):364–375.

Andrew P. Yonelinas and Maureen Ritchey. 2015. The slow forgetting of emotional episodic memories: an emotional binding account. *Trends in Cognitive Sciences*, 19(5):259–267.

Caregivers Exaggerate Their Lexical Alignment to Young Children Across Several Cultures

Thomas Misiek
Aix-Marseille University
thomasmisiek@gmail.com

Abdellah Fourtassi
Aix-Marseille University
abdellah.fourtassi@gmail.com

Abstract

As soon as they start producing their first words, children engage in dialogues with people around them. Recent work has suggested that caregivers facilitate this early linguistic communication via frequently re-using and building on children's own words. This tendency decreases over development as children become more competent speakers. While this pattern has been observed with data of English-learning children, the question remains as to whether this early child-caregiver dynamics is universal vs. culture-specific. We address this question using large-scale data in six languages belonging to both Eastern and Western cultures. We found that the finding generalizes well cross-linguistically, suggesting that caregivers' early "exaggerating" of lexical alignment is likely a scaffolding strategy used across cultures to facilitate children's early linguistic communication and learning.

1 Introduction

Lexical alignment is a phenomenon whereby interlocutors re-use each other's words in a dialog. For example:

- Speaker 1 :"How do you think this **is going?**"
- Speaker 2 :"Yes, I guess it **is going** well!"

Researchers have hypothesized this mechanism to be associated with dialog coordination, facilitating language processing and production and contributing to the collaborative process of building mutual understanding and, thus, communicative success more generally (Pickering and Garrod, 2004, 2006; Brennan and Clark, 1996).

Interestingly, a similar behavior has been documented in child-adult natural dialog, starting from the early stages of the child's language production (Dale and Spivey, 2006; Fernández and Grimm, 2014; Denby and Yurovsky, 2019; Fusaroli et al., 2021; Misiek et al., 2020; Yurovsky et al., 2016; Foushee et al., 2021).

In particular, two large-scale studies — using data from hundreds of children — by Yurovsky et al. (2016) and Misiek et al. (2020) converged on similar conclusions despite the fact they used different measures and focused on different aspects of alignment. The main finding was that caregivers *exaggerate* their re-use of children's early words/expressions when communicating with them. Another finding was that this exaggerated alignment decreases over time and becomes closer to children's own level of lexical alignment (as well as adult-adult alignment rate) by the end of the preschool period. A similar pattern was also observed in the context of second language (L2) learning between tutors and students (Sinclair and Fernández, 2021).

While lexical alignment is sometimes assumed to be largely automatic and priming-like in spontaneous adult-adult dialog (e.g., Pickering and Garrod, 2004), here the observed patterns of alignment suggest otherwise. In particular, the fact that adults align much more to young children (than the other way around), as well as the fact that there is a negative correlation between the adults' alignment and the children's age — and therefore their language proficiency — provide evidence that caregivers actually align as a *scaffolding strategy* to help the younger — less language proficient — children understand and/or learn (e.g., Vygotsky, 1978; Shafto et al., 2014; Yurovsky, 2018). Such a strategy would be less useful to older children with more developed linguistic skills and who need less communicative scaffolding from the caregiver.

1.1 The current study

The study of child-caregiver early lexical alignment dynamics has focused on data from English-learning children. It is still unknown whether the above-mentioned findings generalize to other languages/cultures, especially in the light of research that has pointed out cross-cultural dissimilarities in

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

the way caregivers interact with children early in development (Bornstein et al., 1992; Saint-Georges et al., 2013; Schick et al., 2022).

Addressing this question is of crucial scientific interest: It allows us to determine if the interactions observed between English-learning children and their caregivers reflect more the specificities of their culture (e.g., in terms of parenting style) or whether they represent universal patterns in human development across cultures. The current study is an effort to fill this gap. We conduct a large-scale study of lexical alignment in child-caregiver dialogues, comparing 6 languages: English, Chinese, Spanish, German, Japanese, and French.

2 Methods

2.1 Data

All the data is derived from CHILDES (MacWhinney, 2000; Sanchez et al., 2019), the largest public repository of child-caregiver dialog corpora. First, as shown in Table 1, we ranked all languages based on the size of their aggregated corpora. We aimed at selecting the subset of languages with the largest sizes, making sure we include at least 2 non-western cultures. Japanese was the second largest non-western language (after Chinese) with around 0.5 million words. We included French, which came next, since it had an approximately similar size as Japanese. We did not include the next language in the list since their size dropped significantly.

We focused on development in the pre-school period, ranging from 2 and 5 years old (data in CHILDES becomes too sparse below and above this range). Table 1 provides some summary statistics of the data we use. We note the heterogeneity in terms of the number of transcripts per child across languages, reflecting heterogeneity in data collection procedures (e.g., cross-sectional vs. longitudinal).

2.2 Measure of lexical alignment

Lexical alignment characterizes the speaker's reuse of words from the interlocutor's previous turns in the dialog. Following previous work (e.g., Fernández and Grimm, 2014; Misiek et al., 2020), we quantified this phenomenon by counting the number of shared unigrams (unique words) and bigrams (sequences of two successive words) across adjacent pairs of turns, normalized by the number of all possible ngrams.

Language	Words	Transc.	Children
English	11,801,282	5894	869
German	2,008,317	1073	54
Chinese	1,023,867	508	329
Spanish	665,789	493	63
Japanese	543,495	652	122
French	538,663	724	192
Slavic	385,839		
Afrikaans	288,927		
Romance	230,101		
Scandinavian	168,629		

Table 1: Top 10 languages with largest (aggregated) corpora in CHILDES. We focused on the top 6 with at least 0.5 million words each. For these languages, we show the number of transcripts (dialog sessions) and unique children aged 2 to 5 years.

We computed both `Child alignment` by comparing the child's turn to the adult's *previous* turn and `Adult alignment` by comparing the child's turn to the adult's *following* turn. In both cases, the pairs of turns have to be adjacent. If the same speaker has multiple consecutive utterances, only the first and the last were taken into account since only the first and last are adjacent to the interlocutor's utterances.

Baselines

In addition to the child's and caregiver's alignment measures, we derived two baselines. The first, which we call the `internal baseline`, computes the alignment of pairs of turns (one belonging to the child and the other to the caregiver) sampled randomly from the same transcript/conversation. The second, which we call `external baseline`, compares pairs of child and caregiver turns sampled randomly from the entire corpus (within a given language).

3 Results

Our first goal is to replicate findings for English data as reported in both Misiek et al. (2020) and Yurovsky et al. (2016). The second goal — and the novel contribution of the current study — is to test how previous findings in English generalize cross-linguistically. The results are shown in Figure 1.

We found the following findings both to replicate in English and to generalize well across languages:

1. Children align consistently to their caregiver, starting from the early stages of language production.

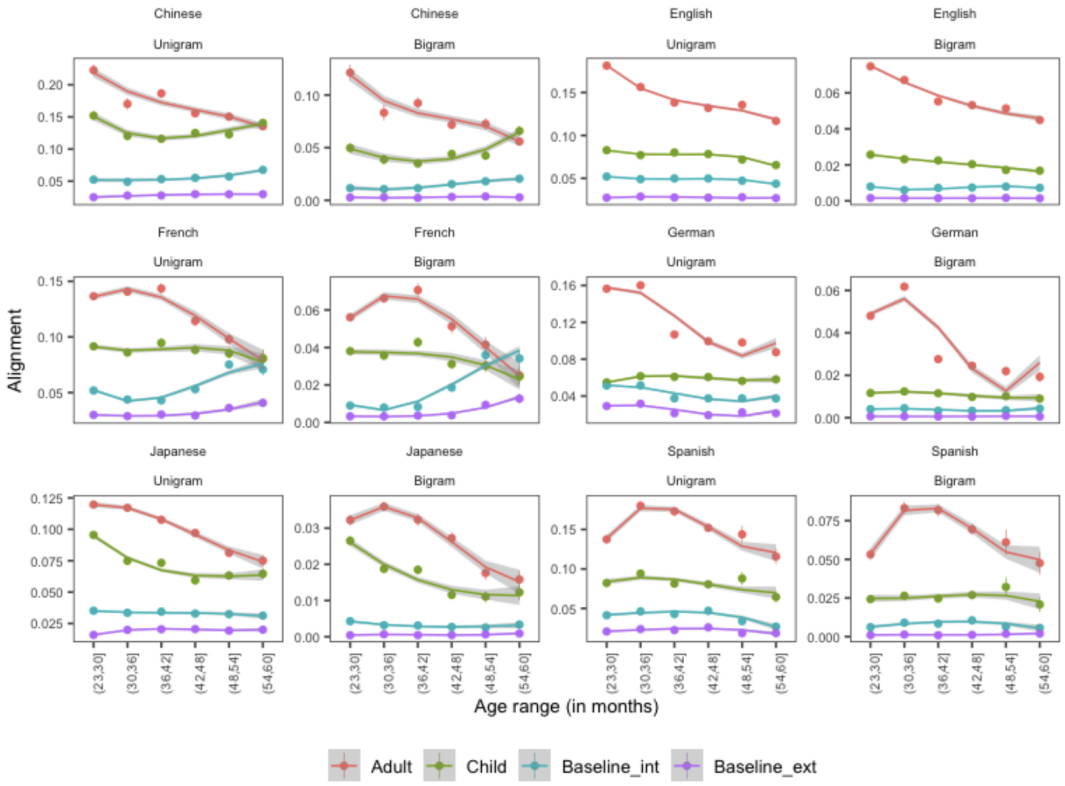


Figure 1: Lexical alignment measured in terms of shared unigrams and bigrams (normalized), as a function of the child's age across languages. The lines are fitted with polynomial splines of degree 3 with 95% confidence intervals.

2. Caregivers align consistently more to children (than the other way around).

3. Caregivers align more when children are younger. Their alignment decreases as children develop.

We corroborate these observations with statistical testing, but first we need to examine the shape of the data and make some simplifications. Figure 2 shows the distribution of (normalized) alignment values. It shows a 0-inflated distribution of a semi-continuous dependent variable. In other words, a substantial chunk of child-caregiver adjacent turns shows no alignment (i.e., the alignment value is exactly 0) and the rest is continuous between 0 and 1.

Standard normality transformations of such data do not solve the zero-inflation issue. One possible solution (to still be able to fit parametric models) is to consider a two-stage approach: a logistic regression predicting the binary 0 vs. non-0 outcome and a linear regression predicting the continuous outcome in the interval]0,1] (e.g., Gelman and Hill, 2006).

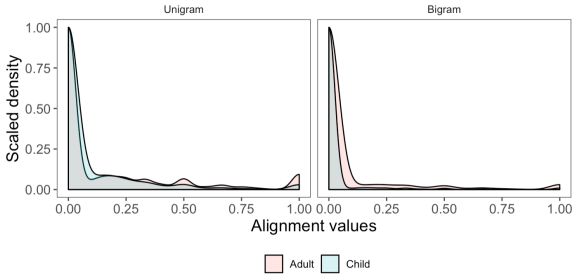


Figure 2: Scaled density plot of normalized alignment values in both unigram- and bigram-based measures, collapsed across all languages.

Here we are interested in comparing data across development, and the linear regression would, however, introduce distortions/biases, e.g., by creating a data size imbalance across ages. The reason is that restricting the data to the]0,1] interval requires selecting, at each developmental stage, only the subset of adjacent turns that include non-zero alignment. This would make it hard to interpret any observed developmental change.

Thus, for simplicity, here we only report results of the logistic regression predicting whether or not adjacent turns have at least one shared lexical un-

igram (for the unigram-based measure) or a least one shared lexical bigram (for the bigram-based measure). The logistic regression (unlike the linear regression on]0,1]) does not require removing data, only reducing its complexity from continuous to binary. This makes the interpretation of developmental *change* much more intuitive.

More precisely, we used mixed-effects logistic regressions, predicting the binary alignment (for both the unigram and bigram measures) as a function of the condition (Child vs. Adult) and age, using the identity of the child and the language as random effects. The results of these two regressions are shown in Table 2.

All predictors were highly significant, confirming the patterns observed in Figure 1: The predictor `Condition` indicates that caregivers align to children to a higher degree (than the other way around). `Age` negatively predicted alignment, showing that alignment decreases with development. The interaction `Condition*Age` shows that caregivers' alignment decreased faster than children's alignment did, confirming the observation that caregivers exaggerate alignment more to younger children than to older ones.

	Alignment	
	Unigram	**Bigram**
`(Intercept)`	-0.658^{***}	-2.207^{***}
	(0.030)	(0.038)
`Condition`	-0.575^{***}	-0.857^{***}
	(0.003)	(0.005)
`Age`	-0.110^{***}	-0.110^{***}
	(0.002)	(0.004)
`Condition*age`	0.125^{***}	0.065^{***}
	(0.003)	(0.005)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2: Estimates of two mixed-effects logistic regressions models predicting the presence of alignment (Unigram and Bigram) in adjacent child-caregiver turns as a function of `Condition` (who is aligning to whom) and child's `Age` (centered and scaled). The model was specified as `Alignment_or_not` $\sim$ `Condition*Age + (1 | child) + (1 | Language)`.

Cross-linguistic differences

In addition to the consistent cross-linguistic similarities, Figure 1 also shows some (minor) differences. For example, we can observe that the caregivers' decreasing alignment matches that of children by 5 years in some languages (i.e., Chinese, French, and Japanese) but not in others (i.e., English, German, and Spanish). In the latter case, it appears that caregivers are still exaggerating alignment despite children's relatively developed linguistic skills by that age.

Another difference concerns the pattern of children's alignment. While the developmental curve is rather stable in most languages, it tends to decrease in Japanese (although at a slower pace than the caregivers' curve does) and to slightly increase in Chinese.

We can also observe that for some languages, especially Spanish and French, the caregivers' curve tends to show an inverted U-shaped curve whereby the youngest children receive less alignment than the slightly older ones (before the curve starts decreasing again). This observation could be due to the fact that younger children have limited language production skills, providing much fewer opportunities for caregivers to align. This interpretation is supported by the fact that the inverted U-shaped curve is more pronounced in the bigram case, i.e., the case where children's utterance has to contain at least two words to provide the opportunities for the caregivers to align at the bigram level; the youngest children do produce much shorter utterances than older children do.

Finally, we observe that in French, the alignment curves become indistinguishable from the random baseline towards the end of the developmental period under study. However, this is likely due to the fact that in French (unlike all other languages), data of the oldest children had a much smaller sample size in CHILDES than the younger ones (data not shown), leading to noisier data by 5 years old.

4 Discussion

Lexical alignment is an important mechanism for dialog coordination in adults. Recent studies suggest it could play a role in child development as well: Adults tend to re-use children's words more frequently in the earlier stages of language production, perhaps scaffolding children's communicative and linguistic skills.

This paper showed that this finding generalizes

well — beyond English — to five different languages, including in three Western cultures (German, Spanish, and French) and two Eastern ones (Chinese and Japanese). The finding was strikingly similar despite variability in how data was collected across languages, and more important, despite the fact that Eastern and Western cultures are sometimes assumed to differ in terms of parenting style (Foo, 2019). The strong similarity among these languages points toward a rather *universal* pattern that characterizes the evolution of child-caregiver dialog dynamics across the first five years of life.

The developmental literature reports several scaffolding mechanisms that may underlie this finding. For example, caregivers tend to build on words and concepts that children already know in order to introduce new, more sophisticated ones, a strategy sometimes called "anchoring." For example, if the child knows/utters the word "rabbit," the caregiver can build on this knowledge to introduce the more abstract word "animal" that the child may not know yet (e.g., "Yes this is a rabbit, a rabbit is a kind of animal!") (Callanan, 1985) (but see Fourtassi et al., 2020).

Further, when children make mistakes, the caregivers tend to repeat the same utterance while correcting the mistake in it, a strategy known as "reformulation" (Chouinard and Clark, 2003). Caregivers also tend to borrow the children's syntactic structures (e.g., by re-using their verbs and function words), which, in turn, facilitate children's processing of the caregiver's next utterance (Yurovsky et al., 2016).

Future work is needed to examine the relative contribution of these strategies (and others) in explaining the "exaggerated lexical alignment" phenomenon and the potential variability of this relative contribution across cultures. In order to address this question at a large scale (which is crucial for more generalizable results), effort should be devoted to the development of automatic algorithms that characterize the caregivers' scaffolding strategies in naturalistic settings (e.g., Hiller and Fernández, 2016; Jiang et al., 2022; Nikolaus et al., 2021). Such an effort would also have applied implications, especially regarding the design of more effective child-oriented conversational AI for first or second language learning (Huang et al., 2022).

Finally, we return to the issue of cross-linguistic differences in the alignment patterns. While we reported several such differences in the results section, they do not necessarily reflect cultural or linguistic differences. The reason is that the corpora varied widely in terms of their sample size, the number of children involved, whether these children were followed or not in time (longitudinal vs. cross-sectional), as well as the multitude of contexts where the data was collected; these contexts were not necessarily similar across languages, perhaps inducing variability in alignment patterns (Dideriksen et al., 2020).

That said, and if anything, this variability makes our findings about cross-cultural *similarities* (i.e., the main claim of this work) stronger, since these similarities are observed *despite* variability in data sizes, collection procedures, and conversational contexts.

Limitations and future work

We only tested a handful of languages (the ones for which sufficient data was available in CHILDES). However more definitive conclusions would only come from the study of a world representative sample of child-caregiver dialogues, including in non-WEIRD[1] cultures (Henrich et al., 2010; Cristia et al., 2019).

Another limitation is that we focused only on one aspect of alignment (lexical repetition) which provides a partial view of how interlocutors align to each other multimodally in social interaction (Rasenberg et al., 2020). A more comprehensive investigation would require using child-caregiver corpora that facilitate the study of multimodal face-to-face conversations (e.g., Bodur et al., 2021, 2022).

Acknowledgments

The authors of this work have been supported by funding from the Institute of Language Communication and the Brain (ANR-16-CONV-0002) and the MACOMIC project (ANR-21-CE28-0005-01).

References

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. Chico: A multimodal corpus for the study of child conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 158–163.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2022. Backchannel behavior in child-caregiver video calls. In *Proceedings of the*

[1]Western, Educated, Industrialized, Rich, and Democratic.

44th Annual Meeting of the Cognitive Science Society.

Marc H Bornstein, Catherine S Tamis-LeMonda, Joseph Tal, Pamela Ludemann, Sueko Toda, Charles W Rahn, Marie-Germaine Pêcheux, Hiroshi Azuma, and Danya Vardi. 1992. Maternal responsiveness to infants in three societies: The united states, france, and japan. *Child development*, 63(4):808–821.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Maureen A Callanan. 1985. How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, pages 508–523.

Michelle M Chouinard and Eve V Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–669.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*, 90(3):759–773.

Rick Dale and Michael J Spivey. 2006. Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3):391–430.

Joseph Denby and Dan Yurovsky. 2019. Parents' linguistic alignment predicts children's language development. In *CogSci*, pages 1627–1632.

Christina Dideriksen, Morten H Christiansen, Kristian Tylén, Mark Dingemanse, and Riccardo Fusaroli. 2020. Quantifying the interplay of conversational devices in building mutual understanding.

Raquel Fernández and Robert Grimm. 2014. Quantifying categorical and conceptual convergence in child-adult dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Koong Hean Foo. 2019. *Intercultural Parenting: How Eastern and Western Parenting Styles Affect Child Development*. Routledge.

Abdellah Fourtassi, Kyra Wilson, and Michael C Frank. 2020. Discovering conceptual hierarchy through explicit and implicit cues in child-directed speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ruthe Foushee, Dan Byrne, Marisa Casillas, and Susan Goldin-Meadow. 2021. Differential impacts of linguistic alignment across caregiver-child dyads and levels of linguistic structure.

Riccardo Fusaroli, Ethan Weed, Deborah Fein, and Letitia Naigles. 2021. Caregiver linguistic alignment to autistic and typically developing children.

Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Sarah Hiller and Raquel Fernández. 2016. A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In *Proceedings of the 20th signll conference on computational natural language learning*, pages 105–114.

Weijiao Huang, Khe Foon Hew, and Luke K Fryer. 2022. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.

Hang Jiang, Michael C. Frank, Vivek Kulkarni, and Abdellah Fourtassi. 2022. Exploring patterns of stability and change in caregivers' word usage across early childhood. *Cognitive Science*, 46(7):e13177.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

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

Mitja Nikolaus, Juliette Maes, Jeremy Auguste, Laurent Prevot, and Abdellah Fourtassi. 2021. Large-scale study of speech acts' development using automatic labelling. In *Proceedings of the 43rd annual meeting of the cognitive science society*.

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

Martin J Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2):203–228.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive science*, 44(11):e12911.

Catherine Saint-Georges, Mohamed Chetouani, Raquel Cassel, Fabio Apicella, Ammar Mahdhaoui, Filippo Muratori, Marie-Christine Laznik, and David Cohen. 2013. Motherese in interaction: at the cross-road of emotion and cognition?(a systematic review). *PloS one*, 8(10):e78103.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51(4):1928–1941.

Johanna Schick, Caroline Fryns, Franziska Wegdell, Marion Laporte, Klaus Zuberbühler, Carel P van Schaik, Simon W Townsend, and Sabine Stoll. 2022. The function and evolution of child-directed communication. *PLoS Biology*, 20(5):e3001630.

Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. 2014. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89.

Arabella J Sinclair and Raquel Fernández. 2021. Construction coordination in first and second language acquisition. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children*, 23(3):34–41.

Daniel Yurovsky. 2018. A communicative approach to early word learning. *New Ideas in Psychology*, 50:73–79.

Daniel Yurovsky, Gabriel Doyle, and Michael C Frank. 2016. Linguistic input is tuned to children's developmental level. In *CogSci*.

Dialogue strategies for... cómo se dice entrenamiento de vocabulario?

Andrea Carrión Del Fresno and **Vladislav Maraev**
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
guscarrian@student.gu.se,
vladislav.maraev@gu.se

Staffan Larsson
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
and Talkamatic AB
staffan.larsson@ling.gu.se

Abstract

We carry out a small-scale empirical study of a dialogue strategy (conversational pattern) found in second language learner dialogues where a language-assisting teacher is present, allowing learners to pick up new words and train on them while maintaining a conversation. We also provide a formal model of the observed conversational pattern including several frequently occurring variants, as well as a demonstration implementation which is able to reproduce the most common variant of the pattern.

1 Introduction and previous work

We are interested in dialogue strategies for vocabulary training in second language learner's dialogues. By finding and analysing recurring patterns in human-human dialogues, we hope to provide a solid empirical basis for the implementation of dialogue strategies in dialogue systems for second language learning.

Varonis and Gass (1985) provide a model for the negotiation of meaning, where the flow of a conversation is described as a linear progression which is interrupted when the communication between the interlocutors is ineffective, causing a "push-down" effect in the normal flow and preventing it from moving forward. These interruptions are the result of a *t*rigger followed by a *r*esponse, where the response serves to indicate a lack of understanding of (or other problem related to) the trigger. For the conversation to resume its linear progress, some negotiation of meaning must take place. An example (from ibid.) is shown below:

A: yeah. How long . . will you be? will you be staying?
B: I will four months (trigger)
A: four months?
B: stay four months here until April

B's answer to speaker A's initial question is not understood properly by A, thus triggering both speakers to try to reach a mutual understanding in order to return to the main topic of the conversation.

Svennevig (2018) provides a CA (Conversation Analysis) style description of a conversational practice used by L2 speakers in acquiring new technical terms in the course of everyday workplace interaction on a construction site. Word search sequences contribute to disrupting the ordinary flow of the conversation. A word search sequence is described as the process by which the learner struggles to produce a full utterance, and is caused by a lack of vocabulary in the target language, thus motivating a request for assistance.

The learner's problems to complete the utterance could be accompanied by pauses and/or hesitation (Schegloff et al., 1977), and in some cases we find clues to indicate the missing information such as descriptions or code-switched explicit questions addressing how to say a specific word in the target language (Greer, 2013). As a collaborative word search sequence, the interlocutor is also involved, being responsible for providing the missing information after which the conversation prior to the interruption is resumed.

When the term is provided, it is repeated, displaying the L2 speaker's ability to pronounce the word. This repeat is treated as a request for confirmation by the L1 speaker, who often also provides further repeats of the word in question. See Figure 1 for an example. When searching for an L2 word that they cannot remember or do not have in their vocabulary, the speaker combines verbal and embodied means (such as gestures) to indicate the missing word and ask the interlocutor what it is called in Norwegian.

Using the terminology of Schegloff et al. (1977), the dialogue patterns described by Varonis and Gass are cases of other-initiated repair, whereas

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

```
1  TOM: e::h o::g (1.9) på khele tak,
        e::h a::nd (1.9) on whole roof,
2       (1.1) ((TOM points to picture on brochure))
3  TOM: ru:r, OG OG hva heter det på norsk. ((points out the window))
        pipes, AND AND what's it called in Norwegian.
4  NIL: >ja häng[ränna.<]
        yeah gutter.
5  DAG:          [  ja   ] ränn- [jaja
                    yeah  gutt- yeah yeah
6  TOM:                         [ja
                                 yeah
7  TOM: og den,
        and that,
8  DAG: ja,
        yeah,
9       (0.6)
10 TOM: o:g tretti år garan (.) [ti:re,
        a:nd thirty years warran (.) ty
11 DAG:                         [ja,
                                 yeah,
```

Figure 1: Dialogue example reproduced from Svennevig (2018)

those described by Svennevig are cases of self-initiated other-repair. Both these dialogue patterns can be expected to occur in second language learner dialogue, and the work presented here started out looking for the type of pattern identified by Varonis and Gass in second language learner corpora. However, such examples were fairly rare; instead, we found several instances of a pattern similar to that described by Svennevig, and attention shifted to this pattern.

This paper combines a small-scale empirical study of a conversational pattern similar to that described in Svennevig (2018) in second language learner dialogues. We also provide a simple computational model and a demonstration implementation which is able to reproduce the most common variant of the pattern. We believe such an implementation can be a very useful addition to conversational systems for second language learners. In Section 2, we describe the corpora and tools used, and then move on to the corpus study in Section 3. The formal model resulting from the corpus study is presented in Section 4, and the implementation based on the formal model is explained in Section 5. In Section 6, we provide conclusions and in Section 7 we describe future work.

2 Resources

The dialogue excerpts used in this paper were extracted from two different second language learner spoken corpora, namely, the European Science Foundation Second Language Databank (ESF)[1] and the Barcelona English Language Corpus (BELC)[2]. Both corpora belong to the SLABank collection[3], a part of TalkBank responsible for providing corpora in order to study the field of second language acquisition and learning.

2.1 ESF

This database collects spontaneous conversations between adults of different nationalities that are learning a second language, including Dutch, English, German and Swedish, and native speakers of those languages. It should be noted that only those conversations where English is the target language were used in this study.

A wide range of topics are covered in these conversations, from descriptions and role-plays to cultural activities. In addition to the transcripts, audio files are also available, which is useful in understanding conversational contributions in cases where the context provided by the transcript is insufficient.

2.2 BELC

The BELC corpus collects speech recordings of Spanish students between the ages of 8 and 18 who are learning English as a second or even third language (Catalan is also spoken in the area where the research was conducted). This corpus contains transcripts of spoken dialogues from four different tasks: written composition, role-play, oral narrative, and oral interview. The dialogue extracts used in this study come from the role-play task, where a pair of students are presented with a real-life

[1] https://slabank.talkbank.org/access/Multiple/ESF/

[2] https://slabank.talkbank.org/access/English/BELC.html

[3] https://slabank.talkbank.org/access/

situation where some negotiation takes place in the target language.

Importantly, there is also an investigator present, providing language support when needed. The investigator interacts with the subject using the target language, although it is shown that the investigator also knows the subject's mother tongue and resorts to it if necessary[4].

2.3 TalkBank browser and SCoRE

In order to access and collect the data, both the TalkBank Browser and SCoRE were used. The former is a browsable database that lets you navigate through transcripts from various corpora as well as watching or listening to any audio or video files attached to them, if available.

SCoRE[5] is a tool for browsing dialogue corpora, originally intended to search the British National Corpus (BNC) but now also able to access other corpora. The web interface allows the user to easily search a corpus with the help of regular expressions. While SCoRE was the main tool for browsing the ESF corpus to collect data, the TalkBank Browser provided access to the corpus' audio files. As for the BELC corpus, the TalkBank Browser was the platform used to navigate through it.

3 Corpus study

This section begins by addressing the process of data collection, from the sources to the methods used to gather the dialogue excerpts. Next, the steps for annotating the data together with a new taxonomy designed for this study are presented.

3.1 Data collection

In an initial exploratory phase, we originally searched for examples similar to those found by Varonis and Gass (1985), but these turned out not to be frequent in our material. Instead, we found numerous occurrences of interaction similar to those found by Svennevig (2018). However, the examples we found were also different from Svennevig's in an important respect. Since an investigator was typically present to provide language support, the learner did not need to go beyond verbal communication to ask for missing words; instead they could ask the investigator using their own first language. Hence, instead of embodied means of indicating a missing word, we found code-switching interactions.

In the data collection phase beyond the initial exploratory phase, we therefore collected dialogue excerpts where a production problem together with code-switching take place. We understand a production problem as those cases where learners fail to find the necessary term or expression in the target language. By code-switching we here mean that in order to get help to find the correct word, they switch to their mother tongue[6].

A number of search expressions were used to make the search for proper dialogue examples more efficient. Given that we are looking for situations where the learner is unable to provide a certain term or expression in the target language, hypothetically we could expect a question from the learner concerning the missing information. Hence, we used some sentences in both the target language (English) and the first languages (Italian, Spanish, Catalan) spoken by the subjects in the cited corpora that could serve to identify those potential examples: "how do you say", "come si dice", "cómo se dice", "com es diu". In addition, we searched for clarification ellipses, i.e. turns that repeated a word from the previous turn and that were understood as questions[7] using the regular expression:

```
*|^\1?$
```

This expression can be read as "A turn containing some thing (a word or expression), followed by another turn starting with that same thing followed by a question mark"[8]. The search process resulted in a collection of 40 suitable dialogue extracts.

[4] It should also be noted that in some cases we also find *two* investigators who complement each other in order to play the same role within the conversation. These examples, although scarce, were included in this study since they did not differ in structure (or otherwise) from the more common dyadic interactions. When including these examples, we did not distinguish between the two investigators.

[5] http://www.eecs.qmul.ac.uk/imc/ds/score.unstable/

[6] The term code-switching is generally defined as "the ability on the part of bilinguals to alternate effortlessly between their two languages" (Bullock and Toribio, 2009). However, in this study code-switching will not be associated with the subject of bilingualism since we deal with second language learners who are still far from becoming proficient in the target language. For this reason, a more appropriate definition of the linguistic phenomenon in the context of this project would be the process of alternating between the native language and the target language mainly due to an insufficient knowledge of the language being learned.

[7] Utterances interpreted by the transcriber as questions are transcribed as ending with "?".

[8] We originally searched for clarification ellipses to capture examples similar to those of Varonis and Gass (1985). However, the search string was also of help in identifying dialogue excerpts similar to those found by Svennevig (2018).

3.2 Data annotation

The target dialogue extracts were manually annotated using a taxonomy of dialogue acts that was created for the purpose of this work but based on previous related taxonomies (Varonis and Gass, 1985; Bondarenko, 2019; Howes et al., 2019; Myrendal, 2019)[9]. Table 1 below shows a detailed description of the annotation tags that make up the taxonomy. We use the following abbreviations:

- S, S1, S2: speaker
- INV: Investigator (teacher)
- SUB: Subject (learner)
- L1: learner's first language
- L2: target language that learner is acquiring
- M: word or phrase in L2 that learner is missing

A sample of 10 dialogue transcripts as well as a description of the annotation tags and some instructions were provided to two annotators in order to ensure inter-rater reliability. Fleiss' kappa test showed a score of 0.812 which indicates good agreement. An example of an annotated dialogue excerpt is shown in Table 2.

4 Formal model

Based on the annotated corpus of dialogue excerpts, we analysed dialogue act sequences looking for recurring patterns with the goal of providing a simple formal model, preferably in the form of a finite state automaton. We found that while a wide variety of dialogue act sequences were used to initiate the repair sequences, they thereafter largely followed a predictable pattern with some minor variations. We therefore split the formal model into two phases where the initial phase (Ask+Info) is separated out from the overall model.

As seen in Figure 2 the formal model presents a finite state automaton with a total of six states, with S and F being the initial and final state respectively. Each action performed by both the subject and the investigator represents the transition from one state to the next one.

State S to 1: The transition between the initial state and state 1 corresponds to the initial 'Ask + Info' phase, which includes some way of asking for a translation of a missing word, about which some information is provided (typically, it's L1 form). We will describe this phase further in Section 4.1. Lines 104 and 105 in the example in 2 above provide an example of behavior in this transition. Specifically, the subject is unable to find the word "traghetto" in English (i.e. "journey"), and consequently he/she asks "what's the name?" of the word in the target language.

State 1 to 2: In this transition the missing information is provided by the investigator as Table 2 shows in line 106 in 2.

State 2 to 3: The subject repeats the information (line 107 in 2) given in the previous state as a way to (a) practice the correct pronunciation and (b) reinforce the acquired knowledge and/or even (c) let the investigator know that the conversation can now proceed.

States 3 to 4 & 4 to 3: These transitions are an optional repetition of M by both the investigator and the subject. This behaviour can take place once or several times, as long as the investigator considers that this repetition is necessary for the subject's proper acquisition of the new information before the conversation can continue. Lines 108 and 109 in 2 illustrate the optional transitions.

State 3 to F: Once the subject successfully repeats the new information, the investigator accepts the subject's contribution to the conversation (line 110 in 2). This can be done overtly using a verbal acknowledgement, or silently[10]. At this point the main conversation is ready to resume (line 111 in 2).

4.1 First stage of the model, Ask+Info

Now we will focus on describing patterns in the 'Ask+Info' stage of the model, that is, dialogue act combinations observed in our dialogue extracts during the transition between the initial state and state 1.

After annotating our data, we found that some tag combinations were more common than others.We refer to the annotations of such sequences as compound tag. Table 3 collects all compound tags identified more than once in the data. It is important to note that the patterns presented here are only based on the dialogue excerpts analyzed in this study and therefore, other possible patterns that

[9]The dialogue act taxonomy used here makes a number of fine-grained distinctions that are beyond the scope of more general dialogue act annotation schemas like DAMSL (Core and Allen, 1997) or the ISO standard (Bunt et al., 2017).

[10]Because we are lacking video recordings of the interactions, we do not know if acknowledgement was provided gesturally, e.g. using a head nod.

Tag	Description	Example utterances
AskL1	SUB asks for word using L1	S: ¿Cómo se llama? <What is the name?> S: ¿Cómo es? <How is it?>
AskL2	SUB asks for a word using L2	S: What is the name?
IndAskL1	SUB indirectly asks for a word using L1	S: no sé el nombre <I don't know the name> S: no sé cómo se dice <I don't know how to say it>
IndAskL2	SUB indirectly asks for a word using L2	S: I don't know what the name is. S: I don't remember how to say the word.
SearchL2	SUB (unsuccessfully) searches for a word or phrase in L2	S: The price of food is... eer... is... S: I only read books and... er... S: We bought tomatoes and... mm...
ProvL1	SUB provides the L1 translation of M	S: ...mesa <table> S: ...¿niña? <girl?>
ProvL2	INV provides M in L2	S: It's called a table.
ProvDesL2	SUB describes M in L2	S: This thing you use for brushing your hair
Rep	INV repeats M	L: Mobile phone T: Mobile phone ←
Test	SUB tries pronouncing M	T: It's called a sprinkling can. L: Sprinkling can. ←
TestC	SUB tries using M in context	T: It's called a sprinkling can. L: We took the sprinkling can to water the plants. ←
Ack	S2 acknowledges previous utterance by S1	S1: We went to the park on Friday. S2: mhm

Table 1: Dialogue act annotation schema

Line	Speaker	Text	Annotation	Stage
104	SUB	yeah... and during the... traghetto <journey>.	SearchL2 + ProvL1	1
105		what's the name?	AskL2	1
106	INV	journey.	ProvL2	2
107	SUB	journey?	Test	2
108	INV	journey.	Rep	2
109	SUB	journey.	Test	2
110	INV	mm.	Ack	2
111	SUB	during the journey.	TestC	2

Table 2: Annotation of dialogue liean24i.1.cha

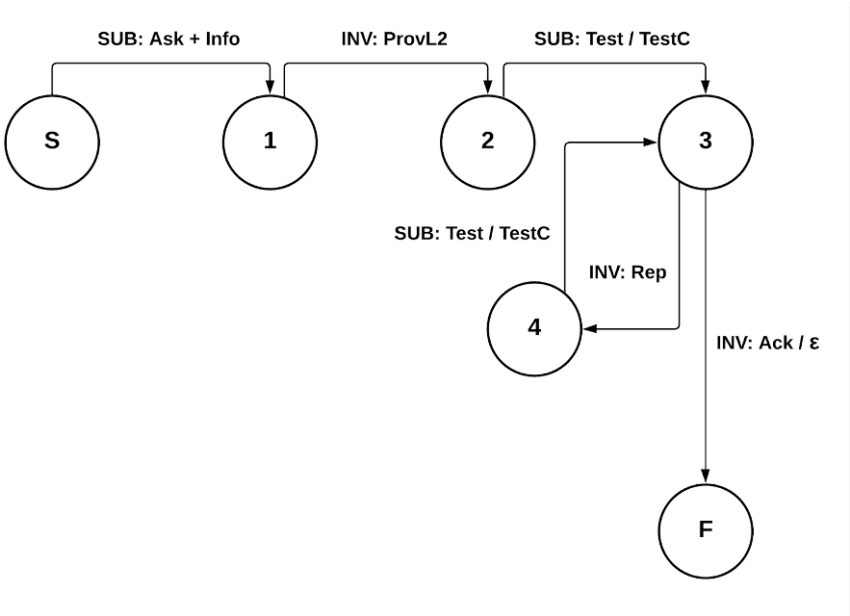


Figure 2: Final-state diagram representing a recurring pattern found in the data.

have not been observed in the data are not ruled out.

Most repair-initiating sequences in our data include signs of a production problems, usually in the form of hesitation sounds (“erm”, “err", “uh"). However, it also happens that the subject immediately asks a question without having attempted to produce an utterance first. Consequently, some patterns are very similar, with the only difference being whether they include this initial (unsuccessful) attempt to produce a whole utterance in the target language.

When asking for help with finding a word, the subjects in our data prefer to do it explicitly, as we have seen in previous examples (“how do you say...?”'). However, there are also instances where an indirect question is used instead (e.g. “I don’t know what it is called”). Interestingly, direct questions tend to be formulated in the learner’s L1, while indirect questions are frequently phrased in L2, the language being learned.

A distinction can be made between excerpts where the missing term/phrase is provided by the subject in the mother tongue (ProvL1) and those where there is no mention of it. This may be due to the word having been mentioned earlier in the conversation, or being inferable from the context. However, it is also possible that in these cases the learner relies on gestural cues (such as pointing at an object), similar to the behaviour described by Svennevig (2018). Indeed, such references are sometimes included in the transcriptions[11] (S=SUB, I=INV):

S: from the. whats name? [makes gesture for ground floor]
I: ground? ground floor ground floor.
S: ground floor mm.
I: okay good.

Some excerpts show how the subject might opt not to specify the missing word explicitly (no ProvL1) but instead using a verbal description:

S: so he didnt he didnt like it.
I: why?
S: no because my mm <pause> mh come si dice <*whats it called*> my principal my chief i dont know.
I: m boss.
S: my boss <pause> understand er if mm he you pay for me <pause> ...

4.2 Second stage of the model

As mentioned, there is a clear recurring pattern occurring from the moment the subject receives the requested information by the researcher, through the subject’s learning of such information, to the time the investigator acknowledges that the acquisition process is complete and the conversation can move forward. Table 4 collects these patterns and

[11] https://sla.talkbank.org/TBB/slabank/Multiple/ESF/EngItal/an/liean13g.cha

Compound tag	#	Example
SearchL2 + AskL1 + ProvL1	8	Young... <unclear> woman is erm... Come si dice ragazza alla pari? <*how do you say au pair girl?*>
AskL1 + ProvL1	5	Come si dice in inglese pioggia? <*how do you say rain in English?*>
ProvL1 + AskL1	3	Ah no pan, mayonesa, ¿cómo se dice? <*ah not bread, mayonnaise, how do you say it?*>
SearchL2 + ProvL1 + AskL1	2	I don't know maybe they they oh dio <*oh god*> <pause> rubare <*steal*> come si dice? <*how do you say it?*>
SearchL2 + IndAskL2	2	And er he er <pause> and him <breath> <pause> try to break the door but is impossible <pause> the black boy <pause> has one idea for go in the kitchen from er the window with one er <pause> I don't know the name.
SearchL2 + AskL2	2	From the <pause> what's name?
AskL1	2	Come si dice in inglese? <*how do you say it in English?*>
IndAskL1 + ProvL1	2	Non so come si dice piu basso <*I don't know how to say lower*>
SearchL2 + AskL1 + ProvDesL2	2	No because my mhm <pause> mhm come si dice? <*how do you say it?*> my principal my chief I don't know

Table 3: Collection of Ask+Info stage tag combinations that were found more than once in the dialogue extracts.

their frequency of appearance in the 40 dialogue extracts. The model fully (100%) covers the dialogue extracts.

The most repeated structure is the sequence ProvL2 + Test + Ack, present in 40% of the data.

S: and straight on in the <pause> street er <pause> the <pause> come si dice la strada principale <*how do you say the main road*>.
I: the main road. [ProvL2]
S: mh the main road. [Test]
I: mh. [Ack]

Slightly different to this pattern is ProvL2 + TestC + Ack (at 20%), where the subject is testing the new information in context. That is, the subject does not just repeat the provided information but uses it to continue the conversation:

S: mm <pause> <pause> ma mi scorde sempre come si dice la porta <*i always forget how to say door*>.
I: door. [ProvL2]
S: mm door <pause> <pause> er <pause> <pause> no open. [TestC]
I: yeah. [Ack]

We may note that in the top 5 patterns in phase 2, Test and TestC are equally frequent at 50% each (40+10% and 20+22.5+7.5%, respectively). Moreover, in 22.5% of dialogue extracts we find a TestC not followed by any verbal acknowledgement from the investigator. We may speculate that the lack of acknowledgement (which was observed much more often after TestC than after Test) could be related to the fact that the subject is demonstrating a correct acquisition of the new information by using it in an utterance, and therefore, if the investigator finds it satisfactory, overt verbal acknowledgement can be omitted.

S: <pause> <pause> come si dice <*what is it called*> three three.
I: <laugh> .
S: m eh <pause> m er.
I: steps. [ProvL2]
S: three steps. [TestC]
I: where?

5 Relation to Traum's (1994) model

It may be instructive to compare our model to Traum's (1994) finite state model of grounding. The model proposed here is to be seen as an amendment to Traum's model, rather than a replacement. Whereas Traum's model is intended as a general account for grounding in dialogue, we are only concerned with a special case.

A full summary of Traum's model is beyond the scope of this paper, but see Table 5 for the complete state transition diagram. For those familiar with the model, we want to point out that there seems to

State 1 - 2	ProvL2	ProvL2	ProvL2	ProvL2	ProvL2
State 2 -3	Test	TestC	TestC	Test	Test
State 3 - 4	-	-	-	Rep	Rep
State 4 - 3	-	-	-	Test	TestC
State 3 - F	Ack	Ack	ϵ	Ack	Ack
#	16	8	9	4	3
%	40%	20%	22.5%	10%	7.5%

Table 4: The 5 most common patterns found in the second stage of the model

be a fairly straightforward mapping of our special-purpouse dialogue acts to the more general ones in Traum's account:

- **SUB: Ask + Info → ReqRepair(I)**:
  Asking for a missing word seems to be a straightforward case of self-initiated (other-) repair. Such requests for repair are abbreviated *ReqRepair* in Traum's model, and *I* is the initiator of the utterance, corresponding to SUB in our model.
- **INV: ProvL2 → Repair(R)**:
  Providing the missing word seems to be a case of repair from the responder *R* (INV in our model).
- **SUB: Test/TestC in state 2 → Continue(I)**:
  Continue is used in Traum's model for continuing an utterance by providing further lexical material (words). Traum gives no particular import to continuations meant to test SUB's mastery of the problematic word, but we do.
- **INV: Rep → ReqRepair(R)**:
  Here, interestingly INV's repetition can be seen as a request from INV for SUB to provide (further) repair. The logic is that INV wants SUB to again repeat the problematic word to make sure SUB sufficiently masters the pronunciation.
- **SUB: Test/TestC in state 4 → Rep(I)**:
  Again, we distinguish testing a word from simply continuing speaking.
- **Ack(nowledgement)**:
  This works the same in both models, although we allow that the vocabulary training episode may end without explicit acknowledgement from the responder (INV)[12].

[12]Traum's model requires acknowledgement from the responder before a discourse unit (roughly, an utterance) can be considered complete. This is not incompatible with our model, as long as one admits that a vocabulary training episode may end before the final discourse unit involved in the episode is complete.

Next act	In state						
	S	1	2	3	4	F	D
Initiate(I)	1						
Continue(I)		1			4		
Continue(R)			2	3			
Repair(I)		1	1	1	4	1	
Repair(R)		3	2	3	3	3	
ReqRepair(I)			4	4	4	4	
ReqRepair(R)		2	2	2	2	2	
Ack(I)				F	1	F	
Ack(R)		F	F			F	
ReqAck(I)		1				1	
ReqAck(R)				3		3	
Cancel(I)		D	D	D	D	D	
Cancel(R)			1	1		D	

Table 5: Traum's (1994) finite state model of grounding

We leave a full integration of our model with Traum's for future work. However, we note that at least on one critical point, our model seems to go substantially beyond Traum's, namely with regard to when a request for repair by the initiator (SUB in our model) is allowed. Traum only allows ReqRepair(I) after some response from the responder R: "...we will also want to allow the possibility of a repair request *after some sort of response by the responder*." (ibid, p. 37; our italics). Concretely, this shows up as an empty space in state 1 for ReqRepair(I) in Tabel 5, meaning that this dialogue act is not allowed in this state. Only in state 2 to F, after a ReqRepair from R, is ReqRepair(I) allowed. In contrast, in our data we frequently find repair sequences initiated (using Ask+Info) by SUB (corresponding to I in Traum's model) without any preceding response from INV. Whether this is occurs also outside of vocabulary training interactions is a question for future research.

6 Implementation

The idea behind the implementation was to reproduce a dialogue strategy frequently observed in our data and embed it in a vocabulary training activity in the second language classroom. Through a conversation, the learner has the opportunity to put into practice the lexicon already acquired and/or even extend it. In this case, the dialogue focuses on vocabulary related to food where the main topic of the conversation revolves around what the learner has had for breakfast. By (verbally) interacting with the dialogue system, the learner is able to reinforce the acquired knowledge of the target language but also learn new lexical items.

For our implementation,[13] we used statecharts (Harel, 1987) which allows to describe the complex behavior of a system using an extended finite state notation. In addition, we chose to work with XState[14] for the model implementation. It is a JavaScript library designed to interpret finite state machines and statecharts in a way defined by Harel and W3C SCXML standard (Barnett et al., 2015). In our case it is a natural way to utilise the interactional structures that we discovered and expressed in a form of finite-state machine in a spoken dialogue system.

The fact that we deal with code-switching in our dialogue excerpts makes it crucial to rely on a bilingual ASR (Automatic Speech Recognition) so that the dialogue system can handle a conversation where the learner alternates between both the first and second language. This is made feasible by setting two individual ASRs with separate confidence scores, corresponding to the native and the target language. In our implementation, Spanish and English are taken as the user's first and target language respectively. At the moment the system can provide translations of single words from a predefined L1-L2 dictionary.

An example dialogue with the system could go as follows:

S: What did you have for breakfast?
U: I had toast with... cómo se dice queso?
S: Cheese.
U: Cheese?
S: Uh-huh. Did you have anything else?

We believe that the implemented model could be used as a tool in the language classroom for practising new words in the context of simulated everyday practical conversations such as making reservations, buying travel tickets, checking in at a hotel, etc..

[13] https://github.com/guscarrian/breakfast_demo

[14] https://xstate.js.org

7 Conclusions and future work

The main goal of the current study was to investigate dialogue strategies for vocabulary learning that could be found in second language learner corpora, and that could be useful in a dialogue system for second language training. The formal model encapsulates a general strategy used among learners at the time of acquiring new vocabulary in the second language, when in the presence of a teacher who can offer language assistance.

We found that learners ask for the L2 word/expression they need, either explicitly or implicitly. Additionally, code-switching occurs frequently as part of these requests. Once the new word/expression is introduced, learners work on repeating it as part of the strategy for acquiring new vocabulary.

The proposed model is based on data from students of English with a poor linguistic competence where the described production problems were common. The vocabulary building activity is designed for learners of a second language at early stages where linguistic support is often needed due to the lack of knowledge in the target language. However, whilst production problems may not be as frequent among advanced learners, the strategy described seems to be still applicable to any learner regardless of their level of linguistic competence. Ultimately, despite the fact that the findings of this study are relevant to the field of second language learning, other research areas could also benefit from them.

Future work includes extending the implementation to cover more variants of the patterns observed in the repair initiation (phase 1). We would also like to explore larger quantities of data provide an even stronger empirical footing. Also, conducting a human evaluation within the second language learning context would be a key component in future attempts to evaluate the model's performance. We would also like to confirm the applicability of the model to other language pairs, and in particular involving second languages other than English.

Acknowledgements

This work was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

J Barnett, R Akolkar, RJ Auburn, M Bodell, D Burnett, J Carter, S McGlashan, T Lager, M Helbing, R Hosn, et al. 2015. State chart XML (SCXML): State machine notation for control abstraction, W3C recommendation.

Anastasia Bondarenko. 2019. *Grounding of names in directory enquiries dialogue. A corpus study of listener feedback behaviour*. MA thesis, Master in Language Technology, Gothenburg University.

Barbara E Bullock and Almeida Jacqueline Ed Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the ISO 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer.

Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Tim Greer. 2013. Word search sequences in bilingual interaction: Codeswitching and embodied orientation toward shifting participant constellations. *Journal of Pragmatics*, 57:100–117.

David Harel. 1987. Statecharts: A visual formalism for complex systems. *Sci. Comput. Program.*, 8(3):231–274.

Christine Howes, Anastasia Bondarenko, and Staffan Larsson. 2019. Good call! grounding in a directory enquiries corpus. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers*.

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

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

Jan Svennevig. 2018. "what's it called in norwegian?" acquiring l2 vocabulary items in the workplace. *Journal of Pragmatics*, 126:68–77.

David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.

Evangeline Marlos Varonis and Susan Gass. 1985. Non-native/non-native conversations: A model for negotiation of meaning. *Applied linguistics*, 6(1):71–90.

How to repair a slip of the tongue?

Andy Lücking[1,2] **and Jonathan Ginzburg**[1]
[1]Université Paris Cité, CNRS,
Laboratoire de Linguistique Formelle (UMR 7110)
[2]Goethe University Frankfurt
luecking@em.uni-frankfurt.de, yonatan.ginzburg@u-paris.fr

Abstract

A slip of the tongue (SoT) is by no means a random occurrence and usually gets self-repaired immediately. The reparandum, however, remains available in context as potential anaphoric antecedent. So at least two puzzles for dialogue theory emerge: (i) how to deal with reparandum anaphora, and (ii) how is immediate repair possible? To provide answers, we make two extensions to the dialogue framework KoS (Ginzburg, 2012): Firstly, we spell out SoT repair as an "intra-utterance move" which utilizes a conversational rule drawing on *intended meaning*; Secondly, by reviewing current cognitive science work, we connect the linguistic types postulated by KoS to a pointer-based neurocognitive architecture and thereby sketch an explanatory dialogical model of SoT repair.

1 Introduction

Besides polished parlance, the domain of natural language use also knows slips of the tongue (SoT), or *lapsus linguae*. A well known example is senator Edward Kennedy's (1), transcribed here following Pincott (2012):[1]

(1) Our national interest ought to be to encourage the *breast* .. the best and the brightest

The SoT in (1) is a *substitution error* where the sound /r/ from *brightest* is anticipated and interferes the production of *best*, leading to an erroneously produced *breast*. However, the SoT is self-monitored and immediately repaired. The transcription used in (1) also exemplifies a methodological problem: why is the SoT transcribed as *breast* instead of, say, homophone *brest*? The reason very likely is just the joy of Freudian interpretations, a presumably dubious construal of speech errors which we do not follow further here—see also Cutler (1981) on "[t]he reliability of speech error data".

Unintentionally produced expressions such as SoTs are somewhat awkward for semantic theorizing: they are (arguably) not licensed by a grammar rule, nor are they part of the intended content of the to-be produced utterance.[2] They may also result in sounds which do not match the phonology of any word in the given language, although they virtually never violate the phonological constraints of that language (Wells, 1951). As a consequence they have been excluded from linguistic theory and competence as "grammatically irrelevant conditions" (Chomsky, 1965, p. 3). However, SoTs nonetheless influence turn-taking, other/self-repair and grounding, and are not on the whole arbitrary (Nooteboom, 1969; Harley, 2006). They can also figure as antecedents of anaphoric expressions: An addressee or overhearer of (1) can pick up the erroneously produced segment by means of a *Wh*-phrase or even a "salience anaphora" (cf. Asher and Wada, 1988).

(2) I heard *what you said first* / *it*.

The successful resolution of an anaphoric relation presupposes that the target *relatum* is available in context—in case of (2) and (1) this is the substitution error segment. Hence, for dealing with anaphora concerning SoTs, we need a notion of context that keeps track of lapses like of other speech items.

SoTs happen in every modality, be it spoken, written, or signed (Fromkin, 1980), but given the temporally detached communication mode in particular of writing, detected errors are usually erased right away—and even more easily so with electronic help—before any text is published.[3] Therefore it may be warranted to "idealize away" speech

[1]The corresponding video recording can be watched here: https://www.youtube.com/watch?v=SVJ0-cWr_PY, accessed 2nd May 2022.

[2]They *can* be used intentionally as part of, say, a joke, however.

[3]"In speaking, however, erasure is a physical impossibility,

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

errors from written, proof-read sentence-oriented grammars, but they are arguably impossible to ignore in (spoken) dialogue theory. Although it may not be part of a speaker's competence to *produce* speech errors, it *is* part of linguistic competence how to deal with them (see Ginzburg et al., 2014, p. 57 for a related argument concerning disfluencies). In fact, about one in three speech errors do get self-repaired (Levelt, 1983, p. 44). Furthermore, as also argued by Ginzburg et al. (2014), self-repaired speech errors pattern with other-repairs, a well-established type of clarification interaction (Schegloff et al., 1977). Hence, a unified account of self- and other-initiated repair needs to be provided by linguistic dialogue theory. We follow Ginzburg et al. (2014) in this respect, but here, following Postma et al. (1990), we distinguish SoTs from other cases of disfluencies/self-repair/self communication management:[4] while the former are proper speech errors in the sense that produced and planned speech diverge, the latter signal problems in the execution of a speech plan; differentiation between both may not be sharp, though.

Dependence on planned speech not only distinguishes SoTs from disfluencies in general, it also induces temporal constraints. Psycholinguistic research on speech lapses focuses on immediate repair (see Sec. 2). But repair detection can be delayed. An anonymous reviewer of SemDial came up with the following example:

(3) A: I think I'll wear my green dress. Can you bring it to me please?
B: OK [leaves to go get dress].
A: Wait, did I say green? Sorry, I meant my red dress.
B: OK, I'll get it. But your original choice was better.

We are somewhat dubious about whether this self-correction should be viewed as a SoT: we think this category should not include errors based on apparent *intention change*, as this one seems to be. We discuss one classification of speech errors below and hypothesize that SoTs do not felicitously allow for editing phrases like 'I didn't mean X', though drawing the line is clearly tricky.

In any case, it is clear that a repair can virtually be delayed for an arbitrary period of time: (speaking to Ann) "Did I really call you 'Barbara' last Christmas?". The temporal range of repair hence seem to be constrained by memory. In this regard, at least three temporal windows can be distinguished:

- immediate repair due to perceptual monitoring (Fig. 2) as in (1);
- repair within the reach of rehearsal of utterances within the phonetic loop within working memory (Baddeley, 2012), as in (3);
- referring to conversations which are stored as episodes within episodic memory (Ginzburg and Lücking, 2020) ("Barbara").

Of course, if a SoT remains unaltered (or undetected) without affecting the ongoing of the actual conversation, repair becomes superfluous; there is a decay of importance of repair, bound up with dialogical relevance. For this reason—memory issues aside—there is a strong prevalence of immediacy of repair. In fact, issuing non-immediate repair needs a special preparation to bring the reparandum into focus again—cf. the "Wait, did I say green?" phrase in (3). This is reminiscent of the pragmatic "one-moment-" or "just-a-minute-test" (Shanon, 1976, p. 248) for addressing presupposed contents. Hence, immediate SoT repair seems to be a uniform articulatory and time-bounded phenomenon which deserves a treatment on its own.

SoT repair usually is self-repair. This follows from its immediacy which is coupled to self-monitoring, but is also due to primary "editing rights" or even obligations of the speaker, as exemplified in (4), taken from a transcript of the TV show *Parks and Recreation*, where the addressee (Tom) claims a SoT concession from the speaker (Jerry):[5]

and its seeming social equivalent is only a polite convention that usually works only superficially. Nevertheless, all of us do try to cover up some of our lapses." (Hockett, 1967, p. 100)

[4] Terminology here involves important presuppositions. In generative linguistics and in NLP, it has been common to use the term 'disfluency' which carries the implication that the phenomena in question are somehow deviant from normal fluency. CA's term 'repair' makes the phenomenon more intentional, in line with works such as (Clark and FoxTree, 2002), which incorporate filled pauses into the lexicon. Allwood's term 'self communication management' goes the whole hog towards intentionalizing the phenomenon. The latter is, arguably, inappropriate for SoTs. We will mostly stick with 'repair', but occasionally use 'disfluencies' where the literature has already established this.

[5] https://tvquot.es/parks-and-recreation/quote/u7lrn5nc/, accessed 26th July 2022.

(4) JERRY: For my murinal, I was inspired by the death of my grandma.
TOM: [laughs] You said "murinal."
JERRY: No, I didn't.
ANN: Yes, you did. You said "murinal". I heard it.

We therefore formulate SoT repair as a speaker-independent linguistic resource below but acknowledge that it is mainly used (if at all) by the producer of a lapsus lingua.

In section 2 we briefly review some common types of SoTs. Section 3.1 introduces the basic ingredients of the formal dialogue theory KoS (Ginzburg, 2012), which is used in section 3.2 to adopt the analysis of backward-looking disfluencies from Ginzburg et al. (2014). Section 3.3 refines the previous analysis and bridges to a neural construal of SoT repair. The neural construal is taken up again in section 4, where simple networks are replaced with current semantic pointer-based architectures within spiking neuron populations. Inasmuch as slips of the tongue exemplify an interface phenomenon of dialogue theory and linguistic processing, a common analysis framework of this kind is needed.

2 Kinds of slips of the tongue

An enormous variety and detail of categories of speech errors has been observed (Crystal, 1997, p. 265). Pfau (2009) classifies the errors he found in his speech error corpus into four kinds (which in turn are partitioned into sub-kinds; examples are his):

- semantic anticipation or perseveration (e.g., substituting *potato* for *onion* or *vice versa*)
- errors involving feature mismatch (e.g., plural verb form following a singular subject noun phrase but which involves a plural genitive)
- stranding or shift of an abstract feature (e.g., perseveration of the plural feature onto a noun)
- errors involving accommodation (post error process where a follow-up error accommodates the error-induced context to grammatical constraints)

The first three classes roughly correspond to the most frequent error types Garnham et al. (1981) observed in the London-Lund corpus, namely substitution and anticipation at segment and word level. Some examples are collected in (5):

(5) a. "taddle tennis" instead of "paddle tennis" (segment, anticipation; Fromkin, 1973a, p. 112)
b. "I can't cook worth a cam" instead of "I can't cook worth a damn" (segment, perseveration; Fromkin, 1973a, p. 112)
c. "Seymour sliced the knife with a salami" instead of "Seymour sliced the salami with a knife" (word reversal; Fromkin, 1973b, Appendix)
d. "Take it out to the porch – eh – verandah." (word, substitution; Laver, 1969, p. 138)

We will therefore mainly focus on these kinds of SoT in the following.

Note that SoTs can also occur in sequence. Weir (2018) retells one of Nazbanou Nozari's—a cognitive scientist—stories about a research participant who was shown a picture of a sheep and called it "wolf". He corrected the incorrect classification to "steep" and then to "sleep". Remarkably, as pointed out by Weir (*op. cit.*), "'Wolf' is related to 'sheep' in meaning, 'steep' is related in sound, and 'sleep' in both meaning and sound." Hence, there are semantic and phonological crossover effects.

The given examples—exceptional cases aside—as well as received knowledge of SoTs show that they are a rather local phenomenon. Harley (2006, p. 740) provides a spot-on summary: "Sounds only exchange across small distances, whereas words can exchange across phrases; words that exchange tend to come from the same syntactic class, whereas sound exchange errors are not constrained in this way, but instead swap with words regardless of their syntactic class." This means that SoTs have to be accounted for *sentence-internally*. Accordingly, they get detected by monitoring mechanism during speech production (Hartsuiker and Kolk, 2001). The repair of a detected SoT follows a common pattern, which is, simplified from Levelt (1983, p. 45), shown in Fig. 1. Following an utterance which contains the *reparandum*, the repair is sometimes prepared by an editing phase[6] and provides the repairing expression, the *alteration*.

[6] In fact, in the corpus study of Switchboard by Hough (2015) fewer than 15% of self-repairs involved an editing phrase.

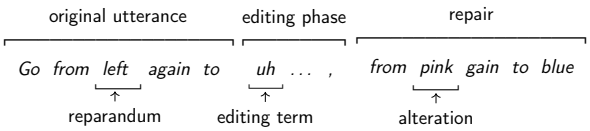


Figure 1: The structure of repair, simplified from Levelt (1983, p. 45).

Indicating correction with the editing phase can happen in several ways, including:

- the use of an editing term as in (5d) and Fig. 1; the editing term may also consist in the repetition of part of the original utterance before the reparandum;
- an aborted production during or after the reparandum and re-start, as in (1)—this is usually accompanied by stressed intonation starting with a guttural sound (Laver, 1969).

Repairing a SoT can happen as both, self-repair and other-repair, although correcting an unintentionally produced lapse of the dialogue partner may be regarded impolite, in particular when the intended utterance is recognized easily.

3 Formal model of SoT repair

3.1 Background: TTR and KoS

Given the intra-sentential domain of SoTs (see above), a highly incremental framework is needed. We use KoS (Ginzburg, 2012) in this respect. KoS is formulated in TTR (Cooper and Ginzburg, 2015; Cooper, 2022), a *Type Theory with Records*. Following the model of (perceptual) classification, a crucial notion of TTR is a *judgement*, this is in general that object a is of type T, notated $a : T$. More complex semantic issues such as an assertion that a situation is of a certain situation type draws on structured entities called *records* (token level) and *record types*. An assertion is then modelled in terms of an *Austinian proposition* as a judgement between *records* and *record types*:

(6) Austinian proposition :=
$$\begin{bmatrix} \text{sit} & : \mathit{Rec} \\ \text{sit-type} & : \mathit{RecType} \end{bmatrix}$$
and true iff $\mathit{Rec} : \mathit{RecType}$, i.e., the situation is of the type specified by the record type—in this case, *Rec* is a *witness* for *RecType*.

Linguistic parsing is construed along this way, too: sign types classify speech event tokens. Following work in *Head-Driven Phrase Structure Grammar* (Pollard and Sag, 1994), the basic sign architecture projects from lexical items to phrases and sentences and is as follows:

(7) *Sign* :=
$$\begin{bmatrix} \text{phon} & : \mathit{List}(\mathit{Phoneme}) \\ \text{cat} & : \mathit{SynCat} \\ \text{constits} & : \mathit{Set}(\mathit{Sign}) \\ \text{dgb-params} & : \mathit{RecType} \\ \text{cont} & : \mathit{SemObj} \end{bmatrix}$$

The *constits* field collects all daughter elements of a (complex) sign, the *dialogue gameboard parameters* (dgb-params) provide an interface to context. Work on dialogue brought about significant refinements of the structure of context extending beyond a speaker addressing an addressee at a given time and place with a speech event, leading to *dialogue gameboards* (DGB; Ginzburg, 2012):

(8) *DGBType* :=
$$\begin{bmatrix} \text{spkr} & : \mathit{Ind} \\ \text{addr} & : \mathit{Ind} \\ \text{utt-time} & : \mathit{Time} \\ \text{s-event} & : \mathit{Rec} \\ \text{c-utt} & : \text{addressing(spkr,addr,s-event,utt-time)} \\ \text{FACTS} & : \mathit{Set}(\mathit{Prop}) \\ \text{Pending} & : \mathit{List}(\mathit{LocProp}) \\ \text{Moves} & : \mathit{List}(\mathit{LocProp}) \\ \text{QUD} & : \mathit{PoSet}(\mathit{InfoStruc}) \end{bmatrix}$$

FACTS represent shared assumptions in terms of a set of propositions. Dialogue moves that are in the process of being grounded or under clarification are the elements of the *Pending* list. Already grounded moves are moved to the *Moves* list. Within *Moves* the first element has a special status given its use to capture adjacency pair coherence and is referred to as *LatestMove*. The current question under discussion is tracked in the *QUD* field. It is structured as a partially ordered set whose topmost element is called *MaxQUD*. QUD not only tracks a question, but also an antecedent focal expression, the *focus establishing constituent* (FEC), hence its contents are objects of type *InfoStruc*:

(9) *InfoStruc* :=
$$\begin{bmatrix} \text{q} & : \mathit{Question} \\ \text{FEC} & : \mathit{LocProp} \end{bmatrix}$$

The sign-based classification of a phonetic speech event is a special kind of Austinian proposition called *locutionary proposition* (*LocProp*), a record–record type-pair consisting of a speech

event u_0 as record and a sign as situation type in such a way that the *phon* value of the sign type correctly classifies the speech event and the entries within *dgb-params* are witnessed in a record w_0. Given these notational conventions, a record type of type *LocProp* has the following structure, where $s_0 = u_0 \wedge_{merge} w_0$ (i.e., s_0 is the *merge*, or unification, of u_0 and w_0):

(10) $$\begin{bmatrix} \text{sit} = s_0 : \mathit{Rec} \\ \text{sit-type} : \mathit{Sign} \end{bmatrix}$$

Updating an information state is licensed by conversational rules, pairs of DGBs of the form *preconditions* and *effects* (sometimes abbreviated as *pre* respectively *eff*):[7]

(11) $$\begin{bmatrix} \text{pre} & : \mathit{DGBType} \\ \text{effects} : \mathit{DGBType} \end{bmatrix}$$

For example, if a question is posed, this question becomes—under smooth development, but not invariably (Łupkowski and Ginzburg, 2017)—the current *question under discussion*:

(12) Ask QUD-incrementation :=:
$$\begin{bmatrix} \text{pre} : \begin{bmatrix} \text{q} & : \mathit{Question} \\ \text{LatestMove}{=}\mathit{Ask}(\text{spkr,addr,q}) & : \mathit{IllocProp} \\ \text{u}_{\text{fec}} \in \text{MaxPending.sit.constits} & : \mathit{LocProp} \end{bmatrix} \\ \text{eff} : \begin{bmatrix} \text{QUD}{=}\begin{bmatrix} \text{q}{=}\text{pre.q} \\ \text{FEC}{=}\text{u}_{\text{fec}} : \mathit{LocProp} \end{bmatrix} : \mathit{InfoStruc} \end{bmatrix} \end{bmatrix}$$

A DGB is the agent-specific structure of context which constitutes the publicized part of information states:

(13) *TotalInformationState* :=
$$\begin{bmatrix} \text{private} : \mathit{PRType} \\ \text{public} : \mathit{DGBType} \end{bmatrix}$$

Given this formal background, an account of SoT repair can be given.

[7] The pair of preconditions and effects notated as a single record type abstracts over deductive and temporal aspects: they are means to *classify* interactions. Seen as processing resources (a point raised by an anonymous reviewer), they can be regarded as functional types (*if* preconds, *then* effects).

3.2 Previous work on backward looking disfluencies

A repair can potentially occur at any place of an ongoing utterance. Hence the preconditions of a dialogical repair are rather weak, presupposing only that Pending is non-empty. Should we add as an additional condition divergence from intended production? This conflicts with repair that involves repetition—a highly pervasive phenomenon (Hough, 2015). In a probabilistic setting, which we are not assuming here, this condition could be formulated as insufficient confidence in the reparandum. In the absence of that, we will not include a divergence condition in the rule for backwards looking repairs, but explicate it in terms of a trigger stated at the level of the private cognitive state. Based on work on meaning-oriented clarification requests, giving rise to a class of conversation rules called *Clarification Context Update Rule* (Ginzburg, 2012), the (potentially accommodated) MaxQUD of the eff(ect) of the repair resource amounts to the issue of *What did the speaker mean by u_{fec}?* This MaxQUD requires the next (if an editing phrase has been produced) or simultaneous (without editing phrase) move (the new LatestMove) to provide an answer—an utterance which is *co-propositional* with u_{fec}. This has been formalized as *Backward Looking Appropriateness Repair* by Ginzburg et al. (2014, p. 42):

(14) $$\begin{bmatrix} \text{pre} : \begin{bmatrix} \text{spkr} : \mathit{Ind} \\ \text{addr} : \mathit{Ind} \\ \text{Pending}{=}\langle p0, \text{rest}\rangle : \mathit{List}(\mathit{LocProp}) \\ \text{u}_{\text{fec}} : \mathit{LocProp} \\ \text{c1} : \text{member}(\text{u}_{\text{fec}}, p0.\text{sit.constits}) \end{bmatrix} \\ \text{eff} : \begin{bmatrix} \text{MaxQUD} = \\ \begin{bmatrix} \text{q}{=}\lambda x?\mathit{Mean}(\text{pre.spkr,pre.u}_{\text{fec}}, x) \\ \text{FEC}{=}\text{u}_{\text{fec}} \end{bmatrix} \\ \quad : \mathit{InfoStruc} \\ \text{LatestMove} : \mathit{LocProp} \\ \text{c2} : \mathit{CoProp}(\text{LatestMove}^{\text{cont}}, \text{MaxQUD}) \end{bmatrix} \end{bmatrix}$$

(The superscript "cont" abbreviates the path LatestMove.sit-type.cont.) What does it mean that a question (MaxQUD) and a general semantic object ($\text{LatestMove}^{\text{cont}}$; including individuals, properties, propositions) are co-propositional? Ginzburg et al. (2014, p. 30) provide the following characterisation in terms of "answerhood" (where "utterances" denotes the range of expressions from fragments to full sentences):

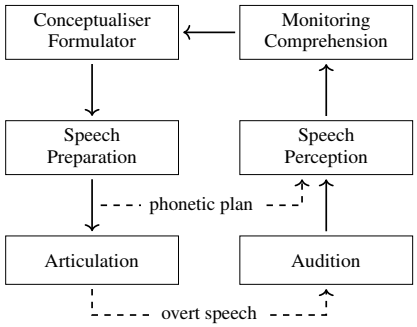


Figure 2: Dual perception loop (simplified from Levelt, 1989; Levelt et al., 1999).

(15) *Co-propositionality*

a. Two utterances u_1 and u_2 are *co-propositional* iff the questions q_1 and q_2 they contribute to QUD are co-propositional.

b. q_1 and q_2 are co-propositional if there exist a record r such that $q_1(r) = q_2(r)$.

That is, an utterance is co-propositional to an intended meaning MaxQud if the question projected from the utterance provides the same result (i.e., answer) when applied to a given record as MaxQud.

The rule in (14) already allows to analyze SoTs: an issue with the reparandum u_{fec} has come up as MaxQUD. The content of the repair provides the alteration of the reparandum. In the aftermath of repair, Pending has to be modified in such a way that within the original utterance the alteration substitutes for the reparandum (*Pending replacement*).

3.3 Application

Since a speech error happens if the verbal utterance diverges from the planned one, speech error correction can occur if the divergence is detected. According to the *Dual Perception Loop* model (Levelt, 1989; Levelt et al., 1999) self-monitoring happens on two routes: the intended utterance is compared to both its phonetic plan ("inner speech") and to the perceived speech output, see Fig. 2.[8] This can be modelled in a pretty straightforward manner by incorporating a phonetic plan into the private share of interlocutors' total information states (*PRType*).

[8]Since in aphasic patients a dissociation between comprehension and error-detection ability has been observed, there is evidence that the fast inner loop does not rest on an internalized speech comprehension monitor but rather uses production signals (Nozari et al., 2011).

(16) $PRType \sqsupset$
$\left[\text{PhonPlan} : List(RecPhon)\right]$

where *RecPhon* is a reduced variant of a locutionary proposition which consists of a (mental) speech event and its phonological classification (according to the model sketched in Fig. 2, syntactic and semantic aspects pertain to the Conceptualiser and Formulator levels).

(17) $RecPhon :=$
$\begin{bmatrix}\text{s-event} & : Rec \\ \text{phon-struc} & : Sign.\text{phon}\end{bmatrix}$
$\sqsubset LocProp$

A observes a speech error iff A.private.PhonPlan.i $\not\sqsubseteq$ A.public.Pending.i, for any list element with index i which is appended to the incrementally increasing list of *RecPhon*s respectively *LocProp*s.

Let us apply these tools in order to analyze example (1). The original utterance before the SoT occurs consists in the speech event e_0 = "Our national interest ought to be to encourage". The parse up to this point (speaking in terms of HPSG) has found an NP ("Our national interest") and an incomplete verb cluster headed by "ought", but the argument structure of "encourage" still requires an NP argument. If an NP argument follows, the verb cluster can be completed by means of a *head-argument-structure* and finally combined with the subject NP into a *head-filler-structure*. We abbreviate the chart loosely following Ginzburg et al. (2020) as T_{natint} as follows, including found (fnd) and still req(uired), anticipated information:

(18) $T_{\text{natint}} =$
$\begin{bmatrix}
e_1 : [\text{Our national interest ought to be to encourage}] \\
e_2 : [\text{Our national interest}] \\
e_3 : [\text{encourage}] \\
e_4 : \begin{bmatrix}
\text{fnd1}{=}e_2 : Sign.\text{cat}{=}\text{NP} \\
\text{fnd2}{=}e_3 : Sign.\text{cat}{=}\text{V} \\
\text{req1}{=}\langle\text{NP,head-arg-struc}\rangle : GramStruc \\
\text{req2}{=}\langle\text{head-cluster-struc}\rangle : GramStruc \\
\text{req3}{=}\langle\text{head-filler-struc}\rangle : GramStruc
\end{bmatrix}
\end{bmatrix}$

Since the chart type in (18) mentions still missing grammatical structures (type *GramStruc*; we only listed the ones needed for the example) it generates hypothesis about its continuation and can be used to construct *Pending* and *QUD* simultaneously and incrementally. K(ennedy)'s DGB therefore can be classified as (19).

(19) K.dgb1 =

$$\begin{bmatrix} \text{spkr}= \text{K} : \mathit{Ind} \\ \text{addr} : \mathit{Set}(\mathit{Ind}) \\ \text{s-event}= e_0 \\ \text{Pending}=\left\langle \begin{bmatrix} \text{sit}= e_0 \\ \text{sit-type}= T_{\text{natint}} \end{bmatrix} \right\rangle \\ \text{Moves}=\langle\,\rangle \\ \text{QUD}=\left\langle \begin{bmatrix} \text{q}=?\text{MaxPending} \\ \text{FEC}=\begin{bmatrix} \text{sit}=[\text{national interest}] \\ \text{sit-type}=\begin{bmatrix} \text{phon}=\langle/\text{national}/, \\ /\text{interest}/\rangle \\ \text{cat}=\text{NP} : \mathit{SynCat} \\ \text{cont} : \mathit{Ind} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

Up to this point Kennedy's *PhonPlan* is satisfied, which continues as in (20):

(20) $\left[\text{K.private.PhonPlan}=\left\langle \text{the,best,and,the,brightest} \right\rangle\right]$

The definite article, the next utterance token, complies with both the PhonPlan and the NP requirement of the chart type. The utterance of the future reparandum could in principle be a noun and therefore complete the sentence:

(21) K.dgb2 =

$$\begin{bmatrix} \text{spkr}= \text{K} : \mathit{Ind} \\ \text{addr} : \mathit{Set}(\mathit{Ind}) \\ \text{s-event}= e_0 \\ \text{Pending}=\left\langle \begin{bmatrix} \text{sit}= e_0 \\ \text{sit-type}= T_{\text{enc}} \end{bmatrix} \right\rangle \\ \text{Moves}=\langle \mathit{Assert}(\text{K},\begin{bmatrix} \text{sit}= e_0 \\ \text{sit-type}= T_{\text{enc}} \end{bmatrix})\,\rangle \\ \text{QUD}=\left\langle \begin{bmatrix} \text{q}=?\text{MaxPending} \\ \text{FEC}=\begin{bmatrix} \text{sit}=[\text{national interest}] \\ \text{sit-type}=\begin{bmatrix} \text{phon}=\langle/\text{national}/, \\ /\text{interest}/\rangle \\ \text{cat}=\text{NP} : \mathit{SynCat} \\ \text{cont} : \mathit{Ind} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle \\ \text{FACTS}=\left\{ \mathit{Classify}(T_{\text{enc}}, e_0) \right\} \end{bmatrix}$$

where T_{enc} is a sentence parse:

(22) $\begin{bmatrix} \text{e}_1 : [\text{Our national interest ought to be to encourage} \\ \quad \text{the bre(a)st}] \\ \text{e}_2 : \left[\text{fnd}=\text{e}_2 : \mathit{Sign}.\text{cat}=\text{S}\right] \end{bmatrix}$

However, the classification of the utterance e_0 by the type T_{enc} is unsatisfying:[9] it either involves a "novel" noun (*brest*), or it is semantically awkward (*breast*). Furthermore, there is a mismatch between K.private.PhonPlan and K.public.Pending following the definite article. Hence, an accommodation of *Backward Looking Appropriateness Repair* is triggered, leading to an update of K.dgb.QUD. The question *What did K mean by "breast"?* (or "*brest*") becomes MaxQUD and has to be addressed first: the following LatestMove—*best and brightest*—is constrained to provide a co-propositional value.

(23) K.dgb3 =

$$\begin{bmatrix} \text{Pending}=\left\langle \begin{bmatrix} \text{sit}= e_0 \\ \text{sit-type}= T_{\text{enc}} \end{bmatrix} \right\rangle \\ \text{u}_{\text{fec}}=\begin{bmatrix} \text{sit}=[\text{the breast}] \\ \text{sit-type} : \mathit{Sign} \end{bmatrix} \\ \text{c1} : \text{member}(\text{u}_{\text{fec}},\text{MaxPending.sit.constits}) \\ \text{Moves}=\langle \mathit{Assert}(\text{K},\mathit{Mean},\text{u}_{\text{fec}}, \\ \quad \begin{bmatrix} \text{sit}=[\text{the best and the brightest}] \\ \text{sit-type}= N_{\text{bab}} : \mathit{Sign} \end{bmatrix})\,\rangle \\ \quad \begin{bmatrix} \text{q}= \lambda x ?\mathit{Mean}(\text{K},\text{u}_{\text{fec}},x) \\ \text{FEC}=\text{u}_{\text{fec}} \end{bmatrix}, \\ \text{QUD}=\left\langle \begin{bmatrix} \text{q}=?\text{MaxPending} \\ \text{FEC}=\begin{bmatrix} \text{sit}=[\text{national interest}] \\ \text{sit-type}=\begin{bmatrix} \text{phon}=\langle/\text{national}/, \\ /\text{interest}/\rangle \\ \text{cat}=\text{NP} : \mathit{SynCat} \\ \text{cont} : \mathit{Ind} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle \\ \text{FACTS}=\left\{ \begin{matrix} \mathit{Classify}(T_{\text{enc}}, e_0) \\ \mathit{Classify}(T_{\text{bab}},\text{u}_{\text{fec}}) \end{matrix} \right\} \end{bmatrix}$$

(Where T_{bab} is the nominal sign type classifying the conjunct *best and brightest*.)

In the aftermath, and if the self-repair is accepted, *Pending Replacement* applies (Ginzburg et al., 2014), leading to a substitution of e_0 and T_{enc} according to the following re-parse:

(24) $\begin{bmatrix} \text{e}_1 : [\text{Our national interest ought to be to encourage} \\ \quad \text{the best and the brightest}] \\ \text{e}_2 : \left[\text{fnd}=\text{e}_2 : \mathit{Sign}.\text{cat}=\text{S}\right] \end{bmatrix}$

Note that the SoT remains in FACTS, from which it can be retrieved as a constits element of T_{enc}, providing, for instance, an antecedent for reparandum anaphora.

The analysis of the SoT from (1) mainly in (23) recognizes two sources: the unlikely sign-based classification of the reparandum on the one

[9] One can express this by assigning the *Classify* relation a probability threshold (Cooper et al., 2015).

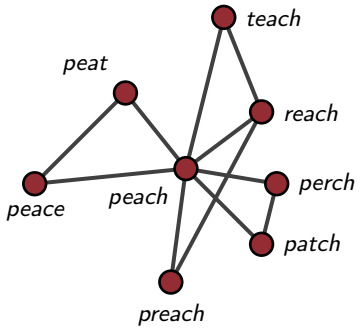


Figure 3: Extract of the phonological network around *peach* (adapted from Vitevitch et al. 2015, p. 32).

hand,[10] and the divergence of phonetic plan and self-monitored speech on the other hand. The *PhonPlan* accesses the mental state of an interlocutor, at least symbolically. We used it to exemplify a formal *post hoc* analysis of SoT repair. However, the non-arbitrary nature of *lapsus linguae* (see section 2) has been explained by activation-spreading models of sentence production (Dell, 1986), which also provide clues for their self-repair (Nooteboom and Hugo, 2020). The underlying rationale can be exemplified by means of a simple, phonetic example in Fig. 3: a word form like *peach* is phonetically similar to *teach* (the only phonetic difference is that the first is produced with an initial bilabial, the latter with an alveolar). Both sounds have not much in common with, say *apple* (although since apples and peaches are both fruits, they are associated in a semantic network). Hence, phonetic distances give rise to a phonological network. Since no exchange of content words with function words have been observed (Harley, 2006, p. 740), we assume that such networks are sorted according to part of speech. Now, if *peach* is to be articulated (cf. Fig. 2) it receives activation. This activation distributes to the neighboring nodes, however, which get co-activated. This co-activation may then lead to choosing the neighboring instead of the planned word and sending it to the articulator.

To summarize:

- How is it that the dialogue proceeds with the corrected utterances but the reparandum is still available as an antecedent? This is because SoT repair amounts to Pending replacement but the original utterance is still available in FACTS.
- Why is the phrase *the best and the brightest* interpreted as a reparandum? In fact, K. could also have uttered the conjunct NP *the breast, the best and the brightest*. Following work on SoT (e.g. Harley, 2006), we assume that the correction interpretation follows from a specific intonation.
- How is the SoT repaired so quickly and seamlessly? Besides the information encoded in the PhonPlan, Nooteboom and Hugo (2020) found evidence that co-activated items not only are the source of lapses but also a cue for their repair: the alteration will also be a node with high activation and therefore more easily accessible.[11]

While the activation spreading model provides crucial explanations for various kinds of *lapsus linguae*, it represents words—either as phonetic forms (segment errors) or as semantic markers ("Freudian" substitutions)—as single nodes. When construed neurally, this is a simplification: lexical items do not correspond to single neurons. Drawing on cognitive science insights, the contour of an integrative framework is emerging, which is sketched in the following section.

4 Activation spreading and semantic pointers

Following recent work in neurocognition, we assume that a *semantic pointer* is a notion that provides a needed level of abstraction within neuronal architectures (see Blouw et al., 2016 for a cognitive science summary). A semantic pointer is a compressed activation of neuronal spiking which is associated with a more elaborate region of neuronal activation, as illustrated in Fig. 4.

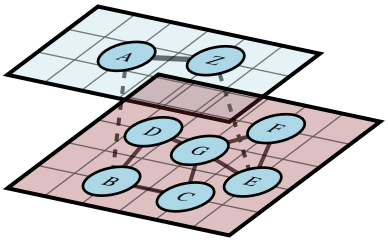


Figure 4: Semantic pointer (top layer): compressed activation of neuronal spiking (bottom layer).

[10] See Oliphint (2022) for a recent metaphysical account of words and the problem of distinguishing them from "non-words".

[11] Since the planned item is available in the PhonPlan and guides the monitoring process, retrieval loops are prevented, as remarked by an anonymous reviewer.

Visual processing, for instance, rests on a large population of spiking neurons which "encode" the visual input.[12] Via neuronal transformations (mathematically modelled as circular convolutions; Eliasmith, 2013) condensed levels of neuronal activation are produced from these large activation patterns which use a (much) less number of spiking neurons.[13] Further processing is sped up by using these abstract semantic pointers. However, the more detailed activation patterns can still be retrieved from the compressed encoding—hence the term semantic *pointer*. Semantic pointers can bind together various levels of activations, such as lexical, perceptual, and motor information. Linguistic forms—(mental representations of) labels such as phonetic strings—can be construed as semantic pointers as well.[14] Such "labeled semantic pointers" provide objects of fast linguistic processing, which may be unpacked in cognitive simulations (Connell, 2019; Goucha et al., 2017). Activation spreading from abstract labels to motor activation patterns implements embodiment (Mahon, 2015). A well-established example of a pointer-based model is the *hippocampal indexing theory* of Teyler and Rudy (2007). The hippocampus captures neocortical activity generated by an observed episode and projects back to these neocortical regions—hence, the hippocampus creates an index which can be unpacked to the full pattern of neocortical activity produced by the episode.

Instead of single nodes representing linguistic forms or meanings, we construe the nodes of such networks as semantic pointers, condensed levels of activation which can be unpacked by larger populations of neuronal spiking from which they are abstracted in the first place. SoT repair as outlined here and forward-looking disfluencies analysed by Ginzburg et al. (2014) appear as two sides of the same coin: a SoT is the result of too much, a disfluency of too little activation—probably everyone has experienced the latter as a tip-of-the-tongue feeling.

[12] But see Brette (2019) for a critical assessment of the encoding metaphor.

[13] Activation patterns are modeled in terms of mathematical vectors and synaptic weights. Now two vectors can be combined into a single one of the same dimensionality and later decomposed again. Hence there is some commonality of brain-based semantic pointer convolutions and popular data-based Deep Learning methods (Rasmussen, 2019).

[14] This view squares with the coordinative role of material symbols in cognition as argued by Clark (2006).

5 Discussion

We offer an account of SoT repair which rests on the notions of co-propositionality and intended meaning clarification from previous work on disfluencies in dialogue (Ginzburg et al., 2014). This account solves two linguistic puzzles which arise from correcting a lapse in a principled way: The utterance containing the reparandum is available within the assumptions shared by the interlocutors (FACTS—since the tongue slipped as a matter of fact) as an antecedent of reparandum anaphora; the repaired move including the alteration becomes MaxPending (the topmost move within the list of pending ones) and contributes to further dialogue progressing. The alteration is immediately retrievable since it is a pre-activated item. We noted that a repair is indicated by an explicit editing phrase, or by a specific intonation pattern, which signal that an utterance provides alterations (expressed by the meaning-pertaining question under discussion (QUD) $\lambda x?Mean(A,u_{fec},x)$) instead of, say, just continuing a dialogical exchange (incorporating phonetic repair-indicating details still has to be worked out, however). It has also been observed that SoTs follow phonetic or semantic constraints and that repair happens on very small time scales, virtually immediately. Psycholinguistic models provide explanations for these observations, mainly in terms of spreading activation architectures. In this respect it has been sketched how TTR types representing signs and *LocProp*s can be construed as labeled semantic pointers that compress larger populations of spiking neurons and are compatible with activation spreading. There is evidence that repair, and immediate repair, is part of dialogical competence (cf. sections 1 and 2). In order to provide an explanation not only of the semantic but also of the temporal aspects of this competence—cf. *Did I say X?* and the failure marking *I meant X* editing phrase for delayed repair—we think that formal models of meaning in dialogue eventually need to draw on processing models. This becomes much more pressing when considering multimodal interaction (which is the default form of dialogue): here temporal alignment of communicative means of various channels occurring both sequentially and (partially) simultaneously give rise to timing as an aspect of interaction *sui generis* (e.g. Lücking and Ginzburg, 2020; Rieser and Lawler, 2020). Timing in language, however, seems to be inextricably bound up with processing.

Using previous work on repair in general, we tried to develop formal tools for analysing SoTs which take the immediacy of repair into account. Besides their analytical properties, the formal tools make the case for a multi-level implementation. The symbolic dialogue theory—if construed cognitively—allows to represent linguistic, label-based compositional processing in a precise way, which is somehow carried out within speakers' brains (Frankland and Greene, 2020). However, linguistic labels are underpinned by statistical associations which help to speed-up processing (Connell, 2019)—but also may lead to production lapses, as reviewed in Sec. 3.3. Following work by Eliasmith (2013) and colleagues, these statistically associated labels can be construed as semantic pointers (Sec. 4); they can be unpacked to retrieve fully grounded semantic models, namely when condensed pointer-based representations in one brain area lead to replay (e.g., in recollection) or simulate (e.g., embodied sensori-motor processing) full representations usually from another brain area (Louwerse and Connell, 2011). The latter is involved in memory-based repair of stored episodes in contrast to immediate SoT repair, as mentioned in Sec. 1. In this sense, our formal, dialogical model of SoT repair suggests a specific interaction of statistical and symbolic semantic approaches, because both seem to target quite different aspects of meaning (Westera and Boleda 2019; see also Lücking et al. 2019): formal semantics provides the analytic backbone for defining semantic ontologies and providing scientifically precise, cognitively potent content representations, statistical regularities add inter-label associations which are important to capture temporal aspects of processing and understanding—and for producing *lapsus linguae* in the first place. To this we add semantic pointers to connect labels to the brain and to distinguish linguistic processing short-cuts from full mental simulations.

On a more general level this means that a (renewed) cooperation of semantics and cognitive science is required. Cognitive science develops processing models, but semantics and pragmatics contribute a precise structuring of the contents and contexts involved in processing. We think that formal dialogue theory, in particular KoS with its focus on spoken language, provides a useful semantic framework in this respect: KoS is already formulated in a way that is close to speech processing models (cf. notions such as LocProp and Pending) and fused with a WM model (Ginzburg and Lücking, 2020). Recent cognitive science work on the other hand seems to narrow down the gap between symbolic and neuronal levels of computation. Phenomena such as SoTs live on the interface of those levels and therefore are a lens into neuronally grounded dialogue semantics.

So how to repair a slip of the tongue? From all the co-activated items, retrieve the alteration which is co-propositional to the focal reparandum and complies with the inner PhonPlan, produce the alteration after an editing sound or phrase, move the original utterance to FACTS, and apply Pending Replacement to substitute alteration for reparandum.

Acknowledgments

We acknowledge the support by a public grant overseen by the French National Research Agency (ANR) as part of the program Investissements d'Avenir (reference: ANR-10-LABX-0083). It contributes to the IdEx Université Paris Cité ANR-18-IDEX-0001. We thank the SemDial reviewers for very helpful comments.

References

Nicholas Asher and Hajime Wada. 1988. A computational account of syntactic, semantic and discourse principles for anaphra resolution. *Journal of Semantics*, 6(1):309–344.

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

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

Romain Brette. 2019. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 42. Published online by Cambridge University Press: 16 July 2018.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50th anniversary edition edition. MIT Press, Cambridge, MA.

Andy Clark. 2006. Material symbols. *Philosophical Psychology*, 19(3):291–307.

Herb Clark and Jean FoxTree. 2002. Using uh and um in spontaneous speech. *Cognition*, 84:73–111.

Louise Connell. 2019. What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, 34(10):1308–1318.

Robin Cooper. 2022. *From perception to communication: An analysis of meaning and action using a theory of types with records (TTR)*. Oxford University Press.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology – LiLT*, 10:1–43.

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

David Crystal. 1997. *The Cambridge Encyclopedia of Language*, 2 edition. Cambridge University Press, Cambridge, UK.

Anne Cutler. 1981. The reliability of speech error data. *Linguistics*, 19(7-8):561–582. Guest editorial.

Gary S. Dell. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3):283–321.

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

Steven M. Frankland and Joshua D. Greene. 2020. Concepts and compositionality: In search of the brain's language of thought. *Annual Review of Psychology*, 71(1):273–303. PMID: 31550985.

Victoria Fromkin, editor. 1980. *Errors in Linguistic Performance*. Academic Press, New York.

Victoria A. Fromkin. 1973a. Slips of the tongue. *Scientific American*, 229(6):110–117.

Victoria A. Fromkin, editor. 1973b. *Speech Errors as Linguistic Evidence*. Mouton, The Hague and Paris.

Alan Garnham, Richard C. Shillcock, Gordon D. A. Brown, Andrew I. D. Mill, and Anne Cutler. 1981. Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics*, 19:805–817.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, UK.

Jonathan Ginzburg, Robin Cooper, Julian Hough, and David Schlangen. 2020. Incrementality and HPSG: Why not? In Anne Abeillé and Olivier Bonami, editors, *Constraint-Based Syntax and Semantics: Papers in Honor of Danièle Godard*. CSLI Publications, Stanford, CA.

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

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

Tomás Goucha, Emiliano Zaccarella, and Angela D. Friederici. 2017. A revival of *Homo loquens* as a builder of labeled structures: Neurocognitive considerations. *Neuroscience & Biobehavioral Reviews*, 81:213–224. The Biology of Language.

T. A. Harley. 2006. Speech errors: Psycholinguistic approach. In Keith Brown, editor, *Encyclopedia of Language & Linguistics*, 2 edition, pages 739–745. Elsevier, Oxford.

Robert J. Hartsuiker and Herman H. J. Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology*, 42(2):113–157.

Charles F. Hockett. 1967. Where the tongue slips, there slip I. In (Fromkin, 1973b), pages 93–119.

Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary, University of London.

John D. M. Laver. 1969. The detection and correction of slips of the tongue. In (Fromkin, 1973b), pages 132–143.

Willem J. M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38.

Max Louwerse and Louise Connell. 2011. A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35(2):381–398.

Andy Lücking, Robin Cooper, Staffan Larsson, and Jonathan Ginzburg. 2019. Distribution is not enough – going Firther. In *Proceedings of Natural Language and Computer Science*, NLCS 6.

Andy Lücking and Jonathan Ginzburg. 2020. Towards the score of communication. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue*, SemDial/WatchDial.

Paweł Łupkowski and Jonathan Ginzburg. 2017. Query responses. *Journal of Language Modelling*, 4(2):245–292.

Bradford Z. Mahon. 2015. What is embodied about cognition? *Language, Cognition and Neuroscience*, 30(4):420–429. PMID: 25914889.

Sieb G. Nooteboom. 1969. The tongue slips into patterns. In (Fromkin, 1973b), pages 144–156.

Sieb G. Nooteboom and Quené Hugo. 2020. Repairing speech errors: Competition as a source of repairs. *Journal of Memory and Language*, 111.

Nazbanou Nozari, Gary S. Dell, and Myrna F. Schwartz. 2011. Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, 63(1):1–33.

Jared S. Oliphint. 2022. Using a two-dimensional model from social ontology to explain the puzzling metaphysical features of words. *Synthese*, 200.

Roland Pfau. 2009. *Grammar as Processor*. Number 137 in Linguistik Aktuell / Linguistics Today. John Benjamins, Amsterdam and Philadelphia.

Jena E. Pincott. 2012. Slips of the tongue. *Psychology Today*.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, CA.

Albert Postma, Herman Kolk, and Dirk-Jan Povel. 1990. On the relation among speech errors, disfluencies, and self-repairs. *Language and Speech*, 33(1):19–29. PMID: 2283918.

Daniel Rasmussen. 2019. NengoDL: Combining deep learning and neuromorphic modelling methods. *Neuroinformatics*, 17:611–628.

Hannes Rieser and Insa Lawler. 2020. Multi-modal meaning – an empirically-founded process algebra approach. *Semantics and Pragmatics*, 13(8):n/a.

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

Benny Shanon. 1976. On the two kinds of presuppositions in natural language. *Foundations of Language*, 14(2):247–249.

Timothy J. Teyler and Jerry W. Rudy. 2007. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus*, 17(12):1158–1169.

Michael S. Vitevitch, Rutherford Goldstein, and Elizabeth Johnson. 2015. Path-length and the misperception of speech: Insights from network science and psycholinguistics. In Alexander Mehler, Andy Lücking, Sven Banisch, Philippe Blanchard, and Barbara Job, editors, *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 29–45. Springer, Berlin and New York.

Kirsten Weir. 2018. Lab work: The secrets behind slips of the tongue. *Monitor on Psychology*, 49(3).

Rulon Wells. 1951. Predicting slips of the tongue. In (Fromkin, 1973b), pages 82–87.

Matthijs Westera and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics – Long Papers*, pages 120–133, Gothenburg, Sweden. Association for Computational Linguistics.

Participants seek shared outlooks in non-canonical disagreements: Evidence from a corpus of dyadic conversation in English

John Duff and **Lalitha Balachandran**
Department of Linguistics
UC Santa Cruz
jduff, lalithab@ucsc.edu

Abstract

We present data and a preliminary analysis of a novel kind of disagreement observed in a corpus of English dyadic conversations. In conversations about a variety of topics, speakers volunteer attitude and speech reports rather than direct answers to Questions Under Discussion. The conversations are challenging to capture with formal pragmatic models, not just because of the mismatches between what is said and the QUD, but also because they lead to apparent disagreements despite a lack of contradiction. We analyze these cases as participants aiming to align non-committal stances, inspired by recent approaches that treat subjective conversation as the coordination of outlooks. Overall, the discussion advocates for flexible models of conversation that allow extra-linguistic goals and pressures to interface with the lower-level dynamics of discourse moves.

1 Introduction

The standard approach to modeling conversations between pairs of participants in formal semantics and pragmatics has involved the notion of *joint commitment*. Conversational participants make and invite assertions with the goal of growing the common ground, i.e. the list of discourse commitments held by both participants (Stalnaker, 1978; Farkas and Bruce, 2010; Farkas and Roelofsen, 2017). These models of conversation primarily deal with the dynamics of information exchange in discourse.

Sometimes participants' discourse commitments stand in the way of information exchange. *Canonical disagreements* like (1) lead to so-called "conversational crisis", seemingly requiring the participants to either retract some commitment or agree to disagree. Information exchange models provide an explanation for why this is: A and B's assertions cannot both be true, as they are directly contradictory, and so they stall the addition of information to the common ground.

(1) A: This tree is taller than it was yesterday.
B: No, it's not!

This paper examines conversations like (2)—intuitively, also a "disagreement", though it lacks direct contradiction. We might imagine that A and B will continue by trying to change each others' minds. But a simple information exchange model doesn't explain these intuitions: here, A and B both make true assertions, and it is unclear why this sequence of exchanges should stall the conversation if participants were simply informing one another of contrasting attitudes with no further consequence.

(2) A: I think this tree is taller.
B: Well, I don't!

Many models assume that discourse is organized according to, usually implicit, *Questions Under Discussion* (QUDs) (Roberts, 1996/2012; Büring, 2003). The QUD structure of a discourse captures patterns of coherence and relevance between participants' conversational moves. But intuitively, conversations do not always follow this idealized template. At times, higher-order conversational goals drive the way in which a particular sequence of discourse moves unfolds, leading to strategies of inquiry that are not directly tied to the QUD in form. This is what we will say about examples like (2), which we will call *non-canonical disagreement*.

In a case study of three dyadic conversations, we show that non-canonical disagreements arise and are resolved with the same basic signature, regardless of the subjectivity of the QUD guiding the discourse. First, participants establish explicitly autocentric viewpoints as in (2). If these viewpoints contrast, they embark on a longer process of justifying their decision process, and attempt to reach a joint outlook on the issue. This process resembles canonical disagreement despite the fact that participants nowhere establish an actual contradiction.

We ultimately propose that non-canonical disagreements (i) arise when participants embark on

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

a somewhat divergent strategy of inquiry concerning questions of shared viewpoint and (ii) may be settled only if they establish parallel *stances* with respect to a proposition p_n via a process of inference. We will argue that both features might be fruitfully derived using outlook semantics (Coppock, 2018) within an augmented version of the Table model of conversation (Farkas and Roelofsen, 2017).

2 Methodology

Undergraduate students at UC Santa Cruz were recruited in 2020 and 2021. They participated remotely, and were compensated with course credit. First, participants privately reviewed four sets of art or media, each set presented with a free response question. Participants then joined each other in a video call, and were instructed to discuss the art and media they saw and the questions they answered for twenty minutes. They were not explicitly instructed to reach an agreement or record revised answers after discussion. Each participant's audio and video was recorded during this conversation, and an automatic aligned transcript was prepared using the text-to-speech service Descript.

Each set appeared with either an objective or subjective question.[1] For instance, one set included three photographs of Bruce Springsteen. Some dyads were asked the (objective) question "Who is depicted in these images?"; others were asked "Which is the worst picture of the musician?", subjective due to the multidimensional adjective *worst*.

We present here data from hand-corrected transcripts of three dyads, all native speakers of English.[2] Participants 11A (23 F) & 11B (19 F) were paired randomly, while participants 12A (19 F) & 12B (19 F) and 13A (19 F) & 13B (20 F) were friends who signed up to participate together.

We identified eleven disagreements in these transcripts, including both explicit contradictions and self-ascriptions of non-cotenable viewpoints. Six were disagreements about the experimental questions (four objective, two subjective). Another five were about questions that came up organically.

The authors annotated these eleven disagreements with the explicit QUDs introduced by the experiment, and implicit QUDs reconstructed such that participants' moves could be construed as intuitively relevant to the current QUD (see §4). The resulting QUD structures provided a framework to understand the organization of each conversation.

3 Data

This section establishes some basic descriptive generalizations from the three conversations, including many cases of what we call non-canonical disagreement. We show that non-canonical disagreements have a consistent profile: participants self-ascribe differing attitudes or judgments, and, finding themselves in a dispute, take turns justifying their positions with the goal of reaching a joint viewpoint.

3.1 How non-canonical disagreements begin

In discussion of the QUDs, dyads typically began by self-ascribing attitudes towards potential answers to the explicit QUDs. We focus here on cases where those attitudes differ, and discuss how these conversations continue in §3.2.

In this initial stage, participants often self-ascribed attitudes in varying tenses: for instance, in (3), 11A reports a past attitude in order to establish a contrast with a present attitude of 11B. When using the past tense, speakers seemed to be speaking of their attitudes as they viewed the stimuli and answered the given questions.

(3) QUD: *Were these (three) videos produced in the same decade?*

11B: ...**number three had some like animation** at the same time, but **I think that's pretty, that's like more advanced**...

11A: Interesting. See now ... I also do not think they're the same decade, but **I thought that clip two was actually the newest**.

11B: Oh.

Participants also very frequently spoke only of their answers, in which case their attitude towards the propositions that formed their answer is expressed only indirectly. Nevertheless, their attitudes still seem to be the main point of such assertions: e.g. in (4), it is perfectly coherent for 13B to respond to 13A's description of an answer by describing a (contrasting) past-tense attitude.

[1] Throughout this paper, we use the word SUBJECTIVE to pick out all and only the kinds of content about which "faultless disagreement" seems to be licensed (Kölbel, 2004, see also §5.1). Subjective questions thus included predicates of personal taste like *tasty* and *beautiful* (Lasersohn, 2005), but also multidimensional adjectives like *good* (Sassoon, 2013).

[2] Full transcripts of these conversations are available at https://osf.io/jwye8. While the other 47 transcripts in the corpus remain to be corrected, we anticipate making the full corpus publicly available for future research.

(4) QUD: *How many people are singing in each of these recordings?*

13A: **I just put**...one person, in the first one, **two in the second one**, maybe three or more in the last one.

13B: Okay, **I thought the second one only had one person**.

In the above cases, the participants' opening assertions are unprompted, but in other cases they were guided by explicit questions, as dyad 13 demonstrates in (5). In (6) we can see a similar case: 13A asks 13B for their attitude. 13B's response, a description of what they wrote during the task, shows another example of verbs like *put* serving as indirect attitude ascriptions.

(5) QUD: *Which artwork is the most beautiful?*

13B: Which one is most beautiful? Which one did you put?

13A: **The first one**.

13B: Oh, really?

13A: What'd you put, the third one?

13B: No, **I put the second one**...

(6) QUD: *Which is the best picture of Alcatraz?*

13A: Which picture did you like better?

13B: Oh, **I put number two**. What about you?

13A: **The first one**.

Note that these patterns of self-ascription were the most common opening for every experimental question we examined, often producing cases where participants found their attitudes aligned (7).

(7) QUD: *When were these films made?*

12B: I said like nineteen forty.

12A: Okay. Yeah. I was ... in that ballpark.

We also observe no clear differences in behavior across objective QUDs like those in (3-4) vs. subjective QUDs like those in (5-6), featuring the PPT *beautiful* and the multidimensional superlative *best*.

3.2 How non-canonical disagreements settle

In the examples above, participants use first-person indexicals to contribute their respective answers to the QUD without directly disagreeing. They never dispute the accuracy of one another's responses. Nevertheless, when they discovered they held differing viewpoints, they entered into lengthy discussions concerning the validity of each participants' viewpoints, beginning by justifying their own positions and then arguing against their interlocutor's. In this process, agreement was often sought, and freely given on matters of simpler evidence or taste.

For instance, early in the process of resolving the disagreement in (5), 13A directly critiques the piece of art which 13B listed as *most beautiful* (8). Note that 13B agrees with 13A's observations without conceding the validity of their own judgment, which they continue to defend after this excerpt.

(8) (QUD: *Why didn't 13A pick the second?*)

13A: I just thought the second was kind of ... it's like mostly the same color.

13B: Yeah, you're right. ... I had to like really look at it to see what it was.

13A: Yeah, like there's no blacks, or like dark, dark colors.

13B: Hm.

13A: Yeah, there's no dark colors.

13B: Yeah, that's true.

Similarly, after (3), 11A goes on to explain the basis of their position regarding the age of dyad 11's video clips. Again, though 11B here accepts 11A's observations (*mhm* and *yeah*), they continue later in the conversation to contest whether 11A's answer is appropriate.

(9) (QUD: *Why did 11A think clip two was the newest?*)

11A: But ... clips one and three had more of that. Like, you know, that graininess, um—

11B: Yeah.

11A: ...which kind of strangely makes me believe that that was produced much, uh—I said later, but later in the time period, so.

11B: Mhm ... yeah.

Participants do not always seek to justify their own position and levy critiques at others; they can explicitly cooperate in their interlocutor's justification. For instance, in the process of resolving the disagreement in (6), 13B sympathizes with 13A's attitudes towards 13A's favored photo, and later invites them to further explain their preference.

Likewise, the participants also seek out points of agreement on questions that we might think are

properly outside of the scope of the disagreement at hand. In dyad 13's same dispute regarding photographs, shortly after each participant has laid out their opinions of each others' favored photos, 13B seeks out a joint opinion about the third photo, which no one chose (10).[3]

(10) 13B: Can we agree that the last one was terrible?

13A: Yeah, I didn't like that one. {LG} There's like a, um, a pole in the middle of it.

As these processes of justification and alignment continue, it is apparent that participants are aiming to avoid an outcome where they maintain different viewpoints. For instance, dyad 13 doesn't move on from the *best picture* and *beautiful art* conversations until they seem to have reached a consensus. The participants reflect explicitly on the pressure they feel to do this (11).

(11) 13A: Okay, **I guess technically the best picture is the second one**.

13B: Okay. Thank you for caving.

Other times, consensus seems to be reached very easily. In (12), after (4), in the face of 13B's argument, 13A readily changes their mind.

(12) (QUD: *How many people are singing in the second recording?*)

13B: ...for the second one, I'm pretty sure they're harmonizing too.

13A: **Okay.**

13B: So I think there's more than one...

13A: I was thinking of instruments. {LG}

But consensus isn't always reached, and in cases of apparently intractable disagreement, participants sometimes moved on without reaching a joint viewpoint. In such cases participants closed the conversation by re-affirming their different attitudes, as dyad 13 do in (13).

(13) (QUD: *Which piece of music is the best?*)

13B: Maybe **I would say the third one** then.

13A: You'll say the third one?

13B: Yeah.

13A: Okay. **I still like the first one**.

13B: Okay, cool.

[3]The annotation {LG} indicates speaker laughter.

This happened even for non-canonical disagreements for objective QUDs, as in (14). After a protracted dispute about the identity of a musician in a series of photographs, 12A uses a reverse image search (visible only to 12A) to obtain evidence, and reports back that they are convinced the man depicted is Bruce Springsteen. 12B nevertheless remains doubtful, and while 12A acknowledges that it is possible the photos depict someone else (*maybe you're right*), the dyad concluded their conversation with very different apparent belief states.

(14) (QUD: *Is this Bruce Springsteen?*)

12B: I don't think it's Bruce Springsteen.

12A: It so is. It came up. (*in the search*)

12B: I, I don't trust it then.

12A: Okay, two of them (= *pictures*) came back saying Bruce Springsteen.

12B: ...Something in my bones is saying ... it's not that. And I am not a Bruce Springsteen expert, but just–

12A: **I don't know, maybe you're right.** Maybe like someone dressed up and tried to impersonate him or something.

12B: ...**I almost don't want to know who he is**.

3.3 Interim summary

Across the three conversations, we see examples of disagreements with the same general properties. They begin with the establishment of autocentric viewpoints. When viewpoints contrast, a longer discussion ensues where participants review their evidence and decision-making processes. The goal of this process seems to be to negotiate which viewpoint they should collectively adopt, each trying to either collect enough evidence to change their own mind or present enough of an argument to change their interlocutor's. When this goal is achieved successfully, i.e. they reach congruent viewpoints, the QUD is notionally resolved. When they fail to reach a consensus, they simply agree to disagree, as with any canonical disagreement where resolution is not successful. These properties held for objective and subjective questions alike.[4]

In the remainder of the paper, we aim to understand why participants establish autocentric view-

[4]We further note that the same patterns show up on a cursory examination of other conversations in the corpus, across participants regardless of gender and age.

points, and why they aim to agree on a shared viewpoint. This behavior is unexplained in basic models of discourse as information exchange: autocentric viewpoints are not appropriately relevant to the apparent QUD, nor are they sufficient commitments to resolve it, nor are contrasting viewpoints clearly problematic in any way. We will propose that these discrepancies arise because participants are making two non-canonical choices in the structure of their conversation, potentially due to insufficient evidence to fully settle the QUD: (i) they follow a strategy of inquiry which is indirectly related to the QUD, and (ii) they transmute the QUD itself into one which is resolved by a joint outlook rather than joint objective commitments.

4 The relevance of establishing viewpoints

4.1 QUD preliminaries

Strict versions of QUD theory maintain that an assertion must be *relevant* to the QUD that dominates it (Roberts, 1996/2012). Given the objective QUD in (7), *When were these films made?*, participants may be relevant by asserting one of the full answers {*They were made in the thirties, They were made in the forties, ...*}. Alternatively, they may project a *strategy of inquiry*, which requires pursuing answers to subquestions (e.g. *When was film {1,2...} made?*) that are entailed by the higher-level QUD .

One advantage of standard QUD theory is that it captures the information structural relationship between assertions in a discourse and their corresponding QUD structures, which is mediated via *focus*. For each proposition in a set of answers to a question, the focus is associated with alternative expressions (*the thirties, the forties, ...*), whereas the backgrounded content remains constant. The assertion-QUD correspondence, then, is often assumed to be fairly direct.

4.2 Establishing viewpoints

The data presented in §3 constitute a puzzle for models of conversation that adopt standard QUD theory: why is it that a pair of autocentric assertions (e.g. *I said nineteen forty* and *Yeah, I was in that ballpark* in the context of the QUD $p?$: *When were these films made?*) contributes information that is treated by the participants as relevant to the overarching QUD? Note that $p?$ does not entail a subquestion about A or B's attitudes. Nevertheless, exchanges such as the one in (7) seem to be coherent, and even successful in addressing the QUD.

There are similar cases where assertions have been argued to indirectly correspond to their QUDs. For instance, Simons (2007) examines examples like (15), where a proposition embedded under a reportative or attitude predicate exhibits a so-called *embedded main point use*.

(15) A: What's the weather like?
B: Jane said that it's raining.

Here, B's embedded content directly answers the QUD, whereas the matrix content serves to provide information concerning the availability or quality of evidence for the embedded proposition, but only provides an indirect answer to the QUD. If B's response were a direct answer, we would expect the QUD, roughly, to be: *What did Jane say (about the weather)?* Note that this question is not entailed by A's question in (15). Nevertheless, the intuition is that B's response is coherent.

The cases examined in §3 take the same general shape as in (15). That is, participants' assertions are not directly relevant to the explicit QUD.

(16) $p?$: *When were these films made?*
A: I said nineteen forty.
$\rightsquigarrow q?$: (*When did A say they were made?*)

(17) B: I said nineteen fifty.
$\rightsquigarrow r?$: (*When did B say they were made?*)

Note that subQUDs $q?$ and $r?$, too, cannot be part of a strategy of inquiry in the sense of Roberts (1996/2012), as they are not entailed by $p?$. To treat them as nevertheless coherent, we may adopt more relaxed constraints on relevance and entailment, following Riester (2019) and others, though we will have to say more about how exactly they satisfy the interlocutors' purposes.

Concretely, we propose that when participants are faced with a QUD that they have insufficient evidence to address directly,[5] they may choose to adopt a mediating strategy of inquiry which involves the projection of individual autocentric, attitudinal subQUDs, e.g. *When did {A,B} say/think they were made?* in (16). Their choice may be driven by competing conversational pressures: here, complying with the Maxim of Quality may override the pressure to maintain relevance. Given the initiation of this strategy by a speaker, the responding participant is likely to adhere to the same

[5] Notably, the one canonical disagreement in these conversations, where participants make full assertions rather than self-ascribing attitudes, is one about a topic which participants clearly have solid prior knowledge.

strategy (i.e. respond by merely establishing a viewpoint), unless they are more confident in the quality of their own evidence.

The proposed strategy of inquiry explains the form and coherence of the participants' assertions within standard QUD theory, but the question remains how this strategy of inquiry is itself relevant to the QUD dominating it. Our intuition is that a conversation about individual viewpoints serves to allow the participants to assemble evidence and arguments towards an answer to the QUD. But we aim also to explain why this process is not typically complete until participants establish parallel attitudes. In §5, we propose that in addition to adopting this strategy of inquiry, participants also alter their overarching objective, shifting from the search for concrete information to the goal of aligning their hypotheses about the world.

4.3 Retractions

In a situation where one participant manages to convince the other to change positions with respect to their attitude on $p?$, we do not see explicit retraction of the compromising participant's original attitude. That is: explicit, linguistically identifiable acknowledgement of a change in commitment, e.g. *I was wrong*, is non-existent in our data set.

This is perhaps due to the fact that retraction is actually unnecessary. One advantage of asserting an attitude with respect to $p?$, rather than directly addressing the issue of $p?$, is that participants are able to concede their original viewpoints without retracting, as they have only committed to an attitude at a particular time. That is, if the proposition *A thinks the first picture is the best one at time* t is in the common ground and A adds the proposition *A guesses at time* t' *that the second picture is technically the best one*, this expresses a change in A's attitude, but A hasn't made contradictory commitments. Given uncertainty about the answer to a QUD, this is a useful strategy, as it allows participants to assert their attitudes without needing to resolve possible later self-contradictions.

5 Resolving non-canonical disagreements

In the previous section, we suggest that these conversations involve strategies of inquiry that project subQUDs about everyone's attitudes. This allows us to capture the local relevance of those viewpoint-establishing moves, but two puzzles remain. First, how can these attitudes satisfy the participants' goals for the conversation? And second, what is different about non-canonical disagreements such that they don't satisfy those goals?

In this section, we review previous approaches to settling QUDs, and propose that participants in these conversations are actually settling a QUD of a non-transparent form, akin to the QUDs in conversations about taste. The proposal can account for how these conversations get settled without assuming that participants have reached an agreement on the actual state of the world.

5.1 Question resolution in discourse

In modern commitment-based discourse-models, when a QUD is on the table,[6] participants cannot treat that QUD as settled until it has been *resolved*.

(18) **QUD Resolution:** A QUD $p?$ is resolved iff participants have collectively committed to one of its possible answers p_n.[7]

On this approach, one of the principal features of a canonical disagreement is that QUD resolution is blocked unless someone retracts one of their commitments. While a QUD remains unresolved, participants must continue working to establish an answer, or else give up the search a joint answer, perhaps engaging in meta-linguistic negotiation to remove or change the QUD (Ginzburg, 2012).

Here lies the problem with non-canonical disagreements: the viewpoint-establishing moves do not generate the commitments needed to resolve the QUD, even when participants establish the same viewpoint. Likewise, even when viewpoints are not aligned, they do not block QUD resolution.

Similar discrepancies between participants' utterances and their discourse effects are at-issue in work on subjective meaning. Consider (19).

(19) A: This chili is tasty.
  a. B: Yes, it is.
  b. B: No, it's not.

One influential approach to subjectivity since Lasersohn (2005) offers a relativist semantics in which predicates like *tasty* might be true or false of the same tasted object within the same world depending on the identity of a judge parameter in the context of assertion. A's assertion would only commit

[6] Classically, table models (Farkas and Bruce, 2010; Farkas and Roelofsen, 2017) track only explicit QUDs. We will assume here that implicit QUDs can also enter the table; see Ginzburg (2012) for a similar proposal.

[7] For Farkas and Roelofsen (2017), once the set of worlds compatible with everyone's commitments entails p_n.

A to an autocentric judgment (see Stephenson, 2007). But on a relativist semantics, three puzzles remain for the pragmatics: (i) why does A's move seem to project a more general QUD about the tastiness of the chili?; (ii) in (19a), why does that QUD seem to be settled?; and (iii) in (19b), why does there seem to be pressure to continue on the same topic until a state like (19a) is reached?

A compelling response to these puzzles that has emerged in the ensuing pragmatic literature is that interlocutors in these conversations are collaborating not to narrow down the set of possible worlds they collectively might inhabit, but the set of what Coppock (2018) calls *outlooks*, refinements on possible worlds that include positions on subjective issues (see also Stephenson, 2007; Egan, 2010; Rudin and Beltrama, 2019). In an outlook-based semantics, we can speak of truth or falsity of a given proposition in the actual outlook of the author of an information state. In a commitment-based conversational model built on top of this, participant's individual discourse commitment sets describe the outlooks they represent themselves as having, while the common ground encompasses joint outlooks. In particular, this accounts for the above intuitions without giving up on the classic intuition of "faultlessness" in disagreements like (19b) (Kölbel, 2004; MacFarlane, 2014): both participants are asserting felicitously based on their outlook, even though by doing so they ultimately block resolution of the QUD.

We will assume this formalism for subjective meaning. As described in §3.2, in the disagreements we are discussing, the profile of participants' reactions to either type of disagreement was largely the same. For these reasons, we will mostly abstract away from the differences between objective and subjective expressions in what ensues.

5.2 Pondering: When stances are enough

Basic models of conversation as information exchange cannot capture why stances about answers to a QUD appear to settle it or prevent settling it. We propose instead that participants in these conversations are not aiming to resolve the questions that were posed to them *per se*, but instead aiming to reach a type of joint outlook on those questions, employing the formalism of Coppock (2018).

In the spirit of conversational models like Ginzburg (2012) that outline the ways participants might negotiate changes to the current parameters of their conversation, we suggest that participants have at their disposal a conventional parameter change procedure we'll call PONDER.

(20) PONDER: When participants think they cannot adequately answer a QUD $p?$ with answers $\{p_1, p_2 ... p_n\}$, they may replace q with an alternative QUD $p'?$ with answers $\{p'_1, p'_2 ... p'_n\}$ such that p'_n is true for a given outlook iff that outlook includes a positive stance towards p_n.

(21) **Positive stance**: An individual has a positive stance towards p_n in the context of a QUD $p?$ iff among the answers of $p?$ they are most willing to entertain that p_n is the case.

We'll call the QUDs PONDER generates *stance-QUDs*. The answers to these stanceQUDs are discretionary propositions as defined by Coppock (2018), which concern the views of the individual(s) committed to them, and can be true or false in a given world, depending on the outlook in question. We might paraphrase a stanceQUD as "Which answer p_n to $p?$ are we most willing to entertain?", though crucially they are dependent not on deictic *we* but a Lasersohnian judge. Lacking evidence of the explicit form of stanceQUDs (or whether a form exists), we might adopt a somewhat liberal position on the nature of QUDs: while there is often a natural correspondence between QUDs on the table and their syntactic form, perhaps this need not always be the case.[8] Minimally, stanceQUDs are abstract goals with resolution conditions that we can represent formally.

To settle the stanceQUD, participants project a strategy of inquiry that is not directly related in form, as discussed in §4. This strategy of inquiry is what invites participants to assert e.g. first-person attitudes, and it is only through inference that these assertions are taken to establish a stance, rather than merely commit to a first-person attitude. It is thus a special pragmatic effect of such assertions in the contexts we are discussing that they also enter an answer to a stanceQUD into the speaker's discourse commitments.[9] Because this approach is

[8] This flexibility would make structural relationships (e.g. Büring, 2003) between stanceQUDs and subQUDs impossible, but note that we have already relaxed that assumption in §4.

[9] First-person assertions live a similar double life with canonical subjective QUDs. In (i), the QUD seems settled, but with a Coppock-style representation, the objective assertions could not directly resolve it. As in our cases, if we take the assertions to imply certain outlooks, we can see why the QUD has been resolved.

somewhat novel and formally complex, we demonstrate in detail how it would model non-canonical agreement and disagreement in Appendix A.

If both speakers indirectly establish p'_n—a positive stance towards answer p_n of the original QUD—they will have formally resolved the stance-QUD, capturing why these conversations seem to be settled. In contrast, if anyone infers a stance for a speaker that differs from their interlocutors', a conversational crisis will ensue. We could assume that participants refrain from inferring stances in this case, but if so, the stanceQUD remains unresolved. To resolve it, participants must continue discussion of the matter until both are willing to establish the same stance. In this way, we capture the pressure for stance alignment observed in §3.2 as a species of the same pressure observed in any conversation, to answer all QUDs on the table.

5.3 Resolving the QUD via inference?

We have proposed that participants are able to implicitly alter the structure of a QUD to introduce a goal with properties more similar to a subjective question, that is resolved through a joint stance. One can imagine another analysis: instead of assuming an implicitly altered QUD, why not assume implicitly strengthened commitments? We'll give one argument against the latter approach.

For viewpoint-establishing moves to resolve the apparent QUD $p?$, it would have to be the case that participants infer that p_n is part of a speaker's discourse commitments when that speaker expresses a positive view of p_n. This is not *prima facie* unreasonable: consider the premise in (22).

(22) **Commitment to attitudes:** A participant in a conversation where p_n is relevant with a positive view of p_n should be committed to p_n for the purposes of the conversation.

With this premise, parallel viewpoint-establishing answers to personalized subQUDs would make participant commitments about the main QUD readily inferable. Once all participants have established a positive view of p_n, p_n will be assumed to be part of all of their discourse commitments, and thus the QUD can be resolved.

But (22) seems to crucially mischaracterize what we usually infer when participants self-ascribe a viewpoint. The argument against it follows the objection Simons (2007) raises against treating cases like (15) as "assertive" (Hooper, 1975). If in (15) B is understood to commit that it's raining, we could understand how their assertion answers A's question. But this analysis misses another classic Gricean implicature of B's utterance, that by avoiding a more direct locution, B gives the impression that they are unwilling to assert that it's raining.[10] The same critique is relevant here. Speakers who merely establish a positive view of p_n are specifically and effortfully avoiding full commitment to p_n. It runs counter to that avoidance to assume they are implicitly committing to p_n.

In contrast, the stanceQUD account manages to capture the ways in which establishing a stance settles a QUD, without dangerously assuming that all participants are representing themselves as committed in full to a particular answer. It is ultimately an empirical question whether participants take stances as evidence for implicit commitments, but until such evidence can be established, we take our proposal to be preferable.

6 Extensions and upshots

6.1 Predicting (in)felicitous responses

We briefly note one piece of evidence to support the validity of the subQUD structure that governs autocentric strategies of inquiry.

(23) $p?$: *(What about It?)*

12B: I don't know if it's weird but I just got like slight *Devil All The Time* vibes from *It*.
$\rightsquigarrow q?$:(*Did B get Devil All The Time vibes?*)

12A: Yes!
$\rightsquigarrow r?$: (*Did A get Devil All the Time vibes*?)

In (23), A responds with the positive polar response particle (PRP) *Yes*. This leads to the "sloppy" interpretation that A also got these vibes from the movie *It*, as opposed to the strict interpretation, which would simply affirm B's assertion. Moreover, this seems to be a general property of PRP responses to attitudes: responding *Yes* to *I hope it rains tomorrow* can only mean that the responding participant also hopes that it will rain. In contrast, a positive PRP response to a non-attitude report such as *I had a bad dream* is infelicitous. The fact that

(i) QUD: *Is this chili tasty?*
A: I like this chili.
B: I do too.

[10] See also Simons (2019) for a more recent argument against a relevance implicature analysis of (15).

PRPs can reference subQUDs $q?/r?$ provides evidence for this subQUD structure in conversations about attitudes, and suggests a way for future work to provide empirical tests for our claims here.

6.2 Subjectivity

Our proposal suggests there are two routes to explain cases of so-called "faultless disagreements." In addition to assertions which are properly judge-dependent, we have argued that stance self-ascriptions can be used to faultlessly disagree with regard to an implicit subjective QUD while being strictly objective in form. Faultless disagreement has been advanced as a diagnostic for the presence of relative truth, not just for predicates of personal taste and their ilk, but also epistemic modality (Stephenson, 2007; MacFarlane, 2011; see Weatherson and Egan, 2011) and statements about the future (MacFarlane, 2003; Giannakidou and Mari, 2018), even though the latter cases fail other diagnostics like *find*-embedding (Coppock, 2018). It's possible that impressions of disagreement for some of these cases come about not through the presence of bona fide judge-dependent meaning, but because in context they are being used to address implicit questions that require joint outlooks. We hope that future work, especially examining naturalistic conversations, might follow up on this possibility.

We also note that the similarities between participants' treatment of objective and subjective QUDs are good evidence for theories of subjective meaning that predict misaligned outlooks to be just as dire as incompatible commitments. We plan to continue looking for differences in behavior on a larger scale as we prepare the complete corpus.

6.3 Outlook congruence

We suggest, tentatively, that the desire for participants to reach a joint outlook may be driven by a general pressure to achieve social cohesion with one's interlocutor (Edwards and Middleton, 1986; Egan, 2010; Coppock, 2018). While not a requirement, this pressure would explain the preference to attempt alignment before leaving the QUD unresolved. The source of this non-linguistic pressure and its empirical validity remain somewhat underexplored, but this idea is consistent with other work on socially-induced QUDs in similar autocentric conversational contexts. For example, Balachandran (2021) argues that a social principle called the *Norm of Reciprocity*, which underlies a pressure for participants to reciprocate in situations involving avowals and conflicts, induces a QUD structure that has the ability to mediate instances of mismatching indexical reference in verb phrase ellipsis (see Chung (2000) and Charnavel (2019) for more detail).

(24) QUD: (*Do A and B love each other?*)

A: I_A love you_B.
$\rightsquigarrow$ subQUD: (*Does A love B?*)

B: Well, I_B don't <love you_A>!
$\rightsquigarrow$ subQUD: (*Does B love A?*)

In (24), A's assertion is taken to project an implicit QUD structure and compel B to respond to the subQUD *Does B love A?* The fact that B's response (24) appears to trigger disagreement is derived pragmatically: violation of the Norm of Reciprocity is taken to lead to interpersonal conflict, but does not block QUD resolution. In contrast, aligned stances are required to settle the QUD under the current analysis. Though both cases aim to derive a pressure for "alignment", here we enshrine this as a proper condition on QUD resolution. This is perhaps desirable, as the nature of these misalignments intuitively seem distinct in some sense, despite their similarities on the surface. Future work should aim to more thoroughly consider a pragmatic analysis of aligning stances.

6.4 Summary

In this article we have provided a description of a conversational phenomenon that proves challenging to treat using the basic toolbox of commitment-based discourse modeling. We suggested adding to that toolbox in two ways to account for these conversations: (i) allowing for implicitly projected strategies of inquiry that are not directly relevant to the current QUD, and (ii) formalizing how participants might pursue a shared hypothesis rather than a complete answer to a QUD. With these components in place, non-canonical disagreements look much like subjective disagreements, raising questions for future work on subjectivity and the role(s) of generalized social alignment in linguistic theories of discourse.

7 Acknowledgements

Jennifer Wheelock contributed to the design of the experiment in which these conversations took place, and assisted in data collection together with Dasha Komissarchik and David Tuffs. Robert Liu, Stephanie Palomares, Vindhya Shigehalli, Jacqueline Wen, and Naomi Wong assisted with transcription and annotation. The authors would like to thank Pranav Anand and Maziar Toosarvandani for helpful comments.

References

Lalitha Balachandran. 2021. Reciprocal questions and the pragmatics of argument reversing verb phrase ellipsis. In *Proceedings of Sinn und Bedeutung 26*.

Daniel Büring. 2003. On D-trees, beans, and B-accents. *Linguistics and Philosophy*, 26(5):511–545.

Isabelle Charnavel. 2019. Supersloppy readings: Indexicals as bound descriptions. *Journal of Semantics*, 36(3):453–530.

Sandra Chung. 2000. Close encounters with pronouns in VP ellipsis. In Sandra Chung, Jim McCloskey, and Nathan Sanders, editors, *Jorge Hankamer's WebFest*.

Elizabeth Coppock. 2018. Outlook-based semantics. *Linguistics and Philosophy*, 41:125–164.

Derek Edwards and David Middleton. 1986. Joint remembering: Constructing an account of shared experience through conversational discourse. *Discourse Processes*, 9(4):423–459.

Andy Egan. 2010. Disputing about taste. In Richard Feldman and Ted A. Warfield, editors, *Disagreement*, pages 247–286. Oxford.

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

Donka F. Farkas and Floris Roelofsen. 2017. Division of labor in the interpretation of declaratives and interrogatives. *Journal of Semantics*, 34.

Anastasia Giannakidou and Alda Mari. 2018. A unified analysis of the future as epistemic modality: The view from Greek and Italian. *Natural Language & Linguistic Theory*, 36:85–129.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford.

Joan B. Hooper. 1975. On assertive predicates. In John P. Kimball, editor, *Syntax and Semantics*, volume 4, pages 91–124. Academic Press.

Max Kölbel. 2004. Faultless disagreement. *Proceedings of the Aristotelian Society*, 104:53–73.

Natasha Korotkova. 2016. Disagreement with evidentials: A call for subjectivity. In *Proceedings of SemDial 20 (JerSem)*, pages 65–75.

Peter Lasersohn. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, 28:643–86.

John MacFarlane. 2003. Future contingents and relative truth. *The Philosophical Quarterly*, 53(212):321–336.

John MacFarlane. 2011. Epistemic modals are assessment-sensitive. In Andy Egan and Brian Weatherson, editors, *Epistemic modality*, pages 144–178. Oxford.

John MacFarlane. 2014. *Assessment Sensitivity: Relative Truth and its Applications*. Oxford.

Arndt Riester. 2019. Constructing QUD trees. In *Questions in Discourse*, volume 2, pages 164–193. Brill.

Craige Roberts. 1996/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics*, 5:6.

Deniz Rudin and Andrea Beltrama. 2019. Default agreement with subjective assertions. In *Proceedings of Semantics and Linguistic Theory (SALT) 29*, pages 82–102.

Galit W. Sassoon. 2013. A typology of multidimensional adjectives. *Journal of Semantics*, 30:335–380.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117:1034–1056.

Mandy Simons. 2019. The status of main point complement clauses. In *Proceedings of SemDial 23 (LondonLogue)*, book section 15.

Robert C. Stalnaker. 1978. Assertion. In Peter Cole, editor, *Syntax and Semantics*, volume 9, pages 78–95. Academic Press.

Tamina Stephenson. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, 30(4):487–525.

Brian Weatherson and Andy Egan. 2011. Introduction: Epistemic modals and epistemic modality. In *Epistemic modality*, pages 1–18. Oxford.

A Modeling stanceQUDs and their resolution

More or less in keeping with Farkas and Roelofsen (2017),[11] we take the state of a dyadic conversation to have at least the following components:

(25) TABLE: A stack of QUDs, represented as sets of answers.

(26) DISCOURSE COMMITMENTS ($DC_{A/B}$): The set of propositions each speaker is committed to for the purposes of the conversation.

(27) COMMON GROUND (CG): $DC_A \cap DC_B$, the set of all propositions the participants share joint commitments to.

We also assume the following:

(28) COMMITMENT SETS ($CS_{A/B}$): For a participant n, $\bigcap DC_n$, the set of all outlooks consistent with their commitments.

(29) CONTEXT SET (CS): $CS_A \cup CS_B$, the set of all outlooks consistent with all participants' commitments.

(30) **QUD Resolution** (formal):
A QUD $p?$ can be removed from the TABLE when $\exists p_n \in p?\,.\,CS \subset p_n$, that is, when participants' commitments entail an answer.

Now consider the toy conversation in (31).

(31) QUD: *Which video is the newest?*

A: I think the first video is the newest.

B: I also think the first video is the newest.

To model the conversation, we'll make reference to the following propositions and set of possible outlooks U.

(32) a. p_n is the set of outlooks where video n is the newest

b. q_n is the set of outlooks where A thinks video n is the newest at time t.

c. r_n is the set of outlooks where B thinks video n is the newest at time t'.

d. p'_n is the set of outlooks which include a positive stance towards p_n

(33) a. $U = \{w_1o_1, w_1o_2, w_2o_1 ... w_8o_2\}$

b. $p_1 = \{w_1o_1, w_1o_2, w_2o_1, w_2o_2, w_3o_1, w_3o_2, w_4o_1, w_4o_2\}$; $p_2 = U \setminus p_1$

c. $q_1 = \{w_1o_1, w_1o_2, w_2o_1, w_2o_2, w_5o_1, w_5o_2, w_6o_1, w_6o_2\}$; $q_2 = U \setminus q_1$

d. $r_1 = \{w_1o_1, w_1o_2, w_3o_1, w_3o_2, w_5o_1, w_5o_2, w_7o_1, w_7o_2\}$; $r_2 = U \setminus r_1$

e. $p'_1 = \{w_1o_1, w_2o_1, w_3o_1, w_4o_1, w_5o_1, w_6o_1, w_7o_1, w_8o_1\}$; $p'_2 = U \setminus p'_1$

We assume the conversation starts as follows:

(34) TABLE $= [\{p_1, p_2 ...\}]$
$CS_A, CS_B = \{w_1o_1 ... w_8o_2\}$
$CS = \{w_1o_1 ... w_8o_2\}$

The participants decide to apply PONDER:

(35) TABLE $= [\{p'_1, p'_2 ...\}]$
$CS_A, CS_B = \{w_1o_1 ... w_8o_2\}$
$CS = \{w_1o_1 ... w_8o_2\}$

They then project a strategy of inquiry which first involves the subQUD *Which video does A think is the newest?*. A's assertion adds the commitment q_1 to DC_A, and B automatically adds q_1 to DC_B, as A is an expert on A's attitudes (see Korotkova, 2016).

(36) TABLE $= [\{q_1, q_2 ...\}, \{p'_1, p'_2 ...\}]$
$DC_A, DC_B = \{q_1\}$
$CS_A, CS_B = \{w_1o_1 ... w_2o_2, w_5o_1 ... w_6o_2\}$
$CG = \{q_1\}$
$CS = \{w_1o_1 ... w_2o_2, w_5o_1 ... w_6o_2\}$

After (36), the subQUD is resolved and can be removed, because $CS \subset q_1$. The participants can also jointly infer from A's assertion that A has a positive stance towards p_1. This positive stance p'_1 is entered into A's discourse commitments, in turn restricting the outlooks the worlds that remain in their commitment set.

(37) TABLE $= [\{p'_1, p'_2 ...\}]$
$DC_A = \{q_1, p'_1\}$
$DC_B = \{q_1\}$
$CS_A = \{w_1o_1, w_2o_1, w_5o_1, w_6o_1\}$
$CS_B = \{w_1o_1 ... w_2o_2, w_5o_1 ... w_6o_2\}$
$CG = \{q_1\}$
$CS = \{w_1o_1 ... w_2o_2, w_5o_1 ... w_6o_2\}$

The next subQUD in the strategy of inquiry is *Which video does B think is the newest?*. B's assertion adds the commitment r_1 to DC_B, and A automatically follows suit.

[11] We crucially allow implicit QUDs on the table.

(38) $\text{TABLE} = [\{r_1, r_2...\}, \{p'_1, p'_2...\}]$
$DC_A = \{q_1, r_1, p'_1\}$
$DC_B = \{q_1, r_1\}$
$CS_A = \{w_1o_1, w_5o_1\}$
$CS_B = \{w_1o_1, w_1o_2, w_5o_1, w_5o_2\}$
$CG = \{q_1, r_1\}$
$CS = \{w_1o_1, w_1o_2, w_5o_1, w_5o_2\}$

After (38), the subQUD is resolved and can be removed, because $CS \subset r_1$. The participants can also jointly infer from B's assertion that B also has a positive stance towards p_1. This positive stance p'_1 is entered into B's discourse commitments, in turn restricting the outlooks the worlds that remain in their commitment set.

(39) $\text{TABLE} = [\{p'_1, p'_2...\}]$
$DC_A = \{q_1, r_1, p'_1\}$
$DC_B = \{q_1, r_1, p'_1\}$
$CS_A = \{w_1o_1, w_5o_1\}$
$CS_B = \{w_1o_1, w_5o_1\}$
$CG = \{q_1, r_1, p'_1\}$
$CS = \{w_1o_1, w_5o_1\}$

After (39), the stanceQUD is resolved and can be removed, because $CS \subset p'_1$. The participants have determined that they share an outlook that contains a positive stance towards p_1. Note nevertheless that they crucially have not determined whether p_1 is true.

We can also model non-canonical disagreements as in (40).

(40) QUD: *Which video is the newest?*
A: I think the first video is the newest.
B: I think the second video is the newest.

This conversation diverges from the one above after 37. B's assertion this time adds the commitment r_2 to DC_B, and A automatically follows suit.

(41) $\text{TABLE} = [\{r_1, r_2...\}, \{p'_1, p'_2...\}]$
$DC_A = \{q_1, r_2, p'_1\}$
$DC_B = \{q_1, r_2\}$
$CS_A = \{w_2o_1, w_6o_1\}$
$CS_B = \{w_2o_1, w_2o_2, w_6o_1, w_6o_2\}$
$CG = \{q_1, r_2\}$
$CS = \{w_2o_1, w_2o_2, w_6o_1, w_6o_2\}$

After (41), the subQUD is resolved and can be removed, because $CS \subset r_2$. The participants can also jointly infer from B's assertion that B also has a positive stance towards p_2. This positive stance p'_2 could be entered into B's discourse commitments, in turn restricting the outlooks the worlds that remain in their commitment set.

(42) $\text{TABLE} = [\{p'_1, p'_2...\}]$
$DC_A = \{q_1, r_2, p'_1\}$
$DC_B = \{q_1, r_2, p'_2\}$
$CS_A = \{w_2o_1, w_6o_1\}$
$CS_B = \{w_2o_2, w_6o_2\}$
$CG = \{q_1, r_2\}$
$CS = \{\}$

But (42) is catastrophic, with no possible joint outlooks remaining in CS. If participants remove inferred stance commitments, they could end up in the state in (43), no longer catastrophic but notably without any answer to the QUD on the table.

(43) $\text{TABLE} = [\{p'_1, p'_2...\}]$
$DC_A = \{q_1, r_2\}$
$DC_B = \{q_1, r_2\}$
$CS_A = \{w_2o_1, w_2o_2, w_6o_1, w_6o_2\}$
$CS_B = \{w_2o_1, w_2o_2, w_6o_1, w_6o_2\}$
$CG = \{q_1, r_2\}$
$CS = \{w_2o_1, w_2o_2w_6o_1, w_6o_2\}$

Because the participants still have pressure to establish a joint stance, and because they are free to make new attitudinal claims for times beyond t and t', a likely continuation is to attempt to convince someone to switch attitudes, and thereby establish a joint stance. For instance, B may eventually commit to some new proposition s_1, that B thinks p_1 at time t'', thereby offering a chance to infer that p'_1 should be added to their discourse commitments. This would result in (44), a late but successful resolution.

(44) $\text{TABLE} = [\{p'_1, p'_2...\}]$
$DC_A = \{q_1, r_2, s_1, p'_1\}$
$DC_B = \{q_1, r_2, s_1, p'_1\}$
$CS_A = \{w_1o_1, w_5o_1\}$
$CS_B = \{w_1o_1, w_5o_1\}$
$CG = \{q_1, r_2, s_1, p'_1\}$
$CS = \{w_1o_1, w_5o_1\}$

Classifying the Response Space of Questions: A Machine Learning Approach

Zulipiye Yusupujiang
Université Paris Cité, CNRS
Laboratoire de Linguistique Formelle
zulipiye.yusupujiang@etu.u-paris.fr

Alafate Abulimiti
INRIA, Paris, France
alafate.abulimiti@inria.fr

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

Abstract

The main goal of this work is to conduct a pilot study on the automatic classification of the response space of questions in English. We aim for a relatively fine-grained understanding of the learning problem of this response space; hence, we conducted classical machine learning studies to automatically identify different response classes based on carefully designed features. Moreover, we compared the results from feature-based classical machine learning algorithms to the classification results obtained from a large-scale pre-trained BERT language model. Experimental results show that the feature-based classical machine learning algorithms can achieve performance results which are close to the results obtained by BERT model on this novel task. The overall trend of the classification results for each response class are also similar in both models. Learnability trends similar to corpus-based studies presented in previous literatures emerge.

1 Introduction

Classifying the response space of questions plays an important role in the design of dialogue systems, particularly systems that can be easily adaptable across domains (Larsson and Berman, 2016). Łupkowski and Ginzburg (2013, 2016) offer an empirical and theoretical characterization of one significant component of the response space of questions, which is responding to a question with a question, which represents more than 20% of all responses to questions found in the British National Corpus (BNC) (Burnard, 2007). Based on a detailed corpus study on the British National Corpus and three other more genre-specific corpora (BEE (Rosé et al., 1999) and AmEx (Kowtko and Price, 1989)) and a sample from CHILDES (MacWhinney, 2000)), Łupkowski and Ginzburg (2013, 2016) provide 7 classes of question responses: `CR`: *clarification requests*, `DP`: *dependent questions*, `MOTIV`: *requests for underlying motivation*, `FORM`: *questions about the form of the expected answer*, `NO ANSW`: *questions raised with the aim of not answering the initial question*, `IND`: *questions providing a potential answer*, and `IGNORE`: *questions raised to ignore the initial question*.

Following the aforementioned research, Ginzburg et al. (2019, 2022) extend the classification of response space to cover all responses to questions. They provide a full response space taxonomy with 9 unique response classes of responses to questions and one OTHER class. They conduct cross-linguistic studies comparing English and Polish.

The main aim of the current work is to conduct a pilot study for automatic classification of response space of questions, based on the taxonomy proposed by Ginzburg et al. (2019, 2022). Such an approach lays a foundation for the automation of response space classification in designing dialogue systems.

This paper is structured as follows: In section 2, we discuss related work on classifying other types of utterances in dialogue. Section 3 contains a discussion of the taxonomy of responses to questions used in this study. In Section 4, we introduce the response space annotation process and labeled dataset. Section 5 presents the experiments on BERT language model and its results. We then introduce the specifically created feature sets, and

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

discuss the results and learnability of different response classes from a classical machine learning algorithm in Section 6. In the last section, we offer some conclusions and discuss future work aimed at improving this study.

2 Related Work

Fernández et al. (2007) propose a taxonomy with 15 classes for Non-Sentential Utterances (NSU) in dialogue, based on a detailed corpus study on BNC. In addition, they also present several results from automatically classifying NSUs using some well-known machine learning techniques. For the machine learning approach, they use the majority class predictor, one-rule classifier, and also the J4.8 decision tree algorithm using the Weka Toolkit (Witten and Frank, 2002). Classification results from the algorithms above served as the baselines of their study. Three other machine learning systems were also used, SLIPPER (Cohen and Singer, 1999), TiMBL (Daelemans et al., 2003), and MaxEnt (Zhang, 2007), in order to conduct a more sophisticated experiment and get a reliable result. To train the machine learning algorithms, Fernández et al. (2007) used three types of feature sets which capture either the properties of NSUs, of the antecedent utterance, or the relations between NSUs and the antecedents. Their results show that machine learning algorithms benefit from utilizing the properties of the antecedent of NSUs and also the relationships between them.

Dragone and Lison (2015) propose an active learning approach to the classification of NSUs, by an extension of the work of Fernández et al. (2007). They extend the feature set from 9 features to a total of 32 features by extracting more features with the PCFG and Dependency Parser from the Standford CoreNLP API (Dragone and Lison, 2015). An active learning method is used to deal with the labelled data scarcity problem. The experimental results show a significant improvement on the classification task when comparing it to the baseline of Fernández et al. (2007). In this study, we use similar methods used to classify NSUs as discussed above.

Clarification requests (CRs) are also common in human dialogue. According to Purver et al. (2003a); Rodríguez and Schlangen (2004), CRs account for 3%-6% of human-human dialogue. CRs are also common in response space taxonomy (4.84% as shown in Table 2). Purver (2006) studies Clarification Requests in details and presented all major forms of CRs and analyzed their readings. He also offered a computational implementation of CRs within a prototype text-based dialogue system - CLARIE.

In addition, Cruz-Blandón et al. (2019) propose a semantic annotation scheme for questions and answers based on the contribution of content and discourse on them. They divided the questions into 5 types: *Yes/No question*, *Completion suggestion*, *Disjunctive question*, *Wh-question*, and *Phatic question*. The authors also categorized answers into 7 different types: *Positive answer*, *Negative answer*, *Feature answer*, *Phatic answer*, *Uncertainty answers*, *Unrelated Topic*, and *Deny the assumption*. They applied this annotation scheme to multiple languages (English, Spanish, and Dutch), and also offered an initial experiment for automating the annotation of question types in English dialogues. Cruz-Blandón et al. (2019) used 8 different hand-designed features and reported the classification results from both statistical machine learning algorithms (`Majority Baseline`: *acc.*=0.47, *F1*=0.31; `Decision Tree`: *acc.*=0.73, *F1*=0.58) and neural networks (`Bag-of-Words`: *acc.*=0.76, *F1*=0.44; `RNN`: *acc.*=0.54, *F1*=0.24).

3 A Taxonomy of Responses to Questions

As mentioned in the previous section, we deploy the corpus-based taxonomy proposed by (Ginzburg et al., 2019, 2022) in our study of automatic classification of response space of questions. They propose that the class of responses to a question q_1 can be classified into three main categories:

(1) a. Q(uestion)–specific: responses directly or indirectly about or subquestions of q_1;

b. MetaCommunicative: responses directly about or subquestions of a question defined in part from the *utterance* of q_1;

c. Evasion: responses directly about or subquestions of a question that is distinct from q_1 and arises from some other component of the context.

The first group is further classified as Direct Answers (DA) which constitute an answer to the initial question, and Indirect Answers (IND) through which one can infer an answer from its content, and also Dependent Questions (DP) where the answer

to the initial q_1 depends on the answer to this query response. The second group is divided into Clarification Responses (CR) which inquire additional information to better understand the initial question, or to clarify some mis-presuppositions addressed in q_1. Acknowledgment (ACK) is the second class under the Metacommunicative group, which signals that the speaker heard and understood the q_1. The last group, Evasion responses, can be further categorized in to four response classes:

1. Ignore (IGNORE) (the utterance does not relate to the question, but to the situation. e.g., *A: So lock erm how would you spell sock? B: <laugh> smelly er smelly* (BNC));

2. Change the topic (CHT) (e.g., *A: Why couldn't they come on Friday? B: What you got me then?* (BNC));

3. Motive (MOTIV) *A: What's the matter? B: Why? (BNC)*;

4. Difficult to provide a response (DPR) (*A: When's the first consignment of Scottish tapes? B: Erm <pause> don't know.*).

The taxonomy is presented in Table 1.

Category	TAG
1. Direct answer	DA
2. Indirect answer	IND
3. Dependent question	DP
4. Clarification response	CR
5. Acknowledgment	ACK
5. The utterance does not relate to the question, but to the situation	IGNORE
6. Utterance signalizes that speaker does not want to answer, s(he) changes the topic, gives an evasive answer	CHT
8. Question about the motivation for the initial question	MOTIV
9. Difficult to provide an answer	DPR
10. Utterance that does not fit in any of the above	OTHER

Table 1: Taxonomy proposed by Ginzburg et al. (2022) and used in this paper

In the following section, we describe our data, annotation process, and also the inter-annotator agreement between annotators.

4 Response Space Annotation

Following the previous studies and the response space annotation guideline provided by Ginzburg et al. (2019, 2022), we annotated question-response pairs (QR-pairs) from different dialogue corpora. We manually annotated dialogues from the British National Corpus (BNC) (Burnard, 2007), Cornell-Movie (Danescu-Niculescu-Mizil and Lee, 2011), Basic Electricity and Electronic Corpus (BEE) collected from dialogue-based tutoring system (Rosé et al., 1999), and HCRC MapTask corpus (Anderson et al., 1991).

We manually annotated 3008 QR-pairs from the BNC corpus, 1172 QR-pairs from the Cornell-Movie, 293 QR-pairs from the HCRC MapTask, and 238 QR-pairs from the BEE corpus. This resulted in 4711 annotated QR-pairs in total. We have a rough estimate that more than 90% of the questions are responded to in the immediately following utterance. This is also in line with the statistics presented in (Purver et al., 2003b) that 94% of the Clarification Requests were answered in the immediately following utterance. Therefore, to facilitate the annotation and data processing for machine learning experiments, we only annotated QR-pairs where the response is the adjacent utterance of the corresponding question. In addition, we did not consider tag questions, such as, *It's too complicated, isn't it?* as a question. Finally, turns with missing text (the BNC's 'unclear') were eliminated from consideration, unless the remaining parts of the utterance provide sufficient information for understanding the meaning of the utterance.

To examine the annotation reliability, we double annotated three files from the BNC, and calculated the inter-annotator reliability based on the Cohen's κ (Carletta, 1996) and Krippendorff's α (Krippendorff, 2011) coefficients. The best inter-annotator agreement scores obtained are 0.8183 and 0.8186 for Cohen's κ and Krippendorff's α respectively. However, the lowest inter-annotator agreement scores are 0.7118 (Cohen's κ) and 0.7128 (Krippendorff's α).

Table 2 shows the distribution of the response space classes in our dataset. As can be observed from the table, the OTHER class is less than 1%, thus the coverage is more than 99%. What's more, the most frequent classes in our dataset are Direct Answers (64.83%), Indirect Answers (10.80%), Difficult to provide answer (5.20%), Change the topic (4.95%), and Clarification Re-

sponses (4.84%). The less frequent classes are DP (0.89%), MOTIV (0.30%), and ACK (3.12%).

The dataset used in this study is highly imbalanced, since the response class DA (64.83%) has significantly more samples than the others, as indicated in Table 2. Therefore, it is important to find a solution to overcome the classification difficulty caused by imbalanced data. In the following section, we introduce the baseline model obtained by the BERT pre-trained English language model (Devlin et al., 2018).

Category	Total	Frequency%
DA	3054	64.83%
IND	509	10.80%
DP	42	0.89%
CR	228	4.84%
ACK	147	3.12%
IGNORE	208	4.42%
CHT	233	4.95%
MOTIV	14	0.30%
DPR	245	5.20%
OTHER	31	0.66%
Total	**4711**	**100%**

Table 2: Overall distribution of response space classes in the dataset

5 Response Space Classification with BERT

To begin with, we set up an experiment with the pre-trained BERT language model, and examined the classification performance of such a large language model on the novel task of response space classification. First of all, we deleted all `OTHER` cases from our annotated dataset, which resulted in a total of 4680 annotated QR-pairs with 9 unique response classes. The distribution of the training, validation, and test sets are 60%, 20%, and 20% respectively. We add 2 special tokens <q> and <r> into BERT tokenizer's vocabulary, and the input of the BERT model is organized as {<q> *question* <r> *response*}.

We conducted two separate experiments: (1). with the full response space taxonomy of 9 unique classes; (2). with a coarser response space taxonomy of only 4 main classes, namely, Direct Answers, Indirect Answers, Clarification Responses, and Evasion. All classes which belong neither to Direct Answers, Indirect Answers, nor Clarification Responses were merged and classified as Evasion. We think that this is a more practical response space taxonomy in designing dialogue systems. In addition, we did not use any resampling techniques when classifying with the BERT language model, since BERT is already trained on a large amount of language data. Therefore, we are interested in seeing how it performs on this response space classification task with a skewed dataset.

Table 3 presents the classification results from the BERT language model on the full response space taxonomy. We use the classification results achieved by BERT model as the baseline for this study, and conduct several experiments to study whether we can obtain similar results as BERT by using classical machine learning algorithms trained with a set of carefully designed features.

As Table 3 shows, the baseline BERT model results in an average weighted f1-score of 0.70 and a macro f1-score of 0.40 on the full taxonomy. Besides, the BERT model achieved roc_auc scores of 0.87 and 0.86 respectively on the full and coarser taxonomy. This signals the very good performance of the BERT model on the response space classification task because they are very close to the perfect roc_auc score of 1.0. The best classified response class among others is the Direct Answers (f1-score: 0.85) as expected, since this is the easiest class to annotate for the human annotators according to the detailed human annotation report in Ginzburg et al. (2022). The next relatively well classified response classes are Clarification Responses (f1-score: 0.74), Acknowledgments (f1-score: 0.52), and DPR (f1-score: 0.59). This is also in line with the relatively higher inter-annotator agreement on these subsets of the full taxonomy, as presented in the previous response-space related literatures. However, the BERT model did not perform well on Indirect Answers, Dependent Questions, and other more evasive response classes, such as IGNORE, CHT, and MOTIV. The f1-scores are below 0.35 for these classes. Such low classification results were anticipated for response classes DP and MOTIV given the very low frequency of such responses in our dataset as shown in Table 2 (they comprise only 0.89% and 0.30% of the overall dataset). As for the response classes Indirect Answers, CHT, and IGNORE, even though their frequencies are higher than other non-major classes (10.80%, 4.95%, and 4.42% respectively), the classification results achieved by BERT language model are still very low (f1-score: 0.32, 0.33,

Classes	Precision	Recall	F1	Support
DA	0.81	0.88	0.85	593
IND	0.33	0.31	0.32	107
DP	0.10	0.20	0.13	5
CR	0.76	0.72	0.74	47
ACK	0.53	0.52	0.52	31
IGNORE	0.14	0.11	0.12	44
CHT	0.39	0.29	0.33	56
MOTIV	0.00	0.00	0.00	3
DPR	0.82	0.46	0.59	50
accuracy			0.70	936
macro avg.	0.43	0.39	0.40	936
weighted avg.	0.68	0.70	0.68	936
roc_auc_score				0.87
DA	0.77	0.95	0.85	595
IND	0.60	0.20	0.30	126
CR	0.70	0.63	0.67	41
Evasion	0.73	0.51	0.60	171
accuracy			0.75	933
macro avg.	0.70	0.57	0.60	933
weighted avg.	0.74	0.75	0.72	933
roc_auc_score				0.86

Table 3: Classification results of BERT language model on full and coarser response space taxonomy

and 0.12 respectively). This can be attributed to the fact that these response classes are intrinsically reliant on deep inference.

The bottom half of the Table 2 presents the classification results from BERT on the coarser taxonomy. The overall classification results improved in terms of the weighted average f1-score (0.75 vs. 0.70) on the coarser taxonomy. This was expected, since classifiers usually perform better on a coarser taxonomy. However, the f1-score on the classification results on Clarification Responses decreased from 0.74 to 0.67, and the Indirect Answers from 0.32 to 0.30. It can be observed that Indirect Answer is still the most difficult response class to be learned by the BERT language model. Finally, the model resulted in a f1-score of 0.60 on the classification of the Evasion response class, which is the new broader response class after merging all other response classes.

6 Classical Machine Learning Approach

In this section, we first introduce the set of carefully designed features for this response space classification task. Then, we present two groups of machine learning experiments: one with the full response space taxonomy, and the other with a coarser taxonomy.

6.1 Features

Similar to the approach used by Fernández et al. (2007), we also divided the features into three main groups: (i) `Response features`, which are related to properties of the response space; (ii) `Question features`, which are properties of the corresponding question; (iii) `Question-Response features`, which keep track of the features related to both question and response, and also similarities between the question and its corresponding response. All the semantic, syntactic, and lexical properties are extracted by using the Python natural language analysis package: Stanza Qi et al. (2020). Stanza is built with highly accurate neural network components that its neural network NLP pipeline can perform various NLP tasks, including tokenization, multi-word token expansion, lemmatization, POS and morphological tagging, dependency parsing, named entity recognition, and also the sentiment analysis of a natural language data. Table 4 presents the response space features and values used in this study.

`Response features` There are 12 different features related to the responses:

- `res_type`, `res_pers`, `res_number`, `res_tense`, `res_entities`, `res_sentiment`. The feature `res_type` has two values *question* and *proposition*, which are intended to capture the query responses and the propositional responses respectively. We encode the person information of the response with the feature `res_pers`. The feature `res_number` encodes the inflectional features of nouns in the response (*singular*, *plural*). `res_tense` records the time line in which the action in the response occurs (*present*, *future*, *past*). The feature `res_pers`, `res_number`, and `res_tense` use a value *empty* wherever the relevant lexical items are absent. Existence of name entities or proper nouns in the response is recorded with the feature `res_entities` (*yes*, *no*). The last feature `res_sentiment` is responsible for encoding the polarity of verbs, adjectives, adverbs, and nouns in the responses, with values *positive*, *negative*, and *neutral*.

- `rsp_aff` encodes the presence of affirmative word *yes* and *no*, we assign a value *empty* if there is no such word. `rsp_dntknow` has a value *yes* if there are phrases such as "I don't know", "dunno", "not sure", etc., and

Feature	Description	Values
res_type	query or propositional response	question, proposition
res_pers	person point of view in the response	1st, 2nd, 3rd, empty
res_number	inflectional feature of nouns	Sing, Plur, empty
res_tense	verb tense in the response	Pres, Fut, Past, empty
res_entities	presence of name entities	yes, no
res_sentiment	sentiment of the response	positive, negative, neutral
rsp_aff	presence of affirmative words	yes, no, empty
rsp_dntknow	presence of words indicating the absence of knowledge	yes, no
rsp_dprel_discourse	presence of "discoure" dependency	yes, no
rsp_dprel_reparandum	presence of "reparandum" dependency	yes, no
rsp_dprel_mwe	different multiword epresion dependency	compound, fixed, flat, empty
rsp_num_content	number of content words	integer
ques_type	wh-question or polar question	what,which...,polar
ques_pers	person point of view in the question	1st, 2nd, 3rd, empty
ques_number	inflectional feature of nouns	Sing, Plur, empty
ques_tense	verb tense in the question	Pres, Fut, Past, empty
ques_entities	presence of name entities	yes, no
ques_sentiment	sentiment of the question	positive, negative, neutral
ques_num_content	number of content words	integer
which_dem	presence of demonstative pronouns in responses utterance to *which* questions	yes, no
who_prs	presence of personal pronouns in responses utterance to *who* questions	yes, no
where_adp	presence of POS-tag "ADP-adposition" in responses to *where* questions	yes, no
wh_discorse	presence of "discourse" dependency in short responses to *wh* questions	yes, no
repeated_words	number of repeated words	integer
common_content_words	number of repeated common words	integer
pos_sequence	lenght of common POS sequence	integer

Table 4: Features of response space and values

no otherwise. rsp_deprel_discourse checks if there is a "discourse" dependency relation in the response utterance. rsp_deprel_reparandum looks for a "reparandum" dependency relation in the response utterance, which indicates disfluencies in the utterance. rsp_deprel_mwe encodes different dependency relations for multi-word expressions, and it has four values: "compound", "fixed", "flat", and "empty". Lastly, rsp_num_content presents the number of content words in the response utterance.

Question features We also use 7 different features to encode the properties of the corresponding questions, namely, ques_type, ques_pers, ques_number, ques_tense, ques_entities, ques_sentiment, and ques_num_content. The feature ques_type is used to differentiate the various types of wh- questions and polar questions. The other 6 features are used in a same way

as the corresponding features in `Response features` described above.

`Question-Response features`

The last 7 features, `repeated_word` and `pos_sequence`, are the numerical features which encode features related to both question and response, and the similarities between the responses and their corresponding questions. The feature `which_dem` records the presence of demonstrative pronouns in a response utterance to a question with *which* `ques_type`. Similarly, the feature `who_prs` records the presence of personal pronouns in a response utterance to a question with *who* `ques_type`, and the feature `where_adp` records the presence of POS-tag "ADP-adposition" in a response utterance to a question with *where* `ques_type`. Besides, the feature `wh_discourse` indicates the presence of "discourse" dependency relation in short responses (less than or equal to two words) to any *wh-* questions. This feature aims to capture utterances such as "Aha", "Well", "Erm", "Mhm", etc, and they are usually classified as Acknowledgment to *wh-* questions. The feature `repeated_word` represents the number of repeated words between responses and questions; `repeated_word` shows the number of common content words in questions and responses; the feature `pos_sequence` records the length of the longest sequence of PoS tags common to responses and questions.

6.1.1 Experiment I: Classification with Over-sampling Method on Full Taxonomy

Data resampling is one of the most widely used methods for dealing with the imbalanced data problem. In this method, training instances are modified in order to produce a more balanced class distribution. One advantage of resampling techniques over other methods is that they are independent of the classifiers (López et al., 2013). The resampling techniques are mainly divided into two groups:

- `Undersampling methods`: this method generates a subset of the original dataset by deleting instances from the majority class. Random undersampling is a very simple non-heuristic method that randomly removes samples from the majority class. However, the drawback of random undersampling is that it may drop some potentially useful data that could be important for the classification.
- `Oversampling methods`: this method outputs a superset of the original dataset through replicating instances from minority classes. The non-heuristic simple random oversampling method balances the class distribution by randomly making exact copies of existing instances of the minority class. Therefore, the disadvantage of random oversampling is that it may cause overfitting.

In this study, we use the `SVM-SMOTE` over-sampling algorithms in the `imbalanced-learn` python package (Lemaître et al., 2017). We do not consider using the under-sampling method because we do not have a huge amount of annotated data at this stage. `SVM-SMOTE` is a special variant of SMOTE algorithm (Chawla et al., 2003), which use an SVM algorithm to detect sample to use for generating new synthetic samples. This over-sampling algorithm resampled all response classes except from the majority class – Direct Answers.

For the classical machine learning task, we use the Support Vector Machine (SVM) classifier from the Scikit-learn library (Pedregosa et al., 2011; Buitinck et al., 2013). The Support Vector Classifier (SVC) internally always uses one-vs-one ('ovo') as a multi-class strategy to train models. However, we use the One-vs-Rest ('ovr') to return the decision function of shape (n_samples, n_classes) as all other classifiers. The One-vs-Rest ('ovr') method turns a multi-class classification into one binary classification problem per class. In addition, the balanced class-weights are used due to the imbalanced characteristics of our data sets.

Evaluation metrics: we report the classification results based on the precision, recall, and f1-score for each response class. Besides, we also show the average classification accuracy of all classes, macro average scores, and also the weighted average scores of precision, recall, and f1-score. Finally, we also present the average accuracy score resulting from 5-fold cross-validation, and also the Area Under the Receiver Operating Characteristic Curve (roc_auc_score) from prediction scores. Again, we use the One-vs-rest configuration to compute the AUC of each class against the rest. This 'ovr' method is sensitive to class imbalance, so it is more suitable for our imbalanced dataset.

Experimental results: Table 5 presents the classification performance of the SVM classifier on

Classes	Precision	Recall	F1	Support
DA	0.73	0.90	0.81	593
IND	0.38	0.19	0.25	107
DP	0.27	0.60	0.37	5
CR	0.67	0.77	0.71	47
ACK	0.33	0.58	0.42	31
IGNORE	0.33	0.02	0.04	44
CHT	0.38	0.09	0.14	56
MOTIV	0.00	0.00	0.00	3
DPR	0.85	0.34	0.49	50
accuracy			0.68	936
macro avg.	0.44	0.39	0.36	936
weighted avg.	0.64	0.68	0.63	936
SVM cv scores				0.85
roc_auc_score				0.79

Table 5: Classification results of SVM classifier on the full response space taxonomy with oversampling

Classes	Precision	Recall	F1	Support
DA	0.72	0.89	0.79	595
IND	0.42	0.04	0.07	126
CR	0.69	0.83	0.76	41
Evasion	0.43	0.34	0.38	171
accuracy			0.67	933
macro avg.	0.57	0.52	0.50	933
weighted avg.	0.62	0.67	0.62	933
SVM cv scores				0.82
roc_auc_score				0.79

Table 6: Classification results of SVM classifier on the coarser response space taxonomy with oversampling

the full response space taxonomy using the SVM-SMOTE oversampling method. As shown in the table, the SVM classifier achieved similar classification results as from the Bert model, in terms of weighted f1-score (0.63 – 0.68) and the macro f1-score (0.36 – 0.40) on the full response space taxonomy. The SVM classifier also performed well on some major response classes, such as Direct Answers (f1-score: 0.81) and Clarification Responses (f1-score: 0.71). However, despite the relatively high frequency of Indirect answers, both models did not perform well on identifying these response classes (f1-score: BERT - 0.32, SVM - 0.25). The overall trend of the classification results for other response class is also similar on both methods. Namely, the response classes such as IGNORE, MOTIV, and CHT are always the most difficult classes for both SVM classifier and BERT models. Moreover, both models can correctly capture nearly half the cases from Acknowledgments and DPR classes. Therefore, we argue that the feature sets designed to capture syntactic and lexical characteristics of responses and the corresponding questions are useful for recognizing some response classes, by merely using the most classical machine learning algorithms.

In addition, we also report the average accuracy from 5-fold cross validation during the training, and also the final roc_auc_score for the SVM classifier on the full taxonomy. The average accuracy from the cross-validation is 0.85%, and the roc_auc score is 0.79, which indicates a very good performance of our classifier. Since the roc_auc score is not affected by the imbalanced distribution of each class in the dataset, we think that roc_auc_score metric can better describe our model on response space classification task with a highly skewed dataset.

6.1.2 Experiment II: Classification with Over-sampling method on a Coarser Taxonomy

In the previous sections, we studied the automatic classification of 9 different response classes as described in Table 2. In this section, we are interested in studying the classification performance of the SVM classifier on a coarser response space taxonomy with only 4 distinct response classes, namely, Direct Answers, Indirect Answers, Clarification Responses, and Evasion.

As shown in Table 6, when classifying with a coarser taxonomy, the SVM classifier achieved a better macro average f1-score than on the full taxonomy (0.50 vs. 0.36). However, when compared to the results achieved by the BERT model (see Table 3) on the coarser taxonomy, the SVM model resulted in a lower weighted average f1-score (0.62 vs. 0.72) and macro average f1-score (0.50 vs. 0.60). The average accuracy for the 5-fold cross-validation while training is 0.82, and the roc_auc score is 0.79, which indicates a good performance of the SVM model. What is more, the overall trend of the classification results for each response class is similar to both the SVM model and the BERT model. Both models achieved similar high f1-scores for the Direct Answers, 0.79 and 0.85 respectively for the SVM and the BERT model. The second-highest performance score goes to the Clarification Responses on both models: f1-score is 0.76, and this is where our SVM model outperforms the BERT model (f1-score is 0.67 for Clarification Responses). However, the SVM model still failed to capture Indirect Answers and returned a 0.07 f1-score for this class. This is much worse than the f1-score of 0.30 achieved by the BERT model. Finally, the Evasion response class also

caused many difficulties for both models, which resulted in f1-scores of 0.38 and 0.60 from the SVM and BERT model.

To conclude, regardless of the full or the coarser taxonomy, the DA response class is learned more easily by both pre-trained BERT language model and the classical machine learning algorithms. Whereas Indirect Answers, IGNORE, and MOTIV cause most difficulties for both models. In addition, the SVM model outperforms the BERT model on identifying Clarification Responses on this coarser taxonomy. Besides, the similar classification trend for each response class on both models suggests that the carefully designed feature sets are useful to capture the main response classes.

7 Conclusions and Future Work

We present a pilot study on the novel task of response space classification of questions in dialogue. We considered the classification results by the large scale pre-trained BERT language model with raw data (questions and responses) as baselines, and conducted experiments with more classical machine learning algorithms (the SVM classifier from the Scikit-learn library). We utilized 26 carefully designed syntactic and lexical features on the SVM classifier, which aim to capture characteristics of responses and question. Since the class distribution in our datasets is highly imbalanced, we first deployed an over-resampling methods to mitigate the imbalanced data problem. Then, we conducted two groups of experiments respectively on both BERT and SVM models: (1) with a fine-grained full response space taxonomy with 9 unique response classes, and (2) with a coarser taxonomy with only 4 main response classes. Finally, we compared the classification results from both models and offered detailed discussions regarding the differences and similarities observed from two models.

The main contributions of this study are threefold: (1) To our knowledge, this is the first study on the automatic classification of response space of questions in dialogue. Such a classification task is of great importance in the design of dialogue systems, particularly systems that can be easily adaptable across domains. (2) We designed 26 different features which help the classical machine learning algorithms to correctly identify different response classes; (3) We provided detailed discussion of the learnability of various response classes by the pre-trained language model and the classical SVM classifier, and observed that the learnability trend is closely in line with that achieved by the human annotators in previous work.

However, we also acknowledge the limitations of the current study and have some initial thoughts for future studies. Firstly, we hope to scale-up the current feature sets used for the SVM model by designing more useful features in terms of syntactic, semantic, and lexical relationships between questions and responses. Secondly, since dialogues are highly context-dependent interactions, we also want to conduct experiments by adding features pertaining to such aspects to the feature set, e.g., the number of common words between previous utterances and questions/responses, the length of the previous utterances etc. Thirdly, a detailed analysis of which features are more informative and which are redundant can also be very useful for the classification task. Lastly, more carefully created features targeting Indirect Answers are necessary to correctly classify this highly inference-based response class.

Acknowledgments

We acknowledge the support by a public grant overseen by the French National Research Agency (ANR) as part of the program Investissements d'Avenir (reference: ANR-10-LABX-0083). It contributes to the IdEx Université Paris Cité ANR-18-IDEX-0001. We also acknowledge that the second author is supported by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We thank the SemDial reviewers for very helpful comments.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth H. Boyle, Gwyneth M. Doherty, Simon C. Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD*

Workshop: Languages for Data Mining and Machine Learning, pages 108–122.

Lou Burnard, editor. 2007. *Reference guide for the British National Corpus (XML Edition)*. Oxford University Computing Services on behalf of the BNC Consortium. Acess 20.03.2017.

Jean Carletta. 1996. Assessing agreement on classification task: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. 2003. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer.

William W Cohen and Yoram Singer. 1999. A simple, fast, and effective rule learner. *AAAI/IAAI*, 99(335-342):3.

Maria-Andrea Cruz-Blandón, Gosse Minnema, Aria Nourbakhsh, Maria Boritchev, and Maxime Amblard. 2019. Toward dialogue modeling: A semantic annotation scheme for questions and answers. *arXiv preprint arXiv:1908.09921*.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. Timbl: Tilburg memory based learner, v. 5.0. Technical report, Reference Guide. Technical Report ILK-0310, University of Tilburg.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pages 76–87. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paolo Dragone and Pierre Lison. 2015. An active learning approach to the classification of non-sentential utterances. In *Proceedings of the Second Italian Conference on Computational Linguistics*, pages 115–119.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, Aleksandra Kucharska, and Paweł Łupkowski. 2022. Characterizing the response space of questions: data and theory. *Dialogue and Discourse (accepted)*. https://drive.google.com/file/d/1AieL7JERQhJnTPlbgn1P_YPDaLP8gGJl/view.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, and Paweł Łupkowski. 2019. Characterizing the response space of questions: a corpus study for english and polish. In *Proceedings of the 20th annual SIGdial meeting on discourse and dialogue*, pages 320–330.

Jacqueline C. Kowtko and Patti J. Price. 1989. Data collection and analysis in the air travel planning domain. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '89, pages 119–125, Stroudsburg, PA, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112.

Staffan Larsson and Alexander Berman. 2016. Domain-specific and general syntax and semantics in the talkamatic dialogue manager. *Empirical Issues in Syntax and Semantics*, 11:91–110.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141.

Paweł Łupkowski and Jonathan Ginzburg. 2013. A corpus-based taxonomy of question responses. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 354–361, Potsdam, Germany. Association for Computational Linguistics.

Paweł Łupkowski and Jonathan Ginzburg. 2016. Query responses. *Journal of Language Modelling*, 4(2):245–293.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, third edition. Lawrence Erlbaum Associates, Mahwah, NJ.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Purver. 2006. Clarie: Handling clarification requests in a dialogue system. *Research on Language and Computation*, 4(2):259–288.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003a. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Matthew Purver, Patrick Healey, James King, Jonathan Ginzburg, and Greg J Mills. 2003b. Answering clarification questions. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 23–33.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*. Citeseer.

Carolyn P. Rosé, Barbara Di Eugenio, and Johanna D. Moore. 1999. A dialogue-based tutoring system for basic electricity and electronics. In Susanne P. Lajoie and Martial Vivet, editors, *Artificial intelligence in education*, pages 759–761. IOS, Amsterdam.

Ian H Witten and Eibe Frank. 2002. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

L Zhang. 2007. Maximum entropy modeling toolkit for python and c++ (online). *http://homepages. inf. ed. ac. uk/s0450736/maxent_toolkit. html*.

The Symbol Grounding Problem Re-framed as Concreteness-Abstractness Learned through Spoken Interaction

Casey Kennington
Boise State University
1910 W University DR
Boise, ID 83725
caseykennington@
boisestate.edu

Osama Natouf
Boise State University
1910 W University DR
Boise, ID 83725
osamanatouf@
u.boisestate.edu

Abstract

The Symbol Grounding Problem points out that the underlying mechanisms of computation are symbolic and, therefore, missing crucial information when they are used for processing natural language until they are somehow able to perceive the world directly. Our goal in this paper is twofold: First, we review some of the recent literature that claims to address (even if just to a small degree) the Symbol Grounding Problem, and explain why it is still yet a problem partially due to a misinterpretation of the problem and that there are more modalilties that symbols need to ground into beyond just pictures, including emotion. Second, we re-frame the problem as a problem of handling concreteness and abstractness because (perhaps surprisingly) computational models of distributional meaning seem to capture abstractness more directly than they do concreteness. We take inspiration from child development and offer a toy example of how one could approach modeling concrete and progressively more abstract words. We conclude by posing some open questions and offering paths for future work.

1 Introduction

The *Symbol Grounding Problem* posits that linguistic meaning cannot be directly encoded in a computational symbol, particularly because the meanings of many words are *grounded* in real-world experience (Harnad, 1990). For example, the word *blue* is a color, but so is *red*, so knowing that they are in the same category of words does not uncover their meaning because both denote different swathes of the color spectrum that is visible to humans, and without experiencing each word used in physical contexts of other people denoting those colors, it is impossible to learn each word's connotation.

Harnad (1990) identified properties of symbolic systems; for example that there are atomic symbols and composed symbol combinations, and that symbols (can be) semantically interpretable, but this is in contrast to how humans can discriminate, manipulate, identify, and describe objects, and humans can even respond to the objects and descriptions of those objects. Put another way, humans interact with and talk about the world, and the cognitive capabilities that humans have are a result of the fact that they do so (Smith and Gasser, 2005).

A recent neurological study gives empirical backing to this proposition where the authors "assessed the extent to which different representational systems contribute to the instantiation of lexical concepts in high-level, heteromodal cortical areas previously associated with semantic cognition" (Fernandino et al., 2022). Their work showed that, though semantic information can be represented by distributional representations and symbolic taxonomies, a clear advantage exists for "experiential representational structures" such as sensory-motor, affective, and other features of phenomenal experience, suggesting that if research is to solve the problem of acquiring, representing, and applying linguistic meaning computationally, then to learn a word's semantics means access to experience with the world.

More than 30 years have elapsed since Harnad (1990), now with over 5,000 citations which suggests, at the very least, that the problem has been considered and taken seriously by scientists. However, fifteen years after the original publication, Taddeo and Floridi (2005) reviewed the literature of the time and concluded that at that point, the problem was far from solved. Since then, the "collectivist" models that seemed promising at the time have evolved to deep learning models that have proven their power on language tasks, with some models showing promise on language and vision tasks. Does this mean the problem has been solved, or will be soon, given the right deep learner?

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

In this paper, we explore some of the recent work on symbol grounding. We observe that the way language is currently modeled suffers from a similar problem that symbolic systems suffered: they are ungrounded (Section 2). Moreover, though vision is an important modality for symbol grounding, it is not the only important one (including emotion). In Section 5 we identify other modalities that are often ignored, but must be part of any model that claims to be holistic. However, not all words need to be grounded into in order to arrive at their meaning; we therefore re-frame the Symbol Grounding Problem in light of an important distinction between concrete and abstract concepts in Section 3 which, we believe, have implications for how meaning can be modeled in existing deep learners. We conclude by offering some suggestions for avenues of future research.

2 The challenge of symbol grounding

Harnad pointed to Searle (1980)'s Chinese Room as a metaphor which challenges the core assumptions that symbols carry meaning on their own. He explains that if he, someone who could not read or speak Chinese, were in a room with a Chinese-Chinese dictionary and had the instuctions to take "input" of one Chinese character, look up the character in the dictionary and then find the "output" character, even if the inputs and outputs were perfectly mapped as observed by an outsider, the person in the room doesn't actually know Chinese, which is the same problem that computers have when they process natural human language.

Yet are words not also symbols? In some ways yes, but we need to be clear here what is meant by *word*. A word is a linguistic unit that carries linguistic meaning on its own and can be used as a placeholder for a concept much like symbols can. For example the word *chair* can denote real chairs, but uttering or writing the word can replace the presence of chairs when someone wishes to talk about the concept of a chair—the word *chair* effectively becomes an abstraction of the connotation. The confusion comes when one assumes that the word *chair* as it is written actually represents the concept itself, but it does not; the concept of *chair* resides in human brains, but because written text is computable and since text is a placeholder for concepts for humans as they communicate with each other, it follows that machines could use text as symbols and text would carry the meaning, but that is precisely what the Symbol Grounding Problem is pointing out does not work because, like symbols, text is ungrounded.

Since 1990, other models of learning and representing linguistic meaning that go beyond the kinds of symbols that Harnad was referring to, most notable embeddings and language models that follow the *distributional hypothesis*, a hypothesis that posits that the meaning of a word can be derived by how it is used in the context of other words within text; the Firthian "you shall know a word by the company it keeps" generally means that words keep company with other words. This led to models such as word embedding vectors (Mikolov et al., 2015) and, more recently, powerful transformer-based language models like BERT (Devlin et al., 2018) that are trained on text alone; the training regime is often a task of guess-the-masked-word in a context of other words. These models both in their time have revolutionized entire research fields. Have they solved the Symbol Grounding Problem?

With attention now to language models instead of symbols, and building on Searle's Chinese Room thought experiment, Bender and Koller (2020) argue that language models do not learn meaning on similar grounds as the Symbol Grounding Problem. They offer *the octopus test* where an octopus "overhears" a conversation between two people on desert islands by tapping into the communication wire that connects them. The octopus learns how to mimic one of the dialogue partners by learning regularities in the kinds of words and phrases they use, and when the octopus has an opportunity to take over the role of that particular dialogue partner, despite being able to learn patterns of words and how they should appear in the context of other words, the octopus cannot answer simple questions because the octopus fails to know the kinds of objects that certain words denote. In other words, despite their success, models that follow from the distributional hypothesis also fail at solving the Symbol Grounding Problem. Furthermore, Herbelot (2013) makes a strong case that text alone cannot possibly be expected to contain the meaning of many words, no matter how much text is used for training. Clearly, however, some degree of meaning can be derived and represented from text, otherwise language models could not possibly work so well on so many natural language processing tasks, which begs what kind of information they are learning (see Rogers et al. (2020) for a review), though it is

clearly not grounded.

Dictionaries, likewise, do not solve the Symbol Grounding Problem even if each word in the dictionary has a corresponding definition that is intended to represent meaning of words, or at least the description of the meaning of words. Harnad explicitly mentioned the "dictionary merry-go-round" of words defining other words, a claim that was empirically tested in Vincent-Lamarre et al. (2016) (work by Harnad and colleagues) who identified a subset of words that all other words are eventually defined by, showing that defining words by other words is indeed useful, but do not capture holistic meaning. Conversely, not all words need to be grounded—meaning can be derived from other words in many cases. The challenge is determining which word meanings that should ground into the physical world and which word meanings that should be derived from lexical context (i.e., text).

3 Reframing the problem: concreteness & abstractness

We argue that framing the Symbol Grounding Problem as a question of *concreteness* vs. *abstractness* puts the research field on better theoretical footing to make the best of what is required for solving the Symbol Grounding Problem and existing computational models that derive meaning from distributional approaches using text. In this section, we explain and give examples of concreteness and abstractness, argue that no current model captures both, and perform a small scale toy experiment to explore what a model that does capture both might look like.

Concrete words are words that denote physical things like objects, shape, and color (e.g., *chair*, *red*), requiring Symbol Grounding to arrive at meaning, whereas abstract words are words that denote ideas (e.g., *democracy*, *travel*), but it should be noted that the distinction between concrete and abstract concepts lies on a continuum, not a binary dichotomy (Della Rosa et al., 2010; Brysbaert et al., 2014). Thus some words are more concrete or abstract than others, some examples that illustrate this are shown in Figure 1. Words range from very concrete (e.g., *ball*) to very abstract (e.g., *utopia*). For more concrete words, corresponding images show clear examples of something that the word can denote visually. However, more abstract words can have aspects of their meaning represented visually, but not fully (e.g., *democracy* includes voting, but voting is only one aspect of the meaning of *democracy*).

That some words need grounding while others do not begs the question *Which words need symbol grounding?* Words that are more concrete like *ball* and *red* clearly need to be grounded. The word *red*, for example, can be understood to some degree without grounding, for example that it is a color and that certain objects can be red (e.g., apples and vehicles), and while it is true that there are metaphorical uses for the word *red*, those metaphorical uses can only be understood after knowledge about *red* as a color is learned (see arguments made in Lakoff and Johnson (2008) about metaphors; see also Bizzoni and Dobnik for discussion on visually grounded metaphors). A fairly simple grounding strategy could be used at the word level to arrive at a grounded representation, for example Schlangen et al. (2016) where each word in a corpus was represented by a binary classifier; the inputs of which were visual features. The model, however, assumes that all words are in fact concrete and visually grounded.

On the other end of the continuum are abstract words like *democracy* and *utopia*. Even though someone could imagine a visual depiction of either of those terms, their meaning is not grounded directly into the physical world, but are rather ideas that are defined by other words. Because the meaning of abstract words can be defined by other words, it is the meaning of abstract words that is captured by distributional methods, such as recent language models like BERT (Devlin et al., 2018). Distributional approaches, as noted above, are trained on text and make the tacit assumption that all words are abstract—ungroudned—even words that show up in the text that are in reality concrete are assumed model to be abstract in how the model captures meaning.

Is there a model that can capture both concreteness and abstractness? The real challenge comes from words that are not obviously concrete nor obviously abstract, rather somewhere in between like *farm*, or *color*. A farm can be observed and denoted visually, but what makes a farm a farm is not represented by an image or a series of images, but rather specific (abstract) properties like growing crops or keeping livestock within a specified land area. However, the words that are required for one to understand the concept of *farm*, one must understand what crops, land, and livestock are, concepts that

Figure 1: Examples of words that are more concrete vs. more abstract. Words that are concrete have physical (in this case, visual) denotations, whereas more abstract words do not physically exist. Concreteness ratings from Brysbaert et al. (2014) resulted in the placement of the words.

are themselves to some degree more concrete and grounded. The other example, *color* might be more illustrative: as a concept, *color* seems concrete because it is a very visual concept that categorizes colors.[1] The meaning of the word *color* can be defined by other terms, but the function of the word itself is to distinguish between other words that are considered colors like *red* and *blue* and words that are not. Thus while *color* itself does not directly ground into the visual world, it does directly connect somehow to words that in turn are grounded in the visual world. So should *color* be learned as an abstract concept or one that categorizes concrete concepts? In the following section we explore the latter with a simple toy example using a handful of categories and related words for each category.

3.1 A toy experiment: grounding into concrete words meanings

We conduct here a small experiment to test the possibility that concrete words can be "grounded into" by more abstract words that are higher on the abstractness scale, where what is grounded into differs depending on the level of abstraction. We use the following train and test set vocabularies for five "abstract" categories (i.e., not fully concrete); each item begins with with a more abstract word in boldface that is a grouping of the other words, which are all more concrete. Note that none of the test words are also in the training set.

Train:

- **color**: red, blue, green, yellow, brown
- **animal**: dog, cow, cat, mouse, bird
- **furniture**: couch, chair, desk, bed
- **vehicle**: car, van, truck, pickup, tractor
- **appliance**: stove, oven, microwave

Test:

- **color**: orange, purple
- **horse**, sheep
- **furniture**: table, sofa
- **vehicle**: taxi, jeep
- **appliance**: mixer

Procedure Following the words-as-classifiers (WAC) approach to grounded semantics (Schlangen et al., 2016), we train a logistic regression classifier (C=0.25, max-iter=1000) for each concrete word using images (we downloaded top 100 images for each word as a search term using Google Image Search) that have been passed as input into the CLIP model (Jia et al., 2021) which yields a vector of size 512 for each image. Negative examples of each word are randomly sampled from images for other words; we use three negative examples for each positive example. This results in a trained binary classifier for each concrete word that can, given a new image (i.e., represented as a CLIP vector), determine how well a trained classifier for a word *fits* the image. This is depicted in the top portion of Figure 2. For example, a trained classifier for *red*, given an image with a lot of red in it, would return a higher probability than if the image had little or no red in it. With our toy example, we therefore have 30 trained classifiers for each of the concrete words in both the train and test sets.

We then train classifiers for the more abstract words (i.e., *color, animal, furniture, vehicle, appliance*) that are defined by how they group together corresponding concrete words in a similar

[1] The concreteness rating for *color* in Brysbaert et al. (2014) is 4.08, which makes it a fairly abstract, compared to *democracy* (1.78) and *chair* (4.58)—higher numbers denote higher concreteness. *Color* is only slightly more abstract than *chair*.

way, though using different features. We hypothesize that the groupings are based on the feature sets that are common to the different categories. We therefore use the *coefficients* of the trained concrete classifiers as input to the abstract word classifiers because, as pointed out by Schlangen et al. (2016), the trained classifiers themselves (which are the logistic function and corresponding coefficients) represent a computational intension of each word; positive examples are the words listed for a category, negative examples randomly sampled from from the other categories; three negative examples for each positive. This is partially depicted in Figure 2 that shows how coefficients from the trained *red* classifier are features for the *color* classifier.

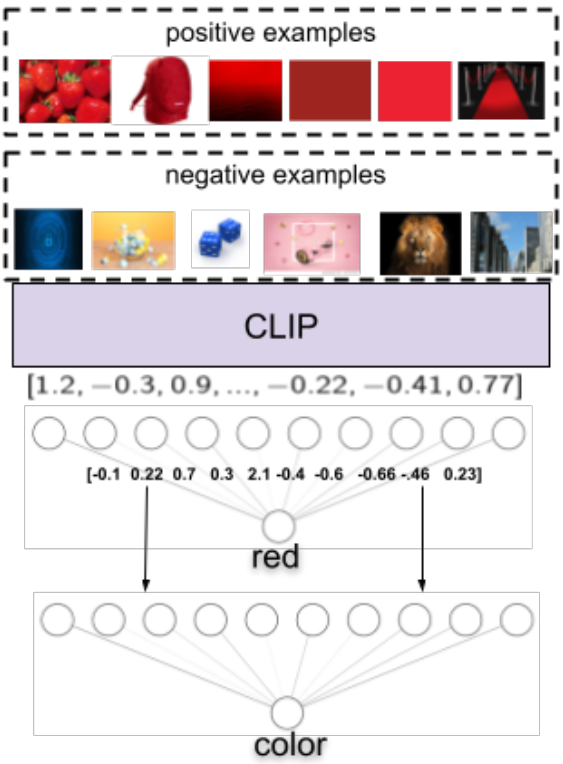


Figure 2: Example of *red* WAC classifier trained on positive negative examples of images represented by CLIP vectors. The coefficients for the *red* classifier are a positive input for the *color* classifier. Figure adapted from Kennington (2021) with permission from the author.

Task & Results We task the five trained abstract category classifiers to correctly classify the words in the test set, resulting in an accuracy metric. To evaluate, we take the concrete words for the test set, then pass their coefficients as input to each of the five classifiers trained for each category and compare the category with the highest probability to the labeled category. The test set only has nine words, but even with a small training set, our evaluation yields 88% accuracy (the only mis-classified word was *mixer* which was mis-classified as *furniture* instead of *appliance*).

This toy experiment shows that it could be possible to build meanings of words that are somewhat abstract by grounding them into words that are concrete by treating coefficients as a level of abstraction, which may solve the grounding problem for some of the words that are closer to the concrete side of the concrete-abstract continuum. One major limitation is that the model assumes that all words are independent of each other in how they are trained; e.g., the word *color* has no knowledge about *appliance* even though appliances can have color—knowledge that could be picked up from text using distributional methods. Another limitation is determining which words are used as positive examples to a more abstract word. In the toy dataset above, the categories are clear, but it is unlikely that an abstract word's meaning can be derived from the coefficients of the words in its dictionary definition as positive examples. In the following section, we explore how one might arrive at a model that does learn from concrete to abstract over time by taking inspiration from child development.

4 Learning meaning from concrete to abstract

With the Symbol Grounding Problem recast as a problem of grounding for concrete words directly into perception, and more abstract words into some kind of meaning representation of more concrete words, and building on the successes of the distributional approaches to modeling language, we now consider what it takes to learn concrete words, followed by more abstract words that build upon them. To do so, we take inspiration from child development where the setting of spoken interaction is crucial to learning concrete terms initially, and how emotion is integral to the process of learning language and is part linguistic meaning itself.

4.1 The setting of spoken interaction

Complementary to Symbol Grounding, *Communicative Grounding* is the process of mediating what words mean between an individual and a language community through active use of language (Clark, 1996). Communicative Grounding is cru-

cial to Symbol Grounding because, as argued by Larsson (2018), Symbol Grounding is a side effect of Communicative Grounding. To illustrate: if two individuals are sitting together in a park and actively observe a kite in the air, and one person utters *kite*, the other person who had never seen a kite before now grounds the word *kite* with the observed object (symbol grounding), and both individuals know that each other has taken part in the interaction (communicative grounding).

Following Kennington (2021), children learn their first language in this highly interactive setting where communicative grounding between caregiver and child takes place as a facilitator for Symbol Grounding, and speech is the primary modality of linguistic interaction. At this early language learning stage, children generally learn words that denote physical objects making them largely concrete (Kuperman et al., 2012; Clark, 2013; Borghi et al., 2019; Ponari et al., 2018). Furthermore, Locke (1995) makes a case that that putting an agent (or, we conjecture, a computational model) in a place where it can only *observe* language—be it text or even referring expressions made to visually present objects—does not bring the child (or a computational models) to language capabilities as much as *participatory interaction*.[2] Before children can comprehend or utter words that carry semantic content in a given language, they experience the world in a profoundly multimodal and interactive setting (Smith and Gasser, 2005), giving children existing experience with the physical world that they can later leverage when learning their first words by categorizing perceptual input and grounding word concepts to those groupings.

Moreover, children tend to move in a learning progression of concrete to abstract over time: Borghi et al. (2019) notes that data indicate that only 10% of the vocabulary of 4-year-olds is composed of abstract words, abstract words represent 25% of 5-year-olds' words and more than 40% of 12-year-olds' vocabulary (see also Ponari et al. (2018)). Put succinctly, the words that children first learn largely require symbol grounding, but meanings of later words that are more abstract can be learned by how they are defined by and used with other words distributionally. This is not to claim in any way that children only learn concrete words early in life, then move to learning only abstract words—humans learn new concrete terms throughout life, and children begin to learn fairly abstract concepts early in development (e.g., greetings). Furthermore, this is not to say that cognition is purely a bottom-up process; clearly there is some degree of cognitive processing that is top down—the natural process of categorization of sensory input is an integral part of cognition whether the categories are innate or not (Harnad, 2017).[3]

4.2 Concrete-affect; abstract-emotion

Missing from the discussion thus far in language learning—both concrete and abstract—is how emotion plays a role that works in parallel to the concrete-to-abstract language learning progression. Early on in a longitudinal project (Alan Sroufe et al., 2009), the authors note that cognitive advances "promote exploration, social development, and the differentiation of affect; and affective-social growth leads cognitive development [...] neither the cognitive nor the affective system can be considered dominant or more basic than the other; they are inseparable manifestations of the same integrated process [...] It is as valid to say that cognition is in the service of affect as to say that affect reflects cognitive processes." In other words, cognition is not disconnected from emotion. Locke (1995) agrees, while tying emotion directly to language: in the real speech of sophisticated speakers, where both linguistic content and vocal affect are present, one type of cue does not preempt the other–and for speech to work this must be the case. Listeners must know both what the speaker is saying and what they intend by saying it. Humans duplexly pick up information about the linguistic content *and* the speaker's affect because the cues to these things are of different sorts and are processed by different brain mechanisms—this is particularly important for children who are learning their first words. Thus, according to Locke, the meaning of an utterance is in the linguistic content, but the *intent* of the speaker who made the utterance is also in the affect and emotion. In fact, children are adept at reading intents of others via affect and emotion, before they can even speak or really understand words (Smith and Gasser, 2005). This

[2] Sachs et al. (1981) explained that two children with normal hearing were born to deaf parents, so the parents did not use speech interaction with their children. Despite watching television with programming for children, their speaking abilities were far behind their peers, which required intervention.

[3] Missing from this discussion is how *affordances* affect perception and categorization, but note that understanding object affordances are an important part of the concept learning process.

suggests that emotional states exist within humans before they can speak; indeed, emotions can facilitate the language learning process for someone who is learning their earliest words (McNeill and Kennington, 2020).

Furthermore, recent empirical work in neuroscience and cognitive science have explored the relationship between language and emotion. Lane and Nadel (2002) explained that the meaning of many words has emotion as part of their connotation, and Mazzuca et al. (2018); Villani et al. (2021) have shown that abstract linguistic concepts are more closely tied to emotion (i.e., interoception) in particular emotional and mental states, as well as social concepts than concrete linguistic concepts are tied to emotion. Moreover, Ponari et al. (2018) showed that the acquisition of abstract concepts is influenced by emotional valence, particularly for children who are at a stage where they are learning abstract words (e.g., 40% of a 12 year old's vocabulary is made up of abstract terms (Borghi et al., 2019), see above). This explains, we conjecture, to some degree why sentiment and emotional valence can be inferred from text in natural language processing tasks, but similar to symbol grounding, emotional valence is inferred from the text, not encoded within it.

Taken together, this suggests that the separation of language from emotion in computational models is going to lead to something that is only an approximation of what a model of language meaning should encode and in that way it is similar to the Symbol Grounding Problem. However, emotion is not just another modality like vision through a camera or haptic sensations through a robotic hand; emotion is communicative on its own, albeit with limited (but important) social signals; pre-linguistic in that it helps scaffold the language learning process especially early on, and emotion is later intertwined with cognitive development and linguistic meaning at an abstract level. Dreyer and Pulvermüller (2018) suggests that representing emotion computationally could be done through the motor system, as done in Moro et al. (2020), which may offer a starting point for bringing emotion into computational models of language (instead of the other way around).

5 Open questions

Resolving the Symbol Grounding Problem has seen real progress, in particular with vision (see below), but it is far from completely solved. There are many modalities to be explored beyond vision, and it is unlikely that the research field will arrive at a solution to representing meaning computationally without some kind of representation of an approximation of emotion. Given the implications of the above sections, in this section we discuss the fact that (besides emotion, discussed above) there are modalities besides just vision that need to be grounded into, pose some open questions, and offer some next steps for the research community.

5.1 Modality questions

Some language models do attempt to model language and vision directly to solve language and vision tasks, for example VilBERT (Lu et al., 2019), CLIP (Jia et al., 2021), FLAMINGO (Alayrac et al., 2022), Dalle 2 (Ramesh et al., 2022), and others. These models are impressive compared to our toy example, but recent work has shown that the models do not quite learn a vision-language mapping in a way that, we argue, actually addresses the Symbol Grounding Problem (Parcalabescu et al., 2020, 2021; Marcus et al., 2022). These language and vision models often force the addition of visual information through robust object detection models that do not capture the true grounding of the words; rather the representation of visual perception is represented symbolically by class labels of the object detection model, but most object detection models do not capture words beyond objects (i.e., nouns). Words like *left* or *red* are also concrete words that an object detection (or region detection) model should not ignore. Moreover, Hendricks et al. (2021) explained that the *quality* of the language (i.e., text) highly affects the visual language models' performance, which seems to suggest that a curriculum not unlike Xu et al. (2020), i.e., by using a training regime that learn with simpler examples first (e.g., that refer to visual objects) then move towards more complex and more abstract examples of language use.

Most recent work has focused on vision, thanks in part to datasets that connect language and vision, but vision isn't the only important modality that humans have access to for grounding linguistic meaning (see (Fernandino et al., 2022; Lynott et al., 2019)): and some have explored grounding into other modalities including modalities that sense the external world like olfactory (Kiela et al., 2015) and sound (Thomason et al., 2018), but also "internal"

(i.e., within the body) modalities such as haptics (Thomason et al., 2018), proprioperception (Moro and Kennington, 2018) and interoception (Moro et al., 2020) (i.e., affect & emotion) as well as in spoken interaction itself (see (Larsson, 2018)).

Grounding into external modalities requires some kind of sensor (e.g., cameras for vision and microphones for sound), but more challenging is grounding into internal modalities like haptics, proprioperception, and interoception because for those some kind of embodiment (e.g., robot or virtual agent) is required. We do not explore here which might be better for computationally modeling linguistic meaning, but, following our inspiration from child development above, we make an obvious point that children have bodies that house the sensors and internal modalities that they use to interact with objects and people in the world. Embodied cognition is not a new idea, but given the discussion above, embodiment may be a requirement for capturing holistic linguistic meaning computationally (Barsalou, 2008; Johnson, 2008; Bisk et al., 2020) and embodiment is not in disagreement with solving the Symbol Grounding Problem. The model described in Hill et al. (2020) may be a step in the right direction, though it remains unclear what degree of concreteness or abstraction the model is learning.

5.2 Modeling questions

Much of the recent literature uses vectors and tensors (i.e., within language models) to computationally represent meaning (grounded or ungrounded), which are convenient for hardware that can parallelize computation of such representations, but are vectors the right representation for learning and modeling meaning, particularly meaning that addresses the Symbol Grounding Problem? One possible alternative are *cognitive architectures*. Is it time to work with cognitive scientists and apply their cognitive architectures in spoken, person-to-person interactive settings? Developmental robotics as a field have done so to an extent (Cangelosi and Schlesinger, 2015), and if we are coming to similar conclusions that embodiment may be necessary, but at they very least interactive learning and sensors are required, then it may be prudent to bring more cognitive scientists into the discussion, where possible. More related to concreteness and abstractness, roboticists have worked on making robot actions composed aggregates of smaller, more concrete actions, which may have implications for modeling language.

5.3 Philosophical Questions

It is clear that when Firth posited that meanings of words can be found in the company they keep, the "company" that Firth meant was company with other words, and researchers often cite Wittgenstein for *language is use in context* which always assumes that *context* means lexical context with other words, but Wittgenstein (2010) brings up color and shape (1.72-74) and that words refer to objects, which themselves have affordances (1.11), and early on mentions that language use is first in reference to deictic (i.e. pointing) gestures. Could Wittgenstein have meant that *context* is not lexical context, but physical context (or some degree of both)? This is an important question because Firth and Wittgenstein have always been called on to motivate distributional methods of language modeling, but words keep company with more than just other words, including words that are more concrete.

6 Conclusion

In this paper we attempted to re-frame the Symbol Grounding Problem as a problem of modeling and learning word meanings from concrete as well as abstract words. How meaning of concrete words are modeled and learned follows directly from symbol grounding, and more abstract words could be learned distributionally.

We will build on our toy example in a large-scale experiment by learning classifiers that are not specifically tied to any known grouping of words, but rather are bottom-up grouping of concepts that are linked to works that are later "heard" by the training regime (e.g., in a similar way that someone may know that colors group together based on their features, but do not yet know the word *color*).

We will also explore how such trained classifiers could be combined with existing language model architectures like BERT. Recent work by Kennington (2021) showed how extracting coefficients from visually-grounded classifiers could enrich a language model, but the enriching took place only in the language model's embedding layer with the assumption that all words were concrete. We will explore using concreteness ratings as a possible signal to determine whether a word's meaning should come from a grounded model or a language model.

Acknowledgements Thanks to the anonymous reviewers for their very useful feedback.

References

L Alan Sroufe, Byron Egeland, Elizabeth A Carlson, and W Andrew Collins. 2009. *The Development of the Person: The Minnesota Study of Risk and Adaptation from Birth to Adulthood*. Guilford Press.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for Few-Shot learning.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, (59):617–645.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Association for Computational Linguistics*, pages 5185–5198.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *arXiv*.

Yuri Bizzoni and Simon Dobnik. Sky + fire = sunset exploring parallels between visually grounded metaphors and image classifiers.

Anna M Borghi, Laura Barca, Ferdinand Binkofski, Cristiano Castelfranchi, Giovanni Pezzulo, and Luca Tummolini. 2019. Words as social tools: Language, sociality and inner grounding in abstract concepts. *Phys. Life Rev.*, 29:120–153.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behav. Res. Methods*, 46(3):904–911.

Angelo Cangelosi and Matthew Schlesinger. 2015. *Developmental robotics: From babies to robots*. MIT press.

Eve V Clark. 2013. *First language acquisition*. Cambridge University Press.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Pasquale A Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, and Stefano F Cappa. 2010. Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 italian words. *Behav. Res. Methods*, 42(4):1042–1048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding.

Felix R Dreyer and Friedemann Pulvermüller. 2018. Abstract semantics in the motor system? – an event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex*, 100:52–70.

Leonardo Fernandino, Jia-Qing Tong, Lisa L Conant, Colin J Humphries, and Jeffrey R Binder. 2022. Decoding the information structure underlying the neural representation of concepts. *Proc. Natl. Acad. Sci. U. S. A.*, 119(6).

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1-3):335–346.

Stevan Harnad. 2017. To cognize is to categorize: Cognition is categorization. In *Handbook of Categorization in Cognitive Science*, pages 21–54.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers.

Aurélie Herbelot. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 321–327, Potsdam, Germany. Association for Computational Linguistics.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and Vision-Language representation learning with noisy text supervision.

Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.

Casey Kennington. 2021. Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 148–157, Online. Association for Computational Linguistics.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 231–236, Beijing, China. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behav. Res. Methods*, 44(4):978–990.

George Lakoff and Mark Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

Richard D Lane and Lynn Nadel. 2002. *Cognitive Neuroscience of Emotion*. Oxford University Press.

Staffan Larsson. 2018. Grounding as a Side-Effect of grounding. *Top. Cogn. Sci.*

John L Locke. 1995. *The Child's Path to Spoken Language*. Harvard University Press.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic visiolinguistic representations for Vision-and-Language tasks.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behav. Res. Methods*, pages 1–21.

Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of DALL-E 2.

Claudia Mazzuca, Luisa Lugli, Mariagrazia Benassi, Roberto Nicoletti, and Anna M Borghi. 2018. Abstract, emotional and concrete concepts and the activation of mouth-hand effectors. *PeerJ*, 6:e5987.

David McNeill and Casey Kennington. 2020. Learning word groundings from humans facilitated by robot emotional displays. In *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, Virtual. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2015. Efficient estimation of word representations in vector space. *In International Conference on Learning Representations (ICLR)*.

Daniele Moro, Gerardo Caracas, David McNeill, and Casey Kennington. 2020. Semantics with feeling: Emotions for abstract embedding, affect for concrete grounding. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtual.

Daniele Moro and Casey Kennington. 2018. Multimodal visual and simulated muscle activations for grounded semantics of hand-related descriptions. In *Proceedings of the 22nd Workshop onthe Semantics and Pragmatics of Dialogue*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. VALSE: A Task-Independent benchmark for vision and language models centered on linguistic phenomena.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2020. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks.

Marta Ponari, Courtenay Frazier Norbury, and Gabriella Vigliocco. 2018. Acquisition of abstract concepts is influenced by emotional valence. *Dev. Sci.*, 21(2).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional image generation with CLIP latents.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv*.

Jacqueline Sachs, Barbara Bard, and Marie L Johnson. 1981. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Appl. Psycholinguist.*, 2(01):33–54.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving references to objects in photographs using the Words-As-Classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223.

John R Searle. 1980. Minds, brains, and programs. *Behav. Brain Sci.*, 3(03):417.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artif. Life*, (11):13–29.

Mariarosaria Taddeo and Luciano Floridi. 2005. Solving the symbol grounding problem: a critical review of fifteen years of research. *J. Exp. Theor. Artif. Intell.*, 17(4):419–445.

Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. 2018. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5520–5527. AAAI.

Caterina Villani, Luisa Lugli, Marco Tullio Liuzza, Roberto Nicoletti, and Anna M Borghi. 2021. Sensorimotor and interoceptive dimensions in concrete and abstract concepts. *J. Mem. Lang.*, 116:104173.

Philippe Vincent-Lamarre, Alexandre Blondin Massé, Marcos Lopes, Mélanie Lord, Odile Marcotte, and Stevan Harnad. 2016. The latent structure of dictionaries. *Top. Cogn. Sci.*, 8(3):625–659.

L Wittgenstein. 2010. Philosophische untersuchungen. In *Sprachwissenschaft*, pages 105–111. De Gruyter.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Comparing Regression Methods for Dialogue System Evaluation on a Richly Annotated Corpus

Kallirroi Georgila
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Los Angeles, CA 90094-2536, USA
kgeorgila@ict.usc.edu

Abstract

We compare various state-of-the-art regression methods for predicting user ratings of their interaction with a dialogue system using a richly annotated corpus. We vary the size of the training data and, in particular for kernel-based methods, we vary the type of kernel used. Furthermore, we experiment with various domain-independent features, including feature combinations that do not rely on complex annotations. We present detailed results in terms of root mean square error, and Pearson's r and Spearman's ρ correlations. Our results show that in many cases Gaussian Process Regression leads to modest but statistically significant gains compared to Support Vector Regression (a strong baseline), and that the type of kernel used matters. The gains are even larger when compared to linear regression. The larger the training data set the higher the gains but for some cases more data may result in over-fitting. Finally, some feature combinations work better than others but overall the best results are obtained when all features are used.

1 Introduction

Dialogue evaluation is an important research topic which over the years has received much attention but still remains an unsolved problem. This is because the quality of a human-machine interaction can be influenced by a large number of factors, such as the genre or domain of dialogue, the design and capabilities of the system and its components, the user expertise and expectations, etc.

In this paper we focus on task-oriented dialogue and our goal is to predict user satisfaction, i.e., user ratings after interacting with the dialogue system. For this purpose we use a richly annotated dialogue corpus with contextual information, and speech act and task labels. This corpus was derived from the original COMMUNICATOR corpus (Walker et al., 2001a) via automatic annotation (Georgila et al., 2005b, 2009). Users of the COMMUNICATOR systems try to book a flight and they may also make hotel or car-rental arrangements. An example dialogue excerpt is shown in Figure 2 in the Appendix.

The original COMMUNICATOR corpus contained speech act and task annotations for the system's side of the conversation based on the DATE scheme (Walker and Passoneau, 2001). Georgila et al. (2005b, 2009) added speech act and task annotations for the user's side of the conversation, as well as information about the dialogue context, e.g., filled slots, filled slots values, grounded slots, speech acts history, etc. The corpus consists of dialogues collected between human users and 8 dialogue systems. We extract domain-independent features from this corpus, and perform regression experiments in order to predict 5 different types of user satisfaction ratings. The corpus and the features we use are discussed in Section 3.

We explore 3 research questions: (i) Which regression method works best and does the choice of kernel matter for kernel-based regression? (ii) What is the impact of varying the training data size? (iii) Which feature combinations work best?

Our contributions are as follows: (1) We compare various state-of-the-art regression methods, in particular, linear regression, linear ridge regression, Support Vector Regression (SVR), and Gaussian Process Regression (GPR). We also vary the kernel type for GPR. To our knowledge, GPR has never been used before for dialogue system evaluation (or generally by the dialogue community) despite the fact that it is considered as the state-of-the-art for regression in other research areas. (2) We vary the size of the training data and report on its impact on performance for all regression methods. (3) We vary the feature combinations used and discuss how the choice of features affects the prediction quality of our models. Our features are domain-independent but are derived from a richly annotated corpus with dialogue context and history, and speech act and task labels. Even though

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

the features we use are domain-independent, our experiments provide valuable insights about the benefits of different feature combinations, including features taking into account dialogue context and dialogue history, as well as features that are not dependent on complex annotations.

Our results show that in many cases GPR leads to modest but statistically significant gains compared to SVR (a strong baseline), and that the type of kernel matters. The gains are even larger when compared to linear regression. The larger the training set the higher the gains but for some cases more data may result in over-fitting. Some feature combinations work better than others but overall the best results are obtained when all features are used.

2 Related Work

Dialogue evaluation is an important area of research, and over the years there have been various surveys recording the state-of-the-art, challenges, and future directions in this research area (Hastie, 2012; Deriu et al., 2021; Mehri et al., 2022).

Prior to the recent advancement of chatbots, most research on dialogue evaluation focused on measuring the quality of human-system dialogue interaction mainly for task-oriented dialogue systems. Dialogue evaluation metrics can be subjective (e.g., user satisfaction, perceived task completion, etc.), or objective (e.g., word error rate, dialogue length, etc.). Interaction logs provide information for calculating objective measures whereas subjective assessments can be collected via surveys and questionnaires (Hone and Graham, 2000).

The most well-known framework for automating the dialogue evaluation process is PARADISE (Walker et al., 2000). PARADISE aims to optimize a desired quality such as user satisfaction by formulating it as a linear combination of various metrics, such as task success and dialogue cost (e.g., dialogue length, speech recognition errors, etc.). The contribution of each factor is determined by weights calculated via linear regression. The advantage of this method is that once a desired quality has been formulated as a realistic evaluation function, it can be optimized by controlling the factors that affect it. Thus, user satisfaction can for example be optimized by increasing task success, and minimizing dialogue length and speech recognition errors. Note however that longer dialogue lengths are not necessarily indicative of poor dialogue quality but depending on the task they may actually indicate user engagement and satisfaction (Foster et al., 2009). Indeed, PARADISE has been shown to be capable of automatically predicting dialogue quality in the travel planning domain (Wright-Hastie et al., 2002). However, it has been argued that PARADISE cannot accurately predict individual user judgements and only covers 40-50% of the variance in the data that it is trained on (Möller and Ward, 2008).

In non-task-oriented dialogue systems (e.g., chatbots) developing robust evaluation metrics can be even harder than for task-oriented dialogue. Here it is not clear what success means and task-specific objective metrics are not appropriate. Instead subjective evaluations for appropriateness of responses can be much more meaningful, which has led to the development of coding schemes for response appropriateness in such cases (Traum et al., 2004; Robinson et al., 2010).

Currently, word-overlap similarity metrics such as BLEU, METEOR, and ROUGE (originally employed in machine translation and summarization) are widely used for measuring chatbot dialogue quality. However, it has been shown that BLEU, METEOR, and ROUGE do not correlate well with human judgements of dialogue quality (Liu et al., 2016). Discriminative BLEU, a variation of BLEU where reference strings are scored for quality by human raters, was found to correlate better with human judgements than standard BLEU (Galley et al., 2015). To address the issues with BLEU, METEOR, and ROUGE, next utterance classification was introduced as a method for evaluating chatbots (Lowe et al., 2016), but the proposed metric recall@k does not take into account the fact that just because a system response is not part of a pre-defined set of appropriate responses it does not mean that it is wrong. Furthermore, topic-based metrics for chatbot evaluation (topic breadth and topic depth) were found to correlate well with human judgements (Guo et al., 2017).

There has also been work on estimating user satisfaction at the system-user exchange level rather than rating the whole dialogue (Engelbrecht et al., 2009; Higashinaka et al., 2010; Ultes and Minker, 2014; Schmitt and Ultes, 2015). Recently, new evaluation metrics have been proposed for open-domain dialogue leveraging pre-trained language models such as BERT and DialoGPT (Ghazarian et al., 2020; Mehri and Eskenazi, 2020a,b).

In this paper, we focus on predicting user sat-

isfaction ratings for the whole dialogue and compare various state-of-the-art regression methods. As mentioned earlier, one of our most important contributions is the introduction of Gaussian Process Regression (GPR) to the dialogue community as a means for dialogue system evaluation. GPR has been used before in the NLP community for machine translation quality estimation (Cohn and Specia, 2013) and emotion prediction (Beck et al., 2014). To our knowledge, in the dialogue community Gaussian Processes (i.e., the GP-SARSA algorithm) have only been used for dialogue policy learning via reinforcement learning (Gašić et al., 2010; Gašić and Young, 2014).

We compare GPR with Support Vector Regression (SVR), which is a strong baseline, and linear regression. Of course linear regression has been used before for dialogue evaluation (e.g., (Walker et al., 2000, 2001b; Georgila et al., 2019, 2020)). Classification based on Support Vector Machines has been used for interaction quality estimation (Ultes and Minker, 2014; Schmitt and Ultes, 2015).

3 Data and Features

The corpus that we use was derived from the original COMMUNICATOR corpus via automatic annotation (Georgila et al., 2005b, 2009). The original COMMUNICATOR corpus contained speech act and task annotations for the system's side of the conversation based on the DATE scheme (Walker and Passoneau, 2001). Georgila et al. (2005b, 2009) added speech act and task annotations for the user's side of the conversation, as well as information about the dialogue context, e.g., filled slots, filled slots values, grounded slots, etc. Georgila et al. (2005b, 2009) present in detail how these fully automatic annotations were generated. Figure 2 shows an example dialogue excerpt including speech act and task annotations, and Figure 3 depicts an example dialogue state corresponding to the dialogue status after user utterance 4 in Figure 2.

The automatic annotations were evaluated with respect to the task completion metrics of the original corpus and in comparison to hand-annotated data, which has verified their validity and reliability (Georgila et al., 2009). Over the years the utility of this annotated corpus has been demonstrated by its use by various researchers for different purposes, mainly, learning dialogue policies (Henderson et al., 2005; Frampton and Lemon, 2006; Henderson et al., 2008) and building simulated users (Schatzmann et al., 2005; Georgila et al., 2005a, 2006). More recently, it was used for system dialogue act selection for pre-training of goal-oriented dialogue policies (McLeod et al., 2019).

The dialogue context annotations are divided into 2 broad categories: logs of the current status of the slots (i.e., 'FilledSlotsStatus', 'FilledSlotsValuesStatus', 'GroundedSlotsStatus'), and logs containing information about how the status of the slots has changed over the course of the dialogue (i.e., 'FilledSlotsHist', 'FilledSlotsValuesHist', 'GroundedSlotsHist'). Because the former inform us about the current status of the slots they may only contain one instance per slot. The latter give us information about the order in which slots have been filled or confirmed and may contain several instances of the same slot, e.g., a slot could be confirmed twice. Thus, if a confirmed slot is refilled with a new value it will remain in the 'ConfirmedSlotsHist' field even though its new value has not been confirmed yet. The history of speech acts and tasks is also included in the annotations.

The annotated corpus (COMMUNICATOR 2001 part) consists of 1683 dialogues collected between human users and 8 dialogue systems but for our experiments we only used dialogues for which all user ratings were available: ATT (157 dialogues), BBN (137 dialogues), CMU (69 dialogues), COLORADO (157 dialogues), IBM (77 dialogues), LUCENT (140 dialogues), MIT (166 dialogues), and SRI (103 dialogues). The first half of the dialogues from each system are included in the training data set (500 dialogues) and the rest are included in the test data set (506 dialogues).

We extract 16 features from this corpus and perform regression experiments in order to predict the following user satisfaction ratings on a Likert scale (1-5, higher is better): ease of the tasks the user had to accomplish (henceforth referred to as 'Task-Ease'), whether it was easy or not to understand the system (henceforth referred to as 'System-Comprehend-Ease'), the user's expertise (henceforth referred to as 'User-Expertise'), whether the system behaved as expected (henceforth referred to as 'System-Behaved-As-Expected'), and if the user would use the system again in the future or not (henceforth referred to as 'System-Future-Use'). We use 16 features divided into 4 categories:

- **duration-related features (9):** overall duration, duration of the system talking part, duration of the user talking part, overall average

duration per utterance, average duration per system utterance, average duration per user utterance, number of overall speech acts, number of system speech acts, number of user speech acts;

- **slots-related features (3):** number of filled slots, number of filled slots without any 'null' values, number of grounded slots (all at the end of the dialogue);
- **slots-history-related features (3):** number of filled slots in the dialogue history, number of filled slots without any 'null' values in the dialogue history, number of grounded slots in the dialogue history (all at the end of the dialogue);
- **word error rate (WER) (1)**: calculated by comparing the speech recognition output to the transcription of the user utterance (this information was included in the original COMMUNICATOR corpus).

We remove all empty ('[]') values, and also distinguish between slots filled with normal versus 'null' values as an extra piece of information (see Figure 3 in the Appendix). Because we only consider numbers of slots, speech acts, and tasks, and not their specific types or values, our features are domain-independent and also automatically extracted from the data. We replace feature values with z-scores, i.e., from each feature value we subtract the mean for that feature and then divide by the standard deviation for that feature. For each feature, the mean and standard deviation are calculated on the training data set.

4 Regression Methods

For our experiments we use various regression methods, specifically, linear regression, linear regression with L2 regularization (also known as linear ridge regression), Support Vector Regression (SVR), and Gaussian Process Regression (GPR). As mentioned above, to our knowledge, GPR has not been used before for dialogue system evaluation, even though GPR is considered as the state-of-the-art for regression and is continually attracting more and more interest.

Gaussian Processes (GPs) are an elegant framework for probabilistic inference incorporating kernels and Bayesian inference (Rasmussen and Williams, 2006). A GP is a probability distribution over possible functions that fit a set of data points. GPs are similar to Support Vector Machines in the sense that they use kernels for non-linear modelling. The main difference is that GPs are probabilistic models and support exact Bayesian inference for regression; approximate inference is required for classification (Rasmussen and Williams, 2006). GPs are also more flexible in terms of fitting the kernel hyperparameters even for complex composite kernels. Because of their probabilistic formulation GPs can also be incorporated into larger graphical models and explicitly model uncertainty.

A kernel is a way of computing the dot product of two vectors in a high dimensional feature space. Thus the kernel function $k(x_i, x_j)$ essentially tells the model how similar two data points (x_i, x_j) are.

For SVR we use scikit-learn[1]. For GPR we use the GPy library[2]. For SVR we experimented with various kernels but using the RBF (radial basis function) kernel resulted in the best performance. For GPR we use the exponential kernel, the rational quadratic kernel, the RBF kernel, the sum of the exponential and the periodic kernel, the sum of the rational quadratic and the periodic kernel, and the sum of the RBF and the periodic kernel.

The RBF kernel is also called the exponentiated quadratic kernel, the squared exponential kernel, or the Gaussian kernel. The rational quadratic kernel is equivalent to adding together multiple RBF kernels with various length scales. For all GPR experiments we varied the length scale and we report results for length scale equal to 1 (the higher the value of the length scale the smoother the learned function). Varying the length scale did not result in significant differences. Note that adding two kernels can be thought of as an OR operation. Thus, the resulting kernel will have a high value if either of the two base kernels has a high value.

All of the above are frequently used kernels for GPR that seem to perform well for various types of data. Training custom kernels may lead to better results but this is a complex process and one of our future work directions. Note that we also experimented with other kernels such as the Matérn 3/2 and 5/2 kernels (Rasmussen and Williams, 2006) as well as the periodic kernel by itself but we do not report these results due to space restrictions. These kernels performed consistently worse.

[1] https://scikit-learn.org/stable/
[2] https://gpy.readthedocs.io/en/deploy/

	linear	linear ridge	SVR RBF	GPR exp	GPR ratq	GPR RBF	GPR exp+per	GPR ratq+per	GPR RBF+per
Task-Ease									
RMSE	1.428	1.376	1.303	1.279	1.281	1.434	1.278	1.281	**1.277**
r	0.349	0.373	0.477	0.493	0.491	0.298	0.494	0.491	**0.498**
ρ	0.425	0.435	0.48	0.506	0.501	0.322	**0.507**	0.501	0.502
System-Comprehend-Ease									
RMSE	1.302	1.242	1.203	**1.161**	1.165	1.246	1.168	1.165	1.178
r	0.161	0.2	0.354	**0.383**	0.378	0.197	0.374	0.378	0.356
ρ	0.242	0.257	0.366	**0.391**	0.383	0.194	0.378	0.383	0.366
User-Expertise									
RMSE	1.405	1.359	1.305	1.297	1.294	1.317	1.297	1.294	**1.283**
r	0.137	0.156	**0.272**	0.252	0.248	0.174	0.253	0.248	0.268
ρ	0.184	0.184	**0.281**	0.266	0.258	0.148	0.267	0.258	0.276
System-Behaved-As-Expected									
RMSE	1.397	1.38	1.295	1.282	1.288	1.419	**1.274**	1.288	1.288
r	0.321	0.333	0.44	0.453	0.447	0.343	**0.462**	0.447	0.445
ρ	0.377	0.382	0.443	0.454	0.451	0.395	**0.465**	0.451	0.451
System-Future-Use									
RMSE	1.492	1.455	**1.397**	1.398	1.398	1.48	1.41	1.398	1.41
r	0.251	0.269	**0.382**	0.376	0.375	0.256	0.342	0.375	0.343
ρ	0.281	0.285	**0.379**	0.362	0.364	0.254	0.333	0.364	0.339

Table 1: Results for RMSE, Pearson's r correlation, and Spearman's ρ correlation, for various regression methods using all the training data and all features; "exp" stands for exponential, "ratq" for rational quadratic, and "per" for periodic kernel. The best values are shown in bold.

5 Results

To measure the predictive power of our models we compare the predictions of each model for each of the 5 user ratings with the ground truth, i.e., the ratings in the test data. We calculate the Root Mean Square Error (RMSE), Pearson's r correlation, and Spearman's ρ correlation.

RMSE measures the average error between the model predictions and the ground truth and its value varies from 0 to 4, given that user ratings were on a scale from 1 to 5. Lower RMSE values are better.

Pearson's r measures the linear relationship between the model predictions and the ground truth and can range from -1 to 1 (the higher the better).

Spearman's ρ is based on the ranked values of the ratings rather than the raw data, which makes sense in our case given that the user ratings can be thought of as some kind of ranking between interactions even though users rated individual interactions. Spearman's ρ determines the degree to which the relationship between the compared variables is monotonic. Spearman's ρ ranges from -1 to 1 (the higher the better).

5.1 Which regression method works best?

Table 1 shows the RMSE, r, and ρ values for the regression methods and kernel types mentioned in Section 4. Here we use all the training data and all features. Clearly SVR and GPR outperform linear and linear ridge regression. For all rating types, GPR results in modest gains compared to SVR, except for 'System-Future-Use'. For 'User-Expertise' SVR results in higher correlation scores than GPR but also higher RMSE. As we will see later, the gains resulting from GPR (compared to SVR) are statistically significant mainly for 'Task-Ease' and 'System-Comprehend-Ease'. For GPR the exponential and rational quadratic kernels outperform the RBF kernel. Adding the periodic kernel to the exponential, rational quadratic, and RBF kernels respectively may lead to improved performance. Adding the exponential and the periodic kernel results in slight gains for 'Task-Ease', 'User-Expertise', and 'System-Behaved-As-Expected'. Adding the rational quadratic and the periodic kernel did not make any difference compared to just using the rational quadratic kernel. Adding the RBF and the periodic kernel led to improved values.

	linear	linear ridge	SVR RBF	GPR exp	GPR ratq	GPR RBF	GPR exp+per	GPR ratq+per	GPR RBF+per
Task-Ease									
20%	1.794	1.446	1.382	1.366	1.479	1.479	1.364	1.364	1.364
40%	1.56	1.415	1.382	1.359	1.349	1.385	1.393	1.39	1.417
60%	1.44	1.397	1.348	1.332	1.331	1.373	1.341	1.347	1.347
80%	1.414	1.369	1.296	1.278	**1.277**	1.31	1.283	1.281	1.305
100%	1.428	1.376	1.303	1.279	1.281	1.434	1.278	1.281	**1.277**
System-Comprehend-Ease									
20%	1.886	1.508	1.249	1.209	1.265	1.265	1.204	1.199	1.199
40%	1.65	1.449	1.222	1.219	1.212	1.312	1.228	1.222	1.231
60%	1.331	1.25	1.188	1.189	1.18	1.241	1.189	1.222	1.222
80%	1.262	1.228	1.172	1.159	**1.154**	1.199	1.161	1.161	1.185
100%	1.302	1.242	1.203	1.161	1.165	1.246	1.168	1.165	1.178
User-Expertise									
20%	1.535	1.499	1.312	1.315	1.329	1.329	1.315	1.309	1.309
40%	1.48	1.395	1.33	1.321	1.319	1.352	1.369	1.399	1.399
60%	1.461	1.418	1.353	1.346	1.34	1.384	1.342	1.366	1.366
80%	1.397	1.361	1.326	1.307	1.299	1.342	1.304	1.31	1.325
100%	1.405	1.359	1.305	1.297	1.294	1.317	1.297	1.294	**1.283**
System-Behaved-As-Expected									
20%	1.777	1.397	1.379	1.431	1.432	1.432	1.431	1.346	1.349
40%	1.506	1.338	1.385	1.333	1.334	1.33	1.422	1.414	1.403
60%	1.404	1.34	1.355	1.305	1.309	1.332	1.313	1.328	1.328
80%	1.383	1.337	1.316	1.287	1.288	1.314	1.29	1.279	1.281
100%	1.397	1.38	1.295	1.282	1.288	1.419	**1.274**	1.288	1.288
System-Future-Use									
20%	1.847	1.643	1.592	1.541	1.541	1.541	1.541	1.5	1.5
40%	1.742	1.524	1.558	1.506	1.47	1.486	1.468	1.494	1.5
60%	1.542	1.489	1.456	1.444	1.443	1.463	1.44	1.444	1.444
80%	1.498	1.461	1.438	1.407	1.41	1.424	1.404	1.411	1.411
100%	1.492	1.455	**1.397**	1.398	1.398	1.48	1.41	1.398	1.41

Table 2: Results for RMSE, for various regression methods using all features, and varying the percentage of training data (20%, 40%, 60%, 80%, 100%); "exp" stands for exponential, "ratq" for rational quadratic, and "per" for periodic kernel. The best values are shown in bold.

5.2 What is the impact of varying the training data size?

Table 2 shows the RMSE values for the regression methods and kernel types mentioned in Section 4. We use all features but vary the percentage of training data (20%, 40%, 60%, 80%, 100%, from each system respectively). Due to space constraints we do not report results on correlation. The values of Pearson's r and Spearman's ρ are consistent with the corresponding RMSE values (the lower the RMSE the higher the correlation).

As expected, for most rating types and methods the larger the size of the training data set the better the performance. However, there are some exceptions when we move from using 80% of the training data to 100% of the training data.

For 'Task-Ease' and for the GPR cases when we add the periodic kernel to the exponential, rational quadratic, and RBF kernels respectively, performance improves or remains stable when we use 100% of the training data but in all other cases it drops. For 'System-Comprehend-Ease', performance improves when we use 100% of the training data only for the GPR case with the sum of the RBF kernel and periodic kernel. For 'User-Expertise' using 100% of the training data outperforms using 80% of the data for all cases except for linear regression. For 'System-Behaved-As-Expected' some-

times adding the last 20% of the data helps but not always. It does not help for linear and linear ridge regression, GPR with the rational quadratic kernel, GPR with the RBF kernel, GPR with the sum of the rational quadratic and periodic kernel, and GPR with the sum of the RBF and periodic kernel. For 'System-Future-Use' using 80% of the training data is better than using 100% of the training data for GPR with the RBF kernel and GPR with the sum of the exponential and periodic kernel. Thus, we can see that in some cases some kind of over-fitting takes place as we add more data.

5.3 Which feature combinations work best?

Table 3 shows the RMSE, r, and ρ values for the regression methods SVR with the RBF kernel and GPR with the exponential kernel. Here we use all the training data but vary the features.

Tables 1 and 2 show that there is not much difference between using GPR with the exponential kernel and GPR with the rational quadratic kernel or their counterparts with the addition of the periodic kernel. For this reason and because of space limitations, for the third research question, we only consider GPR with the exponential kernel and SVR with the RBF kernel. So far we have seen that in many cases GPR outperforms SVR (a strong baseline) but here we also want to see if this is the case for different feature combinations and report on statistical significance.

In terms of feature combinations we get the best results when we use all features except for 'System-Comprehend-Ease' and 'User-Expertise'. As we can see from the first two rows for each rating type, sometimes the duration features are more predictive than the slot features, and vice versa. Combining these features leads to further improvements for all rating types and both SVR and GPR. Adding WER to duration features (dur+WER) always helps except for 'User-Expertise'. Adding slots features to duration features and WER (dur+WER+sl) also always helps. Adding slots history features to WER, slots, and duration features (which is equivalent to using all features) helps in most cases except for 'System-Comprehend-Ease' and 'User-Expertise'. When we remove WER from all features (all-WER) performance improves slightly for 'System-Comprehend-Ease' with SVR, and 'User-Expertise' with both SVR and GPR.

Regarding comparing SVR and GPR, for 'Task-Ease' and 'System-Comprehend-Ease', GPR is almost always significantly better than SVR. For all statistical significance calculations, for comparing SVR and GPR, we use the squared error values and the Wilcoxon signed-rank test with Holm-Bonferroni correction for repeated measures. For 'User-Expertise' and 'System-Behaved-As-Expected', GPR is significantly better than SVR when we use the slots features ($p < 0.01$ and $p < 0.001$ respectively). For 'System-Future-Use', differences between SVR and GPR performance are not significant.

Walker et al. (2001b) also showed the importance of duration and WER for user satisfaction prediction using the original COMMUNICATOR corpus. WER cannot be available unless the user speech is transcribed so an alternative approach would be to use speech recognition confidence scores as a proxy for WER. We also present results assuming that the user's perceived task completion is available (as a high bar for prediction), and as expected, this extra piece of information can significantly improve performance ($p < 0.001$).

We also implemented 5 simple baselines where the model always predicts the same score. Thus, Baseline 1 always predicts the score 1, Baseline 2 always predicts 2, etc. Table 4 shows results for RMSE for the baseline that always predicts the score 3 and the majority baseline for each type of rating, and the best performance of GPR with the exponential kernel (based on Table 3). Figure 1 shows the distributions of values (1 to 5) for each type of rating. The distributions in the training and test data differ, and each type of rating follows different patterns. Based on the distributions for the training data, Baseline 4 is equivalent to the majority baseline for 'Task-Ease', 'System-Comprehend-Ease', 'User-Expertise', and 'System-Behaved-As-Expected', and Baseline 1 is the majority baseline for 'Future-Use'. Baseline 3 generates RMSE values of approximately 1.5 and the only case where the majority baseline works well is for 'System-Comprehend-Ease'. Differences in performance between GPR and all baselines for all rating types are statistically significant ($p < 0.001$).

6 Conclusion

We used regression for predicting user ratings of their interaction with a dialogue system using the richly annotated version of the COMMUNICATOR corpus (Georgila et al., 2005b, 2009). We explored 3 research questions: (i) Which regression method

	SVR-RMSE	SVR-r	SVR-ρ	GPR-RMSE	GPR-r	GPR-ρ	Stat Sign
Task-Ease							
dur	1.37	0.408	0.417	1.319	0.442	0.443	$p < 0.01$
sl	1.377	0.395	0.374	1.357	0.385	0.385	$p < 0.01$
dur+sl	1.334	0.452	0.448	1.292	0.476	0.484	$p < 0.01$
dur+WER	1.327	0.452	0.459	1.292	0.483	0.495	$p < 0.05$
dur+WER+sl	1.316	0.466	0.469	1.281	0.491	0.503	$p < 0.01$
all-WER	1.311	0.473	0.472	1.287	0.484	0.495	$p < 0.05$
all	1.303	0.477	0.48	**1.279**	**0.493**	**0.506**	$p < 0.05$
all+PTC	1.166	0.61	0.605	1.145	0.627	0.636	$p < 0.05$
System-Comprehend-Ease							
dur	1.231	0.269	0.314	1.187	0.339	0.341	$p < 0.01$
sl	1.208	0.349	0.353	1.203	0.343	0.306	n.s.
dur+sl	1.191	0.377	0.382	1.16	0.387	0.387	$p < 0.01$
dur+WER	1.229	0.293	0.318	1.178	0.368	0.376	$p < 0.001$
dur+WER+sl	1.202	0.359	0.364	**1.157**	**0.393**	**0.396**	$p < 0.001$
all-WER	1.191	0.373	0.379	1.162	0.377	0.38	$p < 0.05$
all	1.203	0.354	0.366	1.161	0.383	0.391	$p < 0.001$
all+PTC	1.192	0.386	0.397	1.137	0.434	0.439	$p < 0.001$
User-Expertise							
dur	1.312	0.26	0.29	1.287	0.25	0.262	n.s.
sl	1.317	0.223	0.187	1.305	0.191	0.164	$p < 0.01$
dur+sl	**1.28**	**0.306**	**0.314**	**1.28**	0.275	0.288	n.s.
dur+WER	1.313	0.25	0.27	1.288	0.248	0.263	n.s.
dur+WER+sl	1.287	0.295	0.295	1.283	0.27	0.278	n.s.
all-WER	1.3	0.28	0.3	1.296	0.26	0.274	n.s.
all	1.305	0.272	0.281	1.297	0.252	0.266	n.s.
all+PTC	1.289	0.293	0.315	1.276	0.297	0.325	n.s.
System-Behaved-As-Expected							
dur	1.341	0.398	0.401	1.333	0.392	0.385	n.s.
sl	1.417	0.328	0.307	1.363	0.331	0.322	$p < 0.001$
dur+sl	1.301	0.442	0.436	1.294	0.439	0.432	n.s.
dur+WER	1.309	0.424	0.426	1.301	0.429	0.435	n.s.
dur+WER+sl	1.298	0.439	0.441	1.283	**0.453**	0.453	n.s.
all-WER	1.295	0.447	0.442	1.288	0.446	0.443	n.s.
all	1.295	0.44	0.443	**1.282**	**0.453**	**0.454**	n.s.
all+PTC	1.191	0.568	0.573	1.185	0.572	0.577	n.s.
System-Future-Use							
dur	1.445	0.307	0.298	1.416	0.338	0.323	n.s.
sl	1.446	0.266	0.265	1.446	0.315	0.322	n.s.
dur+sl	1.415	0.357	0.356	1.4	0.364	0.35	n.s.
dur+WER	1.422	0.341	0.333	1.403	0.372	0.364	n.s.
dur+WER+sl	1.41	0.364	0.364	**1.397**	0.374	0.363	n.s.
all-WER	1.405	0.367	0.365	**1.397**	0.37	0.355	n.s.
all	**1.397**	**0.382**	**0.379**	1.398	0.376	0.362	n.s.
all+PTC	1.31	0.489	0.485	1.321	0.49	0.481	n.s.

Table 3: Results for RMSE, Pearson's r correlation, and Spearman's ρ correlation, for SVR with the RBF kernel and GPR with the exponential kernel using all the training data and varying feature combinations; "dur" stands for duration, "sl" for slots, and "PTC" for perceived task completion. The best values are shown in bold. The last column shows statistical significance ("n.s." stands for non-significant).

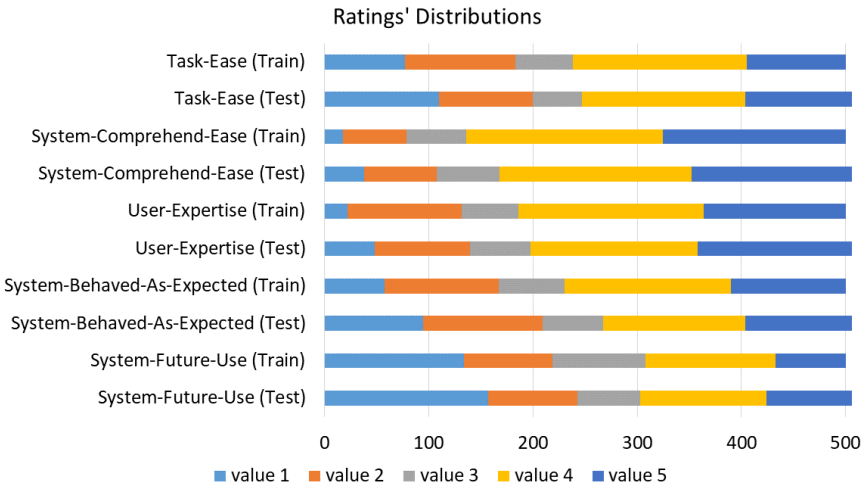


Figure 1: Ratings' distributions for the training and test data.

	Bas 3	Bas maj	GPR
Task-Ease	1.471	1.721	**1.279**
System-Comprehend-Ease	1.421	1.285	**1.157**
User-Expertise	1.431	1.41	**1.28**
System-Behaved-As-Expected	1.433	1.705	**1.282**
System-Future-Use	1.516	2.321	**1.397**

Table 4: Results for RMSE for the baselines and the best performance of GPR with the exponential kernel based on Table 3. The best values are shown in bold.

works best and does the choice of kernel matter for kernel-based regression? (i) What is the impact of varying the training data size? (iii) Which feature combinations work best?

To answer the first question we compared various state-of-the-art regression methods: linear regression, linear ridge regression, SVR, and GPR. We also varied the kernel type for GPR. To our knowledge, GPR has never been used before for dialogue system evaluation (or generally by the dialogue community) despite the fact that it is considered as the state-of-the-art for regression in other research areas. In many cases (mainly for 'Task-Ease' and 'System-Comprehend-Ease'), GPR led to modest but statistically significant gains compared to SVR (a strong baseline), and the type of kernel used mattered. The gains were even larger when compared to linear regression.

To answer the second question we varied the training data size and reported on its impact on performance for all regression methods. The larger the training set the higher the gains but for some methods more data may result in over-fitting.

To answer the third question we varied the feature combinations used for regression and showed how the choice of features affects the prediction quality of our models. Even though the features we used are domain-independent, our experiments provided valuable insights about the benefits of different feature combinations, including features taking into account dialogue context and dialogue history, as well as feature combinations that do not rely on complex annotations. Some feature combinations worked better than others but in most cases the best results were obtained with all features.

Overall the RMSE ranged roughly from 1 to 1.5 depending on the regression method and kernel type, training data size, and feature combination. Predicting individual user judgements is a hard task (Möller and Ward, 2008), and given that we did not use any domain-dependent features our results are promising. For future work we will train custom kernels and measure if performance improves. We also expect performance gains from using domain-dependent features.

Acknowledgements

This work was partly supported by the U.S. Army. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task Gaussian Processes. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1798–1803, Doha, Qatar.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian Processes: An application to machine translation quality estimation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 32–42, Sofia, Bulgaria.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with hidden Markov models. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 170–177, London, UK.

Mary Ellen Foster, Manuel Giuliani, and Alois Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 879–887, Suntec, Singapore.

Matthew Frampton and Oliver Lemon. 2006. Learning more effective dialogue strategies using limited dialogue move features. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 185–192, Sydney, Australia.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. DeltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)– Short Papers*, pages 445–450, Beijing, China.

Milica Gašić, Filip Jurčíček, Simon Keizer, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Gaussian Processes for fast policy optimisation of POMDP-based dialogue managers. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 201–204, Tokyo, Japan.

Milica Gašić and Steve Young. 2014. Gaussian Processes for POMDP-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40.

Kallirroi Georgila, Carla Gordon, Hyungtak Choi, Jill Boberg, Heesik Jeon, and David Traum. 2019. Toward low-cost automated evaluation metrics for Internet of Things dialogues. In *Proc. of the International Workshop on Spoken Dialogue Systems Technology (IWSDS), Lecture Notes in Electrical Engineering 579*, pages 161–175, Singapore.

Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, and David Traum. 2020. Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 726–734, Marseille, France (Online).

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2005a. Learning user simulations for Information State Update dialogue systems. In *Proc. of Interspeech*, pages 893–896, Lisbon, Portugal.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Proc. of Interspeech*, pages 1065–1068, Pittsburgh, Pennsylvania, USA.

Kallirroi Georgila, Oliver Lemon, and James Henderson. 2005b. Automatic annotation of COMMUNICATOR dialogue data for learning dialogue strategies and user simulations. In *Proc. of the Workshop on the Semantics and Pragmatics of Dialogue (SemDial:DIALOR)*, pages 61–68, Nancy, France.

Kallirroi Georgila, Oliver Lemon, James Henderson, and Johanna D. Moore. 2009. Automatic annotation of context and speech acts for dialogue corpora. *Journal of Natural Language Engineering*, 15(3):315–353.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 7789–7796, New York, New York, USA.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2017. Topic-based evaluation for conversational bots. In *Proc. of NIPS Workshop on Conversational AI: Today's Practice and Tomorrow's Potential*, Long Beach, California, USA.

Helen Hastie. 2012. Metrics and evaluation of spoken dialogue systems. In Oliver Lemon and Olivier Pietquin, editors, *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, pages 131–150. Springer.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2005. Hybrid reinforcement/supervised learning for dialogue policies from COMMUNICATOR data. In *Proc. of the IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 68–75, Edinburgh, UK.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed datasets. *Computational Linguistics*, 34(4):487–511.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models. In *Proc. of the International Workshop on Dialogue Systems Technology (IWSDS), Lecture Notes in Computer Science 6392*, pages 48–60, Gotemba, Shizuoka, Japan.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Journal of Natural Language Engineering*, 6(3-4):287–303.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132, Austin, Texas, USA.

Ryan Lowe, Iulian V. Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 264–269, Los Angeles, California, USA.

Sarah McLeod, Ivana Kruijff-Korbayová, and Bernd Kiefer. 2019. Multi-task learning of system dialogue act selection for supervised pretraining of goal-oriented dialogue policies. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 411–417, Stockholm, Sweden.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges. In *arXiv 2203.10012*.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 225–235, Online.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 681–707, Online.

Sebastian Möller and Nigel Ward. 2008. A framework for model-based evaluation of spoken dialog systems. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 182–189, Columbus, Ohio, USA.

Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Susan Robinson, Antonio Roque, and David Traum. 2010. Dialogues in context: An objective user-oriented evaluation approach for virtual human dialogue. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 64–71, Valletta, Malta.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 45–54, Lisbon, Portugal.

Alexander Schmitt and Stefan Ultes. 2015. Interaction Quality: Assessing the quality of ongoing spoken dialog interaction by experts—And how it relates to user satisfaction. *Speech Communication*, 74:12–36.

David R. Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 1699–1702, Lisbon, Portugal.

Stefan Ultes and Wolfgang Minker. 2014. Interaction quality estimation in spoken dialogue systems using hybrid-HMMs. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 208–217, Philadelphia, Pennsylvania, USA.

Marilyn Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, Lynette Hirschman, A. Le, S. Lee, Shrikanth Narayanan, K. Papineni, Bryan Pellom, Joseph Polifroni, Alexandros Potamianos, P. Prabhu, Alexander I. Rudnicky, G. Sanders, Stephanie Seneff, D. Stallard, and Steve Whittaker. 2001a. DARPA Communicator dialog travel planning systems: the June 2000 data collection. In *Proc. of Interspeech/Eurospeech*, pages 1371–1374, Aalborg, Denmark.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with PARADISE. *Journal of Natural Language Engineering*, 6(3-4):363–377.

Marilyn Walker and Rebecca Passoneau. 2001. DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *Proc. of the Human Language Technologies Conference*, pages 1–8, San Diego, California, USA.

Marilyn A. Walker, Rebecca Passoneau, and Julie E. Boland. 2001b. Quantitative and qualitative evaluation of Darpa Communicator spoken dialogue systems. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 515–522, Toulouse, France.

Helen Wright-Hastie, Rashmi Prasad, and Marilyn Walker. 2002. What's the trouble: Automatically identifying problematic dialogues in DARPA Communicator dialogue systems. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384–391, Philadelphia, Pennsylvania, USA.

Appendix

System utterance 1	what trip would you like to take?
System speech act	request_info
System task	top_level_trip
User utterance 1 *ASR output*	i'd like to fly from boston to detroit on november second
User transcription	i'd like to fly from boston to detroit on november second
User speech act	[provide_info,provide_info,provide_info]
User task	[orig_city,dest_city,depart_date]
System utterance 2	on friday, november second
System speech act	implicit_confirm
System task	depart_arrive_date
System utterance 3	flying to detroit
System speech act	implicit_confirm
System task	dest_city
System utterance 4	pardon me
System speech act	apology
System task	meta_slu_reject
System utterance 5	did you say flying from boston instead?
System speech act	explicit_confirm
System task	orig_city
User utterance 2 *ASR output*	yes boston to detroit on northwest airline
User transcription	yes from boston to detroit on northwest airlines
User speech act	[yes_answer,reprovide_info,reprovide_info,provide_info]
User task	[null,orig_city,dest_city,airline]
System utterance 6	flying from boston to detroit
System speech act	implicit_confirm
System task	orig_dest_city
System utterance 7	on northwest
System speech act	implicit_confirm
System task	airline
System utterance 8	instead
System speech act	tbc
System task	tbc
System utterance 9	about what time do you want to leave?
System speech act	request_info
System task	depart_arrive_time
User utterance 3 *ASR output*	early morning july twenty p. m.
User transcription	early morning to arrive around two p. m.
User speech act	[provide_info,provide_info]
User task	[depart_date,depart_time]
System utterance 10	leaving in the early morning
System speech act	implicit_confirm
System task	depart_arrive_time
System utterance 11	pardon me
System speech act	apology
System task	meta_slu_reject
System utterance 12	did you say on friday, july twentieth
System speech act	explicit_confirm
System task	depart_arrive_date
System utterance 13	instead?
System speech act	tbc
System task	tbc
User utterance 4 *ASR output*	no november second
User transcription	no november second
User speech act	[no_answer,provide_info]
User task	[null,depart_date]

Figure 2: An example dialogue excerpt between the system and the user. Both the automatic speech recognition (ASR) output and the transcription of the user's speech are shown. Speech act and task annotations are depicted for both the system and the user. Note that user speech act and task annotations, and dialogue context information are derived from the ASR output.

DIALOGUE LEVEL
Turn: user
TurnStartTime: 991948554.109
TurnEndTime: 991948559.296
TurnNumber: 4
Speaker: user
UtteranceStartTime: 991948554.109
UtteranceEndTime: 991948559.296
UtteranceNumber: 4
DialogueActType: user
ConvDomain: about_task
SpeechAct: [no_answer,provide_info]
AsrInput: no <date_time>november second</date_time>
TransInput: no <date_time>november second</date_time>
Output:
TASK LEVEL
Task: [null,depart_date]
FilledSlot: [null,depart_date]
FilledSlotValue: [no,november second]
GroundedSlot: []
LOW LEVEL
WordErrorRatenoins: 0.00
WordErrorRate: 0.00
SentenceErrorRate: 0.00
KeyWordErrorRate: 0.0
ComputeErrorRatesReturnValue: 0
HISTORY LEVEL
FilledSlotsStatus: [orig_city],[dest_city],[airline],[null],[null],[null],[depart_time],[null],[depart_date]
FilledSlotsValuesStatus: [boston],[detroit],[northwest],[boston],[detroit],[yes],[p m],[no],[november second]
GroundedSlotsStatus: [],[orig_city],[dest_city],[airline],[]
SpeechActsHist: request_info,[provide_info,provide_info,provide_info],implicit_confirm,implicit_confirm,apology,
explicit_confirm,[yes_answer,reprovide_info,reprovide_info,provide_info],implicit_confirm,implicit_confirm,tbc,
request_info,[provide_info,provide_info],implicit_confirm,apology,explicit_confirm,tbc,[no_answer,provide_info]
TasksHist: top_level_trip,[orig_city,dest_city,depart_date],depart_arrive_date,dest_city,meta_slu_reject,
orig_city,[null,orig_city,dest_city,airline],orig_dest_city,airline,tbc,
depart_arrive_time,[depart_date,depart_time],depart_arrive_time,meta_slu_reject,depart_arrive_date,tbc,[null,depart_date]
FilledSlotsHist: [orig_city,dest_city,depart_date],[null,null,null,airline],[depart_date,depart_time],
[null,depart_date]
FilledSlotsValuesHist: [boston,detroit,november second],[yes,boston,detroit,northwest],[july twenty,p m],
[no,november second]
GroundedSlotsHist: [],[orig_city,dest_city,depart_date],[orig_city,dest_city,airline],[]

Figure 3: An example dialogue state generated after user utterance 4 in Figure 2. Note that sometimes empty ('[]') and 'null' values are generated but they do not affect the accuracy of the slot values.

Adjacency Pairs in Common Ground Update: Assertions, Questions, Greetings, Offers, Commands

Manfred Krifka
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
krifka@leibniz-zas.de

Abstract

Dynamic theories of communication focus on the update of the common ground by individual speech acts; for Conversation Analysis, the way that the individual contributions interlock, forming adjacency pairs, are an essential object of study and theorizing. The article proposes a way to enrich dynamic theories by taking into account the possible continuations of speech acts. It focuses on assertions and questions, and extends the treatment to other speech acts.

1 Introduction

Human language communication has been studied from different angles, resulting in quite divergent views that sometimes appear downright incompatible. For instance, on the one hand there are prominent approaches originating in language philosophy, in particular **Speech Act Theory** (Austin 1962, Searle 1969) and the notion of information transfer as **update of Common Ground** (CG) (cf. Stalnaker 1978, 2002). They were successful in describing isolated phenomena, often identified in constructed examples, such as indirect speech acts (Searle 1975), anaphora (Kamp 1981) and projection of presupposition (Heim 1983). On the other hand, there are prominent empirically-driven approaches that pay close attention to actual communicative exchanges, as in **Conversation Analysis** (Sacks et al. 1973, Levinson 2013). They studied phenomena like turn taking that regulate the exchange, the use of backchanneling devices to ensure mutual understanding, and, if that failed, the employment of repair strategies.

A frequent complaint about the first family of approaches is that they put their main focus on the description of single communicative acts, and thus are unable to grasp the dynamics of conversation, where actors plan and shape the direction the conversation should be taking (cf. Levinson 1981, 2017). Approaches of the second type appear far removed from explaining how meaning assignment to complex expressions works and how different aspects of meaning, such as presuppositions, implicatures and alternatives, are woven together. Both approaches exhibit successes, but also have their blind spots. Whether they can be fruitfully combined is an open issue for the authors of Searle et al. (1992). But there are in fact attempts to do so, such as Clark (1996) and Ginzburg (2012), who explicitly combine conversation analysis and CG update.

The current paper presents an **algebraic model of CG update** that is closer to classical speech act theory and accommodates the **sequencing of speech acts** that we observe in communication, thus integrating insights of both research traditions and resulting in a model of communication that takes its interactive nature seriously.

2 Adjacency Pairs

Conversation Analysis offers the notion of **adjacency pairs** as a basic theoretical term to describe the organization of discourse (Schegloff & Sacks 1973). These are conversational moves by one participant, the "first pair part" (FPP), that require corresponding moves of a particular type by the other participant, the "second pair part" (SPP). Examples are greeting-greeting back, question-answer, request-grant (or refusal), proposal-acceptance (or declining). Assertion-confirmation (or rejection), even though not considered adjacency pairs because assertions are said not to require a response, can be seen in similar ways. In case the FPP is not followed by a corresponding SPP, the sequence is felt incomplete, and quite often the initial action will be repeated to achieve success. There are various ways to elaborate on the basic pattern of adjacency pairs by pre-, insert- and post-expansions. Adjacency pairs take on a central role

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

in the textbook by Schegloff (2007), which is evidence for their usefulness for the empirical analysis of conversation.

Early approaches to **sequencing of speech acts** like Kendziorra (1976), Wunderlich (1979) and Ferrara (1980) were not taken up broadly. Searle (1992) considered adjacency pairs to be the most promising aspect of Conversation Analysis to enrich Speech Act theory, but still was skeptical, among other reasons because of the wide variety of appropriate response reactions to a given act.

Speech act theory developed the notion of **felicity conditions** that can be used to specify the **preconditions** that have to be met for a speech act, which often involves the existence of preceding acts. For example, it is a precondition for an answer that a corresponding question was asked. However, preconditions were used in a much wider sense, e.g. for directives, that the addressee is able to carry out the action specified by the directive speech act. For adjacency pairs one would rather need a notion of "postconditions" for speech acts, i.e. how a particular type of speech act is taken up in discourse. By their design, felicity conditions are not suited to capture this forward-looking aspect of speech acts.

Models of **dynamic CG update** did not originally incorporate a notion of interacting conversational moves either, even though such considerations were present in the early work of Hamblin (1971). However, there are more recent approaches that try to represent the dynamics of questions vs. answers, and of assertions vs. (dis)agreements. In particular, the notion of Questions under Discussion provides a tool for modelling this dynamics (cf. Roberts 1996, 2018; Onea 2019). Furthermore, Farkas & Bruce (2010) developed a model that features a negotiation table for updates. Inquisitive Semantics (Ciardelli et al. 2019) provides a CG model for updates with assertions and questions. Also, SDRT (cf. Lascarides & Asher 2009, Hunter et al. 2018) models the intertwining of linguistic discourse and actions, and Murray & Starr (2021) propose a CG model for updates with evidentially modified assertions, commands, and other speech acts.

In this paper I will make use of **Commitment Spaces** (Cohen & Krifka 2014), as this model appears particularly well-suited for dealing with adjacency pairs; its major design feature is the integration of continuations into the notion of CG. Also, it is a rather straightforward extension of the original CG update approach by Stalnaker. Furthermore, it provides an algebraic structure for discourse moves with well-known operations like conjunction, disjunction and denegation.

3 Commitment Spaces

The framework of Commitment Spaces has been developed for pairs of assertions and confirmations or rejections, and for pairs of questions and answers (cf. Krifka 2015, 2022). This article will improve the treatment of assertions and questions, and investigate the potential of the CS framework for modeling adjacency pairs in general.

The CS model starts out with **Commitment States** (**CSts**), which are modeled by non-empty sets of propositions that represent the information about the world and time at which the conversation takes place – more specifically the information that the interlocutors assume to be shared. This contains information about the individual commitments of the participants. If c is such a set of propositions, its conjunction $\bigcap c$ is a set of world-time indices, the "context set" in the sense of Stalnaker (1978). The propositions in a CSts should be consistent (non-contradictory), and also satisfy certain additional **integrity constraints**, some of which we will discuss below.

The notion of **Commitment Spaces** (**CSs**) captures not only information that is shared but in addition the mutual understanding of ways how this shared information can develop in conversation. Hence, a CS is a set of CSts. Disregarding the distinction between informative and performative update (cf. Szabolcsi 1982), **update of a CSt** c with a proposition φ (a function from world-time indices to truth values) restricts c to those indices in which φ is true, cf. (1).

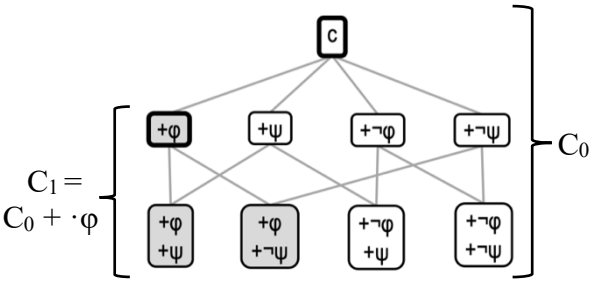
1. $c+\varphi = c \cup \{\varphi\}$, if the integrity constraints for CSts are satisfied, else undefined.

Update of a CS C with a proposition φ restricts C to those CSts c in which φ holds, cf. (2). Here, "·" is an operator that turns a proposition into the corresponding CS update function.

2. $\cdot\varphi(C) = \{c \in C \mid \varphi \in c\}$, also $C + \cdot\varphi$

For example, in (3) a CS consisting of a minimal CSt c and updates by the four propositions φ, ψ. ¬φ and ¬ψ gets updated by φ, resulting in the gray CS.

3. Example: Update of CS C_0 by $\cdot\varphi$



The view of communication as adding information to a CSt is replaced by weeding out those CSts that do not fit to the information that is communicated.

The bold CSt represents the **root** of the CS, the most general CSt that stands for the information accrued so far in the CG; the continuations stand for the ways how the CG can develop. The root of a CS is defined as the set of least informative CSts:

4. $\surd C = \{c \in C \mid \neg\exists c'[c' \in C \wedge c' \subset c]\}$

For example, we have $\surd C_0 = \{c\}$ and $\surd C_1 = \{c+\varphi\}$. Ideally, the root is a singleton, but situations with multiple roots may arise when it is unclear what the shared information actually is. Such multiple roots can be used to model open issues that still have to be resolved, similar to questions under discussion (cf. Kamali & Krifka 2020).

CS updates can be **combined** in various ways. Let A and B be CS updates, then conjunction, disjunction and denegation are defined as follows:

5. $[A \,\&\, B](C) = A(C) \cap B(C)$ conjunction

6. $[A \vee B](C) = A(C) \cup B(C)$ disjunction

7. $[\sim A](C) = C - [A](C)$ denegation

We also have dynamic conjunction (composition) and an operator ? that retains the root of the input CS but restricts the continuations:

8. $[A;B](C) = B(A(C))$ dynamic conjunction

9. $[?A](C) = \surd C \cup A(C)$ restriction

The following examples illustrate these notions with respect to the CS C_0 in (3).

10. $[\cdot\varphi \,\&\, \cdot\psi](C_0) = \{c+\varphi+\psi\} = \{c+\psi+\varphi\}$

11. $[\cdot\varphi \vee \cdot\psi](C_0) = \{c+\varphi,\ c+\psi,\ c+\varphi+\psi,\ c+\varphi+\neg\psi,\ c+\psi+\neg\varphi\}$

12. $[\sim\cdot\varphi](C_0) = \{c,\ c+\psi,\ c+\neg\varphi,\ c+\neg\psi,\ c+\neg\varphi+\psi,\ c+\neg\varphi+\neg\psi\}$

13. $[\cdot\varphi\ ;\ \cdot\psi](C_0) = \{c+\varphi+\psi\}$

14. $[?\cdot\varphi](C_0) = \{c,\ c+\varphi,\ c+\varphi+\psi,\ c+\varphi+\neg\psi\}$

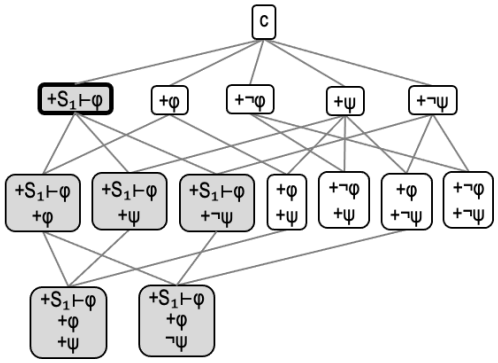
Conjunction (10) and **dynamic conjunction** (13) lead to the same result but achieve this in distinct ways. They differ for anaphoric bindings, as in a dynamic conjunction antecedents in A could bind anaphors in B. **Disjunction** (11) leads to continuations in which either disjuncts are established, which often leads to multiple roots. For example, the root of $[\cdot\varphi \vee \cdot\psi](C_0)$ is $\{c+\varphi,\ c+\psi\}$. **Denegation** (12) removes the possibility that an update occurs, which can be used to model speech acts like *I don't promise to come* (cf. Cohen & Krifka 2014). It typically leaves the root intact, for example the root of $[\sim\cdot\varphi](C_0)$ is $\{c\}$. **Restriction** (14) is like update but retains the CSts in the root, here c.

These are the features of the CS framework in its most basic form. We now set them to work by looking at a model for assertions.

4 Assertions

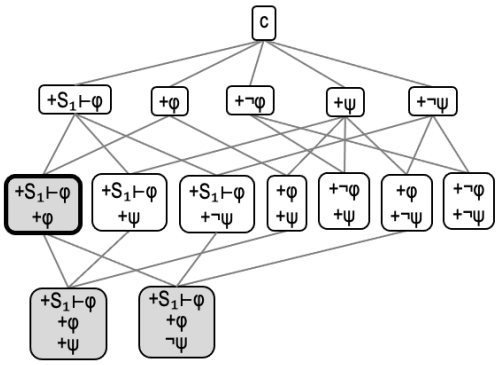
Assertions are not just updates by propositions enforced by a speaker. Rather, the speaker must provide reasons for the addressee to adopt the proposition (cf. Lauer 2013). There is a growing consensus that speakers achieve this by a particular **commitment**, namely by **vouching for the truth of the proposition** (cf. Shapiro 2020; the view can be traced back to Charles S. Peirce, cf. Tuzet 2006; cf. also Brandom 1994). Writing "$S_1 \vdash \varphi$" for the proposition $\lambda i[S_1$ vouches in i that φ is true in i], Krifka (2015) proposes that the characteristic **illocutionary act of assertion** of a proposition φ consists in the speaker S_1 updating the CS by the public commitment of S_1 to the truth of that proposition, i.e., by the proposition $S_1 \vdash \varphi$, with respect to the time of the utterance. This is illustrated in (15):

15. Illocutionary act: $C_2 + \cdot S_1 \vdash \varphi = C_3$



With this backing, the speaker attempts to update the resulting CS by φ itself. This is the intended effect of assertions, their **primary perlocutionary act**: The speaker wants to communicate φ, which is modeled by having it accepted in the CS.

16. Primary perlocutionary act: $C_3 + \cdot\varphi$



5 Accommodating for Reactions

The addressee S_2 has a say in this second move. S_2 can react with *yes* and **confirm** it by also committing to φ, updating with $S_2\vdash\varphi$; or S_2 can say *okay* or **accept** it in other ways, including by not objecting. But S_2 can say *no* and **reject** it by committing to ¬φ, $S_2\vdash\neg\varphi$. It is reasonable to assume an integrity constraint that no CSt c allows for both the propositions φ and $S\vdash\neg\varphi$ be true if S is a participant in conversation. Hence a CS cannot even be updated by $S_2\vdash\neg\varphi$ once φ has been established. The acceptance of φ has to be negotiated – but how should this be modeled?

There are different formal accounts for negotiation in CG update models. For example, Merin (1994) proposes a finite-state automaton representing an "algebra of elementary social acts" that may run in a loop until one of the participants concedes. In their "table" model, Farkas & Bruce (2010) propose that no record of S_1′s initial move is kept if S_2 does not accept it. Krifka (2015) assumes an additional structure, CS developments, allowing for retraction of the most recent move; in case S_2 rejects the attempt of S_1 to assert φ, by saying *no*, the CS will retain the propositions $S_1\vdash\varphi$ and $S_2\vdash\neg\varphi$, hence keep the information that S_1 and S_2 disagree about φ, but not the proposition φ itself.

This article uses the forward-looking feature of CSs, the continuations, to model the effect of rejection without any additional machinery. The overall approach is this: In an assertion, the speaker S_1 first updates the CS with the commitment that the asserted proposition φ is true, rendered as $S_1\vdash\varphi$. This is the illocutionary part. S_1 offers the addressee S_2 not one, but **two continuations**: Either **update with the proposition** φ itself (the intended perlocutionary effect), or a continuation in which S_2 voices **disagreement against update** with φ. I will model the second update by the proposition 'S_2 announces doubts concerning φ', rendered as $S_2\dashv\varphi$, which is incompatible with φ and also with $S_2\vdash\varphi$ by integrity constraints. We assume that the propositions $S_2\dashv\varphi$ and $S_2\dashv\neg\varphi$ can obtain simultaneously in a CSt, they are not ruled out by integrity constraints, different from $S_2\vdash\varphi$ and $S_2\vdash\neg\varphi$. This leads to the following analysis of assertions:

17. Speaker S_1 asserts φ at C_4:
$C_4 + [\cdot S_1\vdash\varphi\ ;\ [\cdot\varphi \vee S_2\dashv\varphi]] = C_5$

This is a dynamic conjunction of an update with the commitment of S_1 to the proposition φ, followed by a disjunction that allows for either the continuation φ or the continuation that S_2 doubts φ. If C_4 is mono-rooted with c_4 as its single CSt, C_5 has a two-element root: $\{c_4 + S_1\vdash\varphi + \varphi,\ c_4 + S_1\vdash\varphi + S_2\dashv\varphi\}$.

Let us consider the possible reactions of S_2 to this disjunction. First, S_2 may **confirm** φ by saying *yes*, updating the CS by $S_2\vdash\varphi$ (where *yes* contains an anaphoric reference to propositions, cf. Krifka 2013). This excludes the disjunct $S_2\dashv\varphi$ due to the integrity constraint mentioned above. The proposition φ is established, and S_2 vouches for it as well:

18. $C_5 + \cdot S_2\vdash\varphi = C_4 + \cdot S_1\vdash\varphi + \cdot\varphi + \cdot S_2\vdash\varphi$

Second, S_2 may just say *okay* and **assent** to φ. This can be interpreted as denegation of $S_2\vdash\neg\varphi$: S_2 indicates non-objection. Under a general rule that objections should be raised as soon as possible (Walker 1996, Faller 2019), even lack of action can be interpreted in this way. Now, the update with $\sim\cdot S_2\vdash\neg\varphi$ is compatible with a CS at which φ is established, but not with a CS at which $S_2\dashv\varphi$ is established. We can assume a plausible integrity constraint for CSs stating that whenever $S_2\dashv\varphi$ is established there must be continuations at which $S_2\vdash\neg\varphi$ gets established – whoever expresses doubt on a proposition might become committed to its negation. Hence update with *okay*, $\sim\cdot S_2\vdash\neg\varphi$, is compatible only with the first disjunct of (17), leading to the establishment of φ:

19. $C_5 + [\sim\cdot S_2\vdash\neg\varphi] = C_4 + \cdot S_1\vdash\varphi + \cdot\varphi$

We did not model the opt-out disjunct in (17) by "weak rejection" of Incurvati & Schlöder (2017),

which amounts to $\neg S_2 \vdash \varphi$, the announcement of non-commitment to φ, as we want to allow for the case of assent, where a proposition φ is in the CG even though not all participants vouch for its truth. The announcement of doubt $S_2 \dashv \varphi$ can be seen as requiring that $S_2 \vdash \neg\varphi$ holds in some continuation.

Third, S_2 may express **dissent** by saying *no*, updating the CS by $S \vdash \neg\varphi$. As this update is not compatible with $\cdot\varphi$ due to an integrity constraint, now the first disjunct of (17) is excluded, resulting in (20). This is a coherent CS in which it is established that S_1 and S_2 do not agree on φ:

20. $C_5 + \cdot S_2 \vdash \varphi = C_4 + \cdot S_1 \vdash \varphi + \cdot S_2 \vdash \neg\varphi = C_6$

What these three reactions have in common is that they **reduce the root** of the CS that was increased by the disjunction in in (17). Multiple roots stand for issues that are still undecided; reducing them not only increases the overall information in a CS but also removes that uncertainty in its root (cf. Kamali & Krifka 2020).

Consent and dissent need not be performed with speech acts involving the very proposition φ or ¬φ. Other assertions that have a bearing on φ or ¬φ, like S_1: *It is raining.* S_2: *I think so too* / *I don't think so*, can be seen as confirming or expressing doubt or dissent as well. This can be dealt with by integrity constraints that rule out, e.g, that both φ and 'x believes ¬φ' ($B_x\neg\varphi$) are part of a CSt, if x is a participant of conversation. For example, update by *I don't think so* results in (21). Here, S_2 commits to $S_2 \vdash B_{S_2}\neg\varphi$ (assuming neg raising), attempting to put $B_{S_2}\neg\varphi$ into the CS (the second disjunct that S_1 doubts this proposition is rather hypothetical as S_1 is not an epistemic authority over S_2's beliefs).

21. $C_5 + [\cdot S_2 \vdash B_{S_2}\neg\varphi \,;\, [\cdot B_{S_2}\neg\varphi \vee \cdot S_1 \dashv B_{S_2}\neg\varphi]]$
$= C_4 + \cdot S_1 \vdash \varphi + \cdot S_2 \dashv \varphi + S_2 \vdash B_{S_2}\neg\varphi$

The update is only compatible with the second disjunct in (17), denegating the commitment of S_2 to φ. In addition, the proposition that S_2 commits to not believing ¬φ is introduced, as well as the proposition that S_2 does not believe ¬φ.

Other reactions to assertions of a proposition φ can express doubts by asserting a proposition ψ that make φ less probable, such as S_1: *It will rain.* S_2: *But the report said it will be fine.* Such assertions of ψ are compatible with both φ, the proposition that S_1 intends to introduce, and $S_2 \dashv \varphi$, that S_2 expresses doubts about φ. Hence they do not decide the issue but leave it open to additional arguments.

In summary, the representation of assertions developed here incorporates **adjacency pairs** into a model of CS change by offering certain continuations after the illocutionary update $S_1 \vdash \varphi$: either φ gets established (by confirming or by assenting, i.e. refraining from dissenting), or $S_2 \dashv \varphi$ gets established (by dissenting). The FPP (17) allows for SPPs like *yes*, *okay* or *no*, but also for other moves that favor one continuation over the other.

6 Retracting Commitments

If conversation leads to a CS that contains both $S_1 \vdash \varphi$ and $S_2 \vdash \neg\varphi$, then neither φ nor ¬φ can be established in the future development of the CS. Either speaker can repeat his or her claims, but this will not move the conversation forwards (cf. Merin 1994). In real life, there are ways out of such quandaries: We can agree to disagree and live with the contradictory claim and turn to other tasks or topics, or one speaker can give up his or her claim. How can this be modeled? We need an account for what happens when speakers **retract** their commitments.

As CSts are modeled as sets of propositions, we can capture such operations as removing a proposition from the CSts of a CS:

22. $C + {}^{-}\varphi = \{c - \{\varphi\} \mid c \in C\}$ retraction

Retraction is a peculiar move. The updates we considered so far restrict the CS they apply to; for such updates A we have $A(C) \subseteq C$. In contrast, retraction is **non-monotonic**: Updating C_1 in (3) by ¬φ results in $\{c+\psi, c+\neg\psi\}$, which is not a subset of C_1. Furthermore, the CS may contain propositions that entail the retracted proposition, which then also would have to be removed.

There is also a move of **addition** of a proposition φ to a CS C that was previously ruled out:

23. $C + {}^{+}\varphi = \{c \cup \{\varphi\} \mid c \in C\}$ addition

The resulting CSts must satisfy the integrity constraints. Such operations require modeling as belief revisions (Gärdenfors 2003), where retraction corresponds to **contraction**, and there is an operation of **revision** $[C + {}^{-}\neg\varphi] + {}^{+}\varphi$ for consistent addition.

Participants are not entitled to remove just any proposition from a CS. But it should be admissible that speakers remove their own commitments or doubts; e.g. S_1 can remove $S_1 \vdash \varphi$ or $S_1 \dashv \varphi$. Even this comes with social costs, as normally people are supposed to stick to their commitments. However,

removing one's commitments should incur higher costs than removing one's doubts.

The communicative impasse in our example can be dissolved by either S_1 giving up $S_1 \vdash \varphi$, as illustrated in (24) for the CS of (20), or alternatively by S_2 giving up $S_2 \vdash \neg\varphi$.

24. $C_6 + {}^{-}S_1 \vdash \varphi = C_4 + \cdot S_2 \vdash \neg\varphi$

S_1 can express this retraction by *okay (you may be right)*. This opens up a way for S_2 to assert φ and introduce φ, in the hope that S_1 will not object the second time around. In (19) we have analyzed *okay* as refraining from committing to the negation of the proposition, $\sim \cdot S_2 \vdash \neg\varphi$; in the present situation, this move presupposes the retraction in (24) and enforce it by accommodation. S_1 may even confirm φ, by asserting it: $[\cdot S_1 \vdash \neg\varphi \ ; \ \cdot\varphi]$, which also presupposes prior retraction of $S_1 \vdash \varphi$.

7 Compositional Interpretation

How do we get from an assertive sentence, like *It is raining*, to its interpretation? Recent proposals assume operators that turn the representation of the proposition into an update with the commitment for this proposition. Krifka (2015), cf. also Miyagawa (2022), has proposed an Act Phrase ActP with head "$\cdot$" and a Commitment Phrase ComP with head "$\vdash$" that takes a Tense Phrase TP as argument which denotes a proposition, resulting in the following interpretation (S_1, S_2 are speaker and addressee, respectively).

25. $[\![[_{\text{ActP}} \cdot [_{\text{ComP}} \vdash [_{\text{TP}}$ *it is raining*$]]]]\!]^{S_1,S_2}$
$= [\![\cdot]\!]^{S_1,S_2}([\![\vdash]\!]^{S_1,S_2}([\![[_{\text{TP}}$ *it is raining*$]]\!]^{S_1,S_2}))$
$= [\![\cdot]\!]^{S_1,S_2}(\lambda x[x \vdash$ 'it is raining'$])$
$= \lambda C[C + \cdot S_1 \vdash$ 'it is raining'$]$

The application of $[\![\vdash]\!]$ to a proposition results in a function from a person x to the proposition that x is committed to the proposition; the application of $[\![\cdot]\!]^{S_1,S_2}$ specifies x as the speaker, S_1, and turns the resulting proposition into a CS update.

However, (25) captures only the illocutionary act of assertion, not the perlocutionary act that puts the proposition into the CS, nor the disjunct that allows for rejection. In fact, it is not even possible to design a compositional interpretation that includes that perlocutionary effect, given the syntactic structure in (25), as the TP proposition is not accessible to $[\![\cdot]\!]$. One option is to assume that the TP introduces a propositional discourse referent, which is independently motivated by the interpretation of response particles like *yes* and *no* that take up such discourse referents (cf. Krifka 2013). This discourse referent is projected to the level of the ActP head "$\cdot$", which can take it together with the TP and create the appropriate meaning. In the representation (26), the discourse referent of a proposition is realized as the first member of a pair with the TP meaning.

26. $[\![\cdot]\!]^{S_1,S_2}([\![\vdash]\!]^{S_1,S_2}(\langle\varphi, \varphi\rangle))$
$= [\![\cdot]\!]^{S_1,S_2}(\langle\varphi, \lambda x[x \vdash \varphi]\rangle)$
$= \lambda C[C + [\cdot S_1 \vdash \varphi \ ; \ [\cdot\varphi \vee \cdot S_2 \dashv \varphi]]$

In (26) the intended perlocutionnary effect $\cdot\varphi$ and its alternative $\cdot S_2 \dashv \varphi$ are built into the interpretation of "$\cdot$". We may doubt that this effect is indeed part of the grammatical meaning: There are assertions that do not intend to inform, but only to commit (e.g. in a confession of religious faith). Alternatively, the continuation $[\cdot\varphi \vee \cdot S_2 \dashv \varphi]$ can be seen as a consequence of a pragmatic rule that is triggered by the introduction of a commitment to a proposition φ, with S_2 as the addressee. Then (25) represents the grammatical meaning of assertions.

8 Polar Questions

Leaving the topic of assertions we turn to questions. In a question, the speaker does not change the factual information present in the CS but indicates that the CS should take a certain development – in the most typical case, that the addressee asserts a proposition that answers the question. Hence questions have been modeled as sets of propositions in one way or other (Hamblin 1973, Groenendijk & Stokhof 1984, von Stechow 1990, Ciardelli et al. 2019). In the commitment space framework, questions are updates that leave the root intact but restrict the continuations (Krifka 2015). This allows to represent **question bias** in a straightforward way.

A simple polar question like *Is the door open?* is typically represented as a set $\{\varphi, \neg\varphi\}$, cf. Hamblin (1973). However, such questions can express a bias towards one proposition. The question *Is the door closed?* differs in this respect from *Is the door open?* (cf. Büring & Gunlogson 2000, Trinh 2014). The commitment space framework offers a way to express this bias by having such questions project only one proposition. Krifka (2015, 2022) implements this in a way that such questions create only one continuation with a commitment by the

addressee to the proposition. Here I assume a refined model that incorporates reactions against the bias of the question as an alternative:

27. $[\![[_{\text{ActP}}\ ?\ \textit{is}\ [_{\text{ComP}} \vdash [_{\text{TP}}\ \textit{it}\ \_\ \textit{raining}]]]]\!]^{S_1,S_2}$
$= [\![?]\!]^{S_1,S_2}([\![\vdash]\!]^{S_1,S_2}([\![[_{\text{TP}}\ \textit{it is raining}]]\!]^{S_1,S_2}))$
$= [\![?]\!]^{S_1,S_2}(\lambda x[x \vdash \varphi])$
$= ?[\cdot[\lambda x[x \vdash \varphi](S_2)] \vee \cdot[\lambda x[x \dashv \varphi](S_2)]]$
$= \lambda C[\sqrt{C} \cup [\cdot S_2 \vdash \varphi \vee \cdot S_2 \dashv \varphi](C)]]$
$= \lambda C[\sqrt{C} \cup C + \cdot S_2 \vdash \varphi \cup C + \cdot S_2 \dashv \varphi]$

Questions have an ActP head ? to which finite copulas and auxiliaries move in standard polar questions in English. This head is interpreted by the restriction operator ?, cf. (14), that is applied to the CS update with the proposition that the **addressee**, here S_2, is committed to the TP proposition, $S_2 \vdash \varphi$, disjoined with the announcement of doubt, $S_2 \dashv \varphi$). The first continuation is the commitment by S_2 to the proposition φ; this represents the bias of the question. The other continuation consists in an update that the speaker doubts φ; this allows for responses like *no* or *I don't know*. As with assertions, the second part may be a pragmatic effect: When speaker S_1 checks if addressee S_2 would commit to φ, S_1 expects that S_2 expresses doubts about φ if S_2 does not want to commit to φ.

Let us consider the effect of different answers. Take C_7 as a CS that becomes updated by the question (27):

28. $(27)(C_7)$
$= [\sqrt{C_7} \cup [\cdot S_2 \vdash \varphi \vee \cdot S_2 \dashv \varphi](C_7)]$
$= [\sqrt{C_7} \cup C_7 + \cdot S_2 \vdash \varphi \cup C_7 + \cdot S_2 \dashv \varphi]]$
$= C_8$

In a **confirming** response, S_2 asserts φ to S_1. As with assertions, with *yes* S_2 picks up the TP proposition, commits to it, and proposes to accept it. The result is an update of the commitment space C_8 with the commitment of S_2 to φ, eliminating the second disjunct in (28), followed by an update with φ. This may be disjoined with an update with $S_1 \dashv \varphi$, but as S_1 gave epistemic authority to S_2 this latter update is hypothetical.

29. $C_8 + [\cdot S_2 \vdash \varphi\ ;\ [\cdot \varphi\ (\vee \cdot S_1 \dashv \varphi])]$
$= C_7 + \cdot S_2 \vdash \varphi + [\cdot \varphi\ (\vee \cdot S_1 \dashv \varphi)]$

In a **dissenting** response, S_2 reacts with *no*, asserting the negated proposition $\neg\varphi$. Now the first disjunct of (28) gets eliminated, resulting in a commitment by S_2 to $\neg\varphi$ and two possible continuations, acceptance of $\neg\varphi$ or assertion of $\neg\neg\varphi$, $= \varphi$.

30. $C_8 + [\cdot[S_2 \vdash \neg\varphi]\ ;\ [\cdot\neg\varphi\ (\vee \cdot S_1 \dashv \neg\varphi)]]$
$= C_7 + \cdot S_2 \dashv \varphi + \cdot S_2 \vdash \neg\varphi + [\cdot\neg\varphi\ (\vee \cdot S_1 \vdash \varphi)]$

Different from Krifka (2015), answers that go against the bias of a question do not require a retraction. There is still a difference to answers that go along with the bias, as they can be achieved by the reaction *yes* that does not require a negation. In case the question is based on a negated proposition, as in *Is it not raining?*, the answer *no* has an assenting reading as it may pick up the non-negated antecedent proposition, cf. Krifka (2013).

Responses like *I don't know* that express **inability to answer** can be dealt with as well as they are not compatible with $S_2 \vdash \varphi$. but with $S_2 \dashv \varphi$. Representing this proposition 'S_2 knows φ' as $K_{S_2}\varphi$, (which entails $B_{S_2}\varphi$) when uttered by S_2, we have to invoke the integrity constraint that rules out $S_2 \vdash \varphi$ and $\neg K_{S_2}\varphi$. This is illustrated in (31). We treat the second disjunct $S_1 \dashv \neg K_{S_2}\varphi$ as irrelvant, as S_1 has no epistemic authority over S_2's knowledge.

31. $C_8 + \cdot[S_2 \vdash \neg K_{S_2}\varphi]\ ;\ [\cdot\neg K_{S_2}\varphi\ (\vee \cdot S_1 \dashv \neg K_{S_2}\varphi)]$
$= C_7 + \cdot S_2 \dashv \varphi + \cdot S_2 \vdash \neg K_{S_2}\varphi + \cdot\neg K_{S_2}\varphi$

In case S_2 reacts with the assertion of an **irrelevant** proposition, such as *It's Monday.,* the effect is that the question still stays active, as both disjuncts of (28) can be updated with it. More specifically, such updates result in root multiplication:

32. $C_8 + [\cdot S_2 \vdash \psi\ ;\ [\cdot\psi \vee \cdot S_1 \vdash \neg\psi]]$
$= C_7 + \cdot S_2 \vdash \varphi + \cdot S_2 \vdash \psi + [\cdot\psi \vee \cdot S_1 \vdash \neg\psi]$
$\cup\ C_7 + \cdot\neg S_2 \vdash \varphi + \cdot S_2 \vdash \psi + [\cdot\psi \vee \cdot S_1 \vdash \neg\psi]$

9 Other Questions

We have dealt with simple polar questions, called **monopolar** by Krifka (2015), as they put one proposition in the foreground. **Alternative questions** such as *Is it raining or not?* and *Is it raining or snowing?* are disjunctions of such questions:

33. $[\![[[_{\text{ActP}}\ ?\ \textit{is}\ [_{\text{ComP}} \vdash [_{\text{TP}}\ \textit{it}\ \_\ \textit{raining}]]]\ \textit{or}$
$[_{\text{ActP}}\ ?\ \textit{is}\ [_{\text{ComP}} \vdash [_{\text{TP}}\ \textit{it}\ \_\ \textit{not raining}]]]]]\!]^{S_1,S_2}$
$= [?\cdot S_2 \vdash \varphi \vee ?\cdot S_2 \vdash \neg\varphi]$
$= \lambda C[\sqrt{C} \cup [\cdot S_2 \vdash \varphi \vee \cdot S_2 \dashv \varphi \vee$
$\cdot S_2 \vdash \neg\varphi \vee \cdot S_2 \dashv \neg\varphi](C)]$

The difference to the monopolar question (27) is that the update $\cdot S_2 \vdash \neg\varphi$ is mentioned explicitly, and also introduces a propositional discourse referent. Hence this question is **non-biased**, with the answers *Yes, it is* and *No, it isn't* equally prominent.

Biezma (2009) observes that alternative questions based on a proposition and its negation come with a **cornering effect**: The addressee is forced to give a non-evasive answer. This can be derived from (33) under a preference for strongest disjunctive alternatives. Observe that $\cdot S_2 \vdash \varphi$ is stronger than $\cdot S_2 \dashv \neg\varphi$, in the sense that whenever a CS is updated with the former, the latter update does not add new information, due to the integrity constraint of commitment consistency that rules out $x \vdash \varphi$ and $x \dashv \varphi$. In the same way, $\cdot S_2 \vdash \neg\varphi$ is stronger than $S_2 \dashv \varphi$. This preference strengthens (33) to $\lambda C[\surd C \cup [\cdot S_2 \vdash \varphi \vee \cdot S_2 \vdash \neg\varphi]]$, which does not leave S_2 an option to evade the question.

Constituent questions like *When did it rain?* can be analyzed as generalized disjunction over the alternatives provided by the *wh*-constituent:

34. $[\![[_{\text{ActP}}$ *when* ? *did* $[\vdash [_{\text{TP}}$ *it* _ *rain* _ $]]]]\!]^{S_1,S_2}$
$= V_{t \in \text{TIME}}\ [?\cdot S_2 \vdash \varphi[t] \vee ?\cdot S_2 \dashv \varphi[t]]$

Possible answers specify one or more of the disjuncts, e.g. *It rained at noon*, or *It rained at noon and in the evening*, or *It rained at noon or in the evening.* Also, answers like *It did not rain at noon* (which implies $\neg S_2 \vdash \varphi$[noon]) can be handled. Answers to constituent questions typically are understood as exhaustive, which can be modeled by focus-induced alternatives in the answer, such as *It rained at* $[NOON]_F$ (cf. Kamali & Krifka 2020 for a proposal within the CS model).

Modeling assertions as in (17) or (26) with the help of a disjunction of the intended enrichment of the CS with the proposition φ and a commitment to its negation looks similar to an **assertion with question tag**, as in *It is raining, isn't it?* However, such cases can be transparently interpreted as a disjunction of an assertion with a question (cf. Krifka 2015, 2022). This disjunction can be expressed overtly, as e.g. in *It is raining, or not?*

35. $[\![[[_{\text{ActP}} \cdot\ [_{\text{ComP}} \vdash [_{\text{TP}}$ *It is raining*$]]]$
$[_{\text{ActP}}$? *is* $[_{\text{ComP}} \vdash [_{\text{NegP}}$ *n't* $[$*it* _ ~~*raining*~~$]]]]]]\!]^{S_1,S_2}$
$= \lambda C[\cdot S_1 \vdash \varphi\ ;\ [\cdot\varphi \vee \cdot S_2 \dashv \varphi]](C) \vee$
$[\surd C \cup [\cdot S_2 \vdash \neg\varphi \vee \cdot S_2 \dashv \neg\varphi]](C)]$

In this move, the speaker S_1 vouches for the truth of φ, trying to introduce φ, or alternatively, the addressee vouches for the truth of ¬φ. As the second part is a question, the root does not change in this overall move. In case S_2 confirms with *yes*, both S_1 and S_2 vouch for φ, and φ gets established. In case S_2 rejects with *no*, then S_1 is not committed to φ due to the second disjunct in (35). This differs from the plain assertion, *It is raining*, where the speaker commitment to the proposition remains even if the other speaker rejects this move with *no*. In a sense, question tags like the one in (35) have the effect that the speaker is committed to the proposition only under the condition that the addressee does not disagree.

10 Greetings

Having discussed assertions and questions, we turn to the classical adjacency pair of greetings. What is a greeting, as a speech act? In general, it is an acknowledgement of the presence of another person or group of persons, making them participants of the conversation. Particular greetings often incorporate the time of the day, express emotional involvement, and confirm the social relation between speaker and addressee as being familiar, distant, symmetric, or asymmetric. Greetings may be pure recognitions, such as *Hi!*, they may be derived from wishes as in *Good morning!*, or be based on questions about the current state of the other person such as *How are you?* (cf. Jucker 2017). There are non-linguistic greetings such as waving, eyebrow raises and whistles, and greetings are similar to callings (vocatives).

For the current purpose it is sufficient to assume a proposition λi[x greets y in i], in short G(x,y), which holds if x recognizes y. Adding this proposition to the CS presupposes that x is a participant, and makes y a participant as well. Example:

36. $[\![Hi!]\!]^{S_1,S_2} = \cdot G(S_1,S_2)$

This does not involve any commitment operator ⊢ as the speaker does not commit to the truth of the proposition $G(S_1,S_2)$ but simply creates it in the CS. This is similar to explicit performative speech acts like *I hereby open the buffet* or *The buffet is hereby open*, which also do not communicate about the world with the help of truth commitments but create new facts in the world (cf. Searle 1976, Szabolcsi 1982)

Greetings expect a counter-greeting, which ensures that the greeting was recognized. This expectation can be modeled by the restriction operator ?:

37. $C_9 + [\![Hi!]\!]^{S_1,S_2} = C_9 + \cdot G(S_1,S_2)\ ;\ ?\cdot G(S_2,S_1)$
$= C_{10}$

Here, the input CS is first modified by the greeting of S_2 by S_1, and then the greeting of S_1 by S_2 is

established as the preferred continuation. If S_2 greets back, the conversation goes on smoothly:

38. $C_{10} + [\![Hi!]\!]^{S_2,S_1} = C_9 + \cdot G(S_1,S_2)\ ;\ \cdot G(S_2,S_1)$

But what happens if S_2 does not recognize S_1? Then the effect of S_1's greeting obviously does not obtain. This can be modeled by assuming a disjunction between the effect of the countergreeting, and the **removal** of the effect of the first greeting:

39. $[\![Hi!]\!]^{S_1,S_2} \cdot [G(S_1,S_2)\ ;\ [?\cdot G(S_2,S_1) \vee -G(S_1,S_2)]]$

Again, if S_2 greets back, the conversation goes on as intended. If S_2 fails to do so, the effect of the first greeting is removed, that is, it is not part of the CG that S_1 recognized S_2. In this situation, S_1 can greet S_2 again in a second attempt to enrich the CG by mutual recognition.

In the case of assertions, the opt-out move was not specified as a removal of the commitment of the first speaker, $S_1 \vdash \varphi$. The reason for this is that the commitment of the speaker remains even if the speaker's move is not taken up.

11 Offers and Commands

The final interactional pair we consider are offers (commissives), in which the speaker promises to do something, such as *I promise to do the dishes*, and commands (directives), in which the speaker obliges the addressee to do something, such as *Do the dishes!* They differ from assertions about future actions or deontic propositions (*I will do / you must do the dishes*), insofar the speaker does not commit to a proposition that is independently true of the utterance itself.

However, these future clauses can also be used as performatives (optionally marked by *hereby*). This provides a novel way of modeling offers and commands as performative speech acts that add propositions about future actions. This is different from the analysis of imperatives as performative deontics in Kaufmann (2012) but related to the analysis by Barker (2011) as imposing future actions. The addressee has an option to decline the offer or to reject the command, which again can be expressed by a disjunction. Let WD(x) be the proposition 'x will do the dishes':

40. $[\![I\ promise\ to\ do\ the\ dishes]\!]^{S_1,S_2}$
$= \cdot WD(S_1)\ ;\ [?\cdot S_2 \vdash WD(S_1) \vee -WD(S_1)]$

41. $[\![Do\ the\ dishes!]\!]^{S_1,S_2}$
$= \cdot WD(S_2)\ ;\ [?\cdot S_2 \vdash WD(S_2) \vee -WD(S_2)]$

In (40) the speaker S_1 introduces the proposition that S_1 will do the dishes but this depends on confirmation by S_2, here rendered as an assertion; otherwise the proposition is removed. The situation is similar in (41) except that now S_1 places an obligation on the addressee S_2 that can be confirmed or dismissed by S_2. For example, if S_2 reacts with *No*, asserting $S_2 \vdash \neg WD(S_2)$, this is only compatible with the second disjunct in (41). Both speech acts could be expressed by performatively interpreted future propositions, but there are idiomatized forms for commissives and grammaticalized forms for directives (cf. Gärtner 2020).

12 Conclusion

This paper developed an algebraic model that allows for the modeling of adjacency pairs in a framework of common ground update. It made use of the commitment space (CS) model that incorporates a forward-looking dimension in CG updates. The essential idea is that the possible reactions to a particular update are represented in these possible continuations. It is crucial that the commitment states that make up a CS satisfy pragmatic integrity constraints that restrict the possible moves.

There are a number of issues that this approach raises, some of which mentioned by the reviewers. One concerns the psychological plausibility, given modelling by infinite sets. Appendix 2 argues that a representational variant is possible that works with an interpreted language. Another is the fact that conversation often requires collaboration and the recognition of long-term intentions beyond mere adjacency pairs (Clark 1996). The CS model with its focus on continuation is actually a promising framework for such wider-reaching conversational plans. Another is the fact that conversations often interleave with real actions; this necessitates a notion of CSts and CSs that includes aspects of shared attention beyond language (cf. Clark 1997, Hunter et al. 2018). Finally there is the conception of CSs as a representation of the CG that is supposed to be shared. Participants may have different ideas about what the CG is, which may necessitate private versions of the CG such as the dialogue gameboards of Ginzburg (2012), but see Gregoromichelaki et al. (2020) in defense of a common space of interactions.

Acknowledgments

I am grateful to Anton Benz, Friderike Buch, Hans-Martin Gärtner, Marvin Schmitt and Tue Trinh, as well as the discussants of presentations of this material in Berlin and in Austin. I am particularly grateful for the helpful and detailed comments by three anonymous reviewers. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme, ERC Advanced Grant 787929 SPAGAD: Speech Acts in Grammar and Discourse.

Appendices

Integrity constraints

The theoretical approach presented here relies on integrity constraints for Commitment States (CSts) In particular, update c+φ results in c ∪ {φ} only if the integrity constraints are satisfied. These constraints represent rational communicative behavior that participants expect from each other in conversation. The constraints used in the text are listed here as combinations of propositions that are ruled out for well-behaved CSts, where x stands for a participant in conversation, P for sets of propositions, ⇒ for logical consequence, ⊢ for public commitment to the truth of a proposition and ⊣ for announcement of doubt to a proposition.

1. * φ ∈ c, ∃P⊆c[P ⇒ ¬φ] logical consistency
2. * x⊢φ, x⊢¬φ ∈ c claim consistency
3. * x⊢φ, ¬φ ∈ c claim/proposition consistency
4. * x⊢φ, x⊣φ ∈ c claim/doubt consistency
5. * x⊣φ, φ ∈ c doubt/proposition consistency
6. * B_x¬φ, φ ∈ c belief/proposition consistency
7. * B_x¬φ, x⊢φ belief/claim consistency

The following two integrity constraint do not restrict commitment states but commitment spaces:

8. All commitment states in a commitment space satisfy the integrity constraints for commitment states.
9. If there is a c ∈ C, with x⊣φ ∈ c, then there is a c′∈c with c⊆c′ such that x⊢¬φ ∈ c′.

The latter states that if x commits do doubt about φ then x does not rule out to commit to ¬φ.

Representation of Commitment States / Spaces

The framework to conversation presented here follows Stalnaker's approach to Common Ground updates insofar as CGs were captured by propositions (sets of propositions for CSts, sets of sets of propositions for CSs). In this it is similar to frameworks such as Farkas & Bruce (2010) and Ciardelli et al. (2019). But relying on propositions as sets of world-time indices, and on sets (of sets) of such sets, may be psychologically and implementationally implausible (cf. Ginzburg 2012). But representational versions of the framework presented here can be developed that achieve a compact formulation of commitment spaces:

As for CSts, instead of being modelled by sets of **propositions** φ they can be represented by sets of **formulas** φ in an interpreted language that state the truth conditions of these propositions, $[\![\varphi]\!] = \varphi$.

As for CSs, instead of being modelled by sets of sets of propositions that represent possibly infinite continuations, a CSs C can be represented by the CSts in its root √C, potentially extended by one continuation level in the case of questions. We can derive C as the union of all expansions E(R) of a possibly extended root set R of CSts that satisfy the integrity constraints, if we add certain formulas.

10. $\cdot\varphi(R) = \{c \cup \{\varphi\} \mid c \in R\}$ if integrity constraints are satisfied
11. $[?\varphi](R) = R \cup \cdot\varphi(R)$ restriction
12. $[A\ ;\ B](R) = B(A(R))$ dynamic conjunction
13. $[A \vee B](R) = A(R) \cup B(R)$ disjunction
14. $[\sim\varphi](R) = \{c \cup \{\sim\varphi\} \mid c \in R\}$ denegation

Denegation instructs expansion E not to include φ. This is mediated by an integrity constraint:

15. * $\sim\varphi$, $\varphi \in c$

In this blocking of φ, $\sim\varphi$ has a similar effect as negation $\neg\varphi$, but notice that $\sim\varphi$ is not interpreted: If $\sim\varphi \in c$ then c leaves it open whether φ holds or not; if $\neg\varphi \in c$ then c rules out that φ holds. Hence, retraction of $\sim\varphi$, as required by addition of φ, does not change the truth conditions of a CSt, and is a monotonic operation on this level.

The formulas $x\vdash\varphi$ and $x\dashv\varphi$ also have a blocking effect, on ¬φ. In this case, we can assume that the retraction of $x\dashv\varphi$ occurs no social costs to x, in contrast to the retraction of $x\dashv\varphi$.

References

Austin, J. L. 1962. *How to do things with words.* Oxford: Clarendon Press.

Barker, C. 2012. Imperatives denote actions. *Sinn und Bedeutung* 16. Cambrdige, Mass: MITWPL., 57-70.

Biezma, M. 2009. Alternative vs. polar questions: the cornering effect. *SALT* 19. LSA Open Journal Systems, 37-54.

Brandom, Robert B. 1994. *Making it explicit. Reasoning, representing, and discourse commitment.* Cambridge, Mass.: Harvard University Press.

Büring, D. & C. Gunlogson. 2000. Aren't positive and negative polar questions the same? *LSA Annual meeting.*

Ciardelli, I., J. Groenendijk & F. Roelofsen. 2019. *Inquisitive Semantics.* Oxford University Press.

Clark, H. H. 1996. *Using language.* Cambridge University Press.

Cohen, A. & M. Krifka. 2014. Superlative quantifiers and meta-speech acts. *Linguistics and Philosophy* 37: 41-90.

Faller, M.. 2019. The discourse commitment of illocutionary reportatives. *Semantics & Pragmatics* 12, 1-46.

Farkas, D. F. & K. B. Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics* 27: 81-118.

Ferrara, A. 1980. Appropriateness conditions for entire sequences of speech acts. *Journal of Pragmatics* 4: 321-340.

Gärdenfors, P. 2003. Belief revision: An introduction. In: Gärdenfors, P, (ed), *Belief revision*. Cambridge University Press, 1-28.

Gärtner, H.-M. 2020. On the utility of the promissive signal and the "promissive gap". *Chicago Linguistic Society* 56. 123-135.

Ginzburg, J. 2009. *The interactive stance. Meaning for Conversation.* Oxford University Press.

Gregoromichelaki, E. et al. 2020. Affordance competition in dialogue: the case of syntactic universals. *SemDIAL (WatchDIAL) 2020.*

Groenendijk, J. & M. Stokhof. 1984. *Studies on the semantics of questions and the pragmatics of answers.* Doctoral Dissertation. University of Amsterdam.

Hamblin, C. L. 1971. Mathematical models of dialogue. *Theoria* 37: 130-155.

Hamblin, C. L. 1973. Questions in Montague English. *Foundations of Language* 10: 41-53.

Hunter, J., N. Asher & A. Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics* 11.

Incurvati, L. & J. J. Schlöder. 2019. Weak assertion. *The Philosophical Quarterly* 69: 741-770.

Jucker, A. H. 2017. Speech Acts and speech act sequences: Greetings and Farewells in the History of American English. *Studia Neophilologica* 89: 39-58.

Kamali, B. & M. Krifka. 2020. Focus and contrastive topic in questions and answers, with particular reference to Turkish. *Theoretical Linguistics* 46: 1-71.

Kamp, H. 1981. A theory of truth and semantic representation. In: Groenendijk, J.A.G. et al. (eds), *Formal Methods in the Study of Language*. Amsterdam: Mathematical Centre Tracts 135, 277-322.

Kaufmann, Magdalena. 2012. *Interpreting imperatives.* Heideberg: Springer.

Kendziorra, E. 1976. Sequenzierung von Sprechakten. In: Weber, H. & H. Weydt, (eds), *Sprachtheorie und Pragmatik*. Tübingen: Max Niemeyer Verlag, 357-366.

Krifka, M. 2013. Response particles as propositional anaphors. *SALT* 23. LSA Open Journal Systems, 1-18.

Krifka, M. 2015. Bias in Commitment Space Semantics: Declarative questions, negated questions, and question tags. *SALT* 25. LSA Open Journal Systems, 328-345.

Krifka, M. 2021. Modeling questions in commitment spaces. In: Cordes, M. (ed), *Asking and answering*. Tübingen: Narr, 63-95.

Lascarides, A. & N. Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of Semantics* 26: 109-158.

Lauer, S. 2013. *Towards a dynamic pragmatics.* Doctoral dissertation. Stanford University.

Levinson, S. C. 1981. The essential inadequacies of speech act models of dialogue. In: Parret, H. et al., (eds), *Possibilities and limitations of pragmatics*. John Benjamins, 473-492.

Levinson, S. C. 2013. Action formation and ascription. In: Sidnell, J. & T. Stivers, (eds), *The Handbook of Conversation Analysis*. London: Blackwell,

Levinson, S. C. 2017. Speech acts. In: Huang, J. (ed), *The Oxford Handbooks Online*. Oxford.

Merin, A. 1994. Algebra of elementary social acts. In: Tsohatzidis, S. L., (ed), *Foundations of speech act*

theory. Philosophical and linguistic perspectives. London: Routledge, 234-266.

Murray, S. E. & W. B. Starr. 2020. The structure of communicative acts. *Linguistics and Philosophy* 1-50.

Miyagawa, S. 2022. *Syntax in the treetops.* Cambridge, Mass.: MIT Press.

Onea, E. 2016. *Potential questions at the semantics-pragmatics interface.* Leiden: Brill.

Roberts, C. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. In: Yoon, J. H. & A. Kathol, (eds), *OSU Working Papers in Linguistics 49: Papers in Semantics*. Columbus: The Ohio State University, 91-136.

Roberts, C. 2018. Speech acts in discourse context. In: Fogal, Daniel, Daniel W. Harris & Matt Moss, (eds), *New Work on Speech Acts*. Oxford University Press, 317-359.

Sacks, H., Schegloff, E. & G. Jefferson. 1972. A simplest systematics for the organization of turn-taking in conversation. *Language* 50: 596-735.

Schegloff, E. & H. Sacks. 1973. Opening up closings. *Semiotica* 8: 289-327.

Schegloff, E. A. 2007. *Sequence Organization in Interaction.* Cambridge: Cambridge University Press.

Searle, J. R. 1969. *Speech acts. An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Searle, J. R. 1976. A classification of illocutionary acts. *Language in Society* 5: 1-23.

Searle, J R. 1992. Conversation. In: Searle, J. R., H. Parret & J. Verschueren, (eds), *(On) Searle on conversation*. John Benjamins, 7-30.

Searle, J. R., H. Parret & J. Verschueren, (eds). 1992. *(On) Searle on conversation*. John Benjamins.

Shapiro, L. 2020. Commitment accounts of assertion. In: Goldberg, S, (ed), Oxford Handbook of Assertion. Oxford: Oxford University Press, 73-97.

Stalnaker, R. 1978. Assertion. In: Cole, P. (ed), *Pragmatics*. New York: Academic Press, 315-323.

Stalnaker, R. 2002. Common ground. *Linguistics and Philosophy* 25: 701-721.

Szabolcsi, A. 1982. Model theoretic semantics of performatives. In: Kiefer, F. (ed), *Hungarian linguistics*. Amsterdam: John Benjamins, 515-535.

Trinh, T. 2014. How to ask the obvious: A pre-suppositional account of evidential bias in English yes/no questions. In: Crnič, L. & U. Sauerland, (eds), *The Art and Craft of Semantics: A Festschrift for Irene Heim*. 227-249.

Tuzet, G. 2006. Responsible for Truth? Peirce on judgement and assertion. *Cognitio* 7: 317-336.

von Stechow, A. 1990. Focusing and backgrounding operators. In: Abraham, Werner, (ed), *Discourse particles*. Amsterdam: John Benjamins, 37-84.

Walker, M. 1996. Inferring acceptance and rejection in dialog by default rules of inference. *Language and Speech* 39: 265-304.

Wunderlich, D. 1979. Was ist das für ein Sprechakt? In: Grewendorf, G. (ed), *Sprechakttheorie und Semantik*. 275-324..

Rational Speech Act models are utterance-independent updates of world priors

Jean-Philippe Bernardy **Julian Grove** **Christine Howes**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
firstname.lastname@gu.se

Abstract

A popular framework for modelling pragmatic effects is the "rational speech act" (RSA) model introduced by Frank and Goodman (2012). The idea behind RSA is that, to interpret an utterance, a rational (pragmatic) listener reasons about a speaker who chooses their utterance by reasoning about the listener, using a literal semantic model. In the present work, we take the RSA model at face value, but we reformulate it in information-theoretic terms. We find that the pragmatic listener model can be reconceived as an update of the prior over worlds that can be provided independently of the speaker's actual utterance. This update consists in a preference for world states which are the most specific to a given utterance in the set of possible utterances given by the pragmatic context. Our reformulation allows us to deduce general properties of pragmatic reasoning problems. As an example, we show that RSA does not predict certain quantity implicatures in the presence of bell-curve priors.

1 Introduction

The "rational speech act" (RSA) model introduced by Frank and Goodman (2012) recasts a broadly Gricean view of language in Bayesian probabilistic terms. As in the work of Grice (1975), the core ideas underlying RSA are that dialogue participants are rational agents who communicate efficiently by reasoning over each other's beliefs and the shared communicative goal. The core assumption in the RSA model is "... that listeners view speakers as having chosen their words informatively — that is, relative to the information that they would transfer to a naive listener" (Frank and Goodman, 2014, p.84).

The basic RSA model claims that a rational (pragmatic) speaker will take into account how a naive (literal) listener interprets an utterance, assuming it is true. The ideal (pragmatic) listener reasons, in turn, about the pragmatic speaker, thus also taking into account the nested reasoning over the literal listener.

This model is meant to account for human decision making. Much of the support for RSA comes from restrictive communication games in which participants must pick a speaker's intended referent from a set of objects which may match or differ on particular attributes (such as shape or colour) given only a one word utterance. In certain circumstances, a pragmatic listener must take into account both the shared features across objects that are consistent with a given utterance, as well as those features which are not shared, in order to disambiguate among referents for which the utterance is ambiguous. According to Frank and Goodman (2012), the predictions of the RSA model correlate strongly with human behaviour in such one shot referential games. The RSA framework has since been applied to a variety of linguistic puzzles of ambiguity and optionality, including whether plural predications will receive a distributive or collective reading (Scontras and Goodman, 2017), and whether null versus overt pronouns are chosen in constructions which may feature pro-drop (Chen et al., 2018), among others.

In the present work, we take the RSA model at face value, but we reformulate it in explicitly information-theoretic terms by calling on the notion of information gain between the prior and posterior distributions. Our reformulation provides the following insights:

- While a common, algorithmic interpretation of the RSA model suggests that agents reason over each other's reasoning states (listener-speaker-listener), this formulation is not only implausible, but unnecessary, as we show. That is, one can reason in much more direct terms.

- RSA does not, in fact, make correct predictions about the implicatures expected in par-

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

ticular conversational contexts, according to the Gricean underpinnings of RSA, and given reasonable (bell-curve) priors.

In particular, we show that any given RSA model, whether of a pragmatic listener or a pragmatic speaker, may be presented merely as a *filter* on what we will call a "pragmatic prior"; that is, a prior over worlds or utterances which has been reconceptualised in information-theoretic terms, in order to incorporate notions of specificity and informativeness. Given such a pragmatic prior distribution, any given occasion of interpreting an utterance (or choosing an utterance, given some intended message) requires only that the listener/speaker renormalise this distribution with any incompatible values removed.

There is some precedent for our proposal in work by Scontras et al., which provides an information-theoretic reformulation of the pragmatic speaker model. More precisely, the authors characterise the model in terms of the following formula (which, as we will explain in the next section, incorporates a parameter α setting the model's temperature):

$$\begin{aligned} P_{S_1}(u|w) \;\propto\; & \mathrm{Truth}(u,w) \\ & * \mathrm{Informativeness}(u)^{\alpha} \\ & * \mathrm{Economy}(u)^{\alpha} \end{aligned}$$

Here, the term $\mathrm{Truth}(u,w)$ is a filter on the distribution determined by the other two terms; that is, it is valued as 1 if utterance u is true at world w, and as 0 if it is false. In the present work, we take the next logical step by showing that the pragmatic listener model can be subject to the same kind of reformulation. Consequently, both the pragmatic speaker model and the pragmatic listener model can be reduced to mere filters on their respective pragmatic priors.

2 Background: RSA

RSA, as proposed by Frank and Goodman (2012), assumes a set of possible utterances $\mathcal{U}$ and a set of world states $\mathcal{W}$. World states w come with a prior probability $P(w)$, and utterances u come with a cost $C(u)$. Additionally, we have a relation l on $\mathcal{U}$ and $\mathcal{W}$ such that $l(u,w) = 1$ if utterance u is true at world state w, and $l(u,w) = 0$ otherwise. We say that the tuple $(\mathcal{U}, \mathcal{W}, P, C, l)$ constitutes a *pragmatic interpretation problem*. A solution to such a problem consists in a specification of a pragmatic listener, which is a function from $\mathcal{U}$ to distributions over $\mathcal{W}$. Given an utterance $u \in \mathcal{U}$, it is assumed that the posterior distribution of the pragmatic listener is computed on the assumption that u is literally true. Thus we model the pragmatic listener as taking for granted that its interlocutor is adhering to the Maxim of Quality.

In its most common formulation, RSA models a pragmatic listener as an agent which reasons about a speaker, which, in turn, reasons about a literal listener. To illustrate how this works, let us consider a situation in which there is a box which, at one point, contained 7 cookies, and it is known that John ate at least 5 of them. Thus $\mathcal{U} = \{\text{'John ate } x \text{ cookies'} \mid x \in [5,7]\}$. (We let the cost function C be constant across utterances.) The set of possible world states corresponds to those where some number w of cookies has actually been eaten. We choose the literal semantics to allow for more cookies to have actually been eaten than stated:

$$l(\text{'John ate } x \text{ cookies'}, w) = w \geq x$$

Thus considering only relevant values of w, the literal meaning l can be represented by the following table.

w	5	6	7
'John ate 5 cookies'	1	1	1
'John ate 6 cookies'	0	1	1
'John ate 7 cookies'	0	0	1

2.1 The literal listener model

The literal interpretation of u is given by a Bayesian update of P by l, which thus acts as a filter on P:

$$P_{L_0}(w \mid u) \propto l(u,w) \times P(w) \qquad (1)$$

In our example, we consider the prior P to be uniform, and thus, the family of distributions $P_{L_0}(w \mid u)$ is obtained by normalising each row of the above table:

w	5	6	7
'John ate 5 cookies'	1/3	1/3	1/3
'John ate 6 cookies'	0	1/2	1/2
'John ate 7 cookies'	0	0	1

2.2 The speaker model

According to RSA, the speaker **S** is modelled as an agent which produces a distribution over utterances for each world state w that **S** might wish to convey:

$$P_{S_1}(u \mid w) \propto \exp[\alpha \times (\log(P_{L_0}(w \mid u)) - C(u))]$$

or, equivalently:

$$P_{S_1}(u \mid w) \propto \frac{P_{L_0}(w \mid u)^{\alpha}}{e^{\alpha C(u)}} \qquad (2)$$

Each parameter $C(u)$ represents the cost of uttering u. The role of the parameter α (where it is assumed that $\alpha > 1$), is to exacerbate the differences of literal fit among utterances. For α tending to infinity, **S** chooses the utterance with the highest utility $U(u, w) = \log(P_{L_0}(w \mid u) - C(u))$ with a probability of 1 (i.e., stochastic certainty).

In our example, if we let $\alpha = 4$, then we obtain the family of distributions $P_{S_1}(u \mid w)$ by, first, exponentiating by 4, and, second, normalising each *column*. Dividing by the exponentiated cost has no effect on the resulting distribution because it is a constant that vanishes after normalising.

w	5	6	7
'John ate 5 cookies'	1	0.16	0.01
'John ate 6 cookies'	0	0.84	0.06
'John ate 7 cookies'	0	0	0.93

2.3 The pragmatic listener model

The pragmatic listener model P_{L_1} refers to the above speaker model when updating the distribution over world states:

$$P_{L_1}(w \mid u) \propto P_{S_1}(u \mid w) \times P(w) \qquad (3)$$

Since P is uniform in our example, we obtain the family of distributions $P_{L_1}(w \mid u)$ by once more normalising each row.

w	5	6	7
'John ate 5 cookies'	0.85	0.14	0.01
'John ate 6 cookies'	0	0.93	0.07
'John ate 7 cookies'	0	0	1

This example illustrates some noteworthy points. First, *the core aspect of computing RSA models is the application of normalisation steps*. While the normalisation factors of the nested speaker and listener models are therefore crucial, they are left implicit by the usual formulaic presentation of RSA (Eqs. (1) to (3)). We will see below that making these factors explicit brings insight.

Second, the formulation of RSA in terms of a listener who reasons about a speaker who reasons about a literal listener makes it difficult to build an intuition of what the model predicts. For instance, in our example, RSA predicts, to a large extent, that 'John ate x cookies' implicates 'John ate exactly x cookies'. But does it predict a similar implicature for variations of the same example, for instance using another prior? One might intuitively expect it to do so, but, as we demonstrate below, it does not always make this prediction.

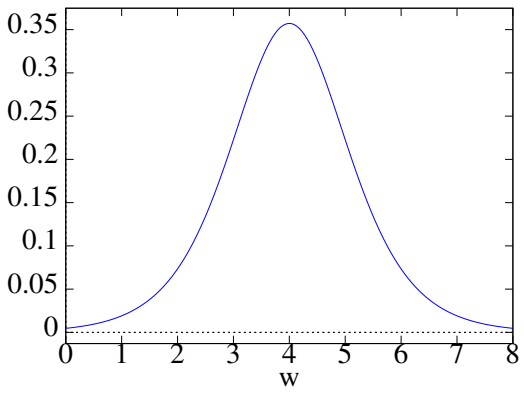
Third, a pragmatic listener is conceived of as reasoning about all possible combinations of world states and utterances simultaneously, which is large for any non-trivial example. Indeed, it is psychologically implausible that such a process is at play in the listener's mind.

The purpose of the next section is to reformulate RSA in terms that are easier to grasp, while addressing these weaknesses.

3 Information-theoretic reformulation

We carry out our reformulation in terms of information theoretic concepts; in particular, information gain. To illustrate our points, we use a variation of the example from the previous section, in which the alternative utterances differed along some numerical value which provided a lower bound on compatible world states. The main differences in our current example will be the following:

- The relevant numerical variable is now continuous. Thus alternative utterances now have the form 'John ran x kilometres'.
- The prior distribution over world states (i.e., over the number of kilometres John ran) is no longer uniform. We instead use a logistic distribution, defined below. (A logistic distribution is similar to a normal distribution, but it simplifies our calculations.[1] Our qualitative conclusions will hold just as well in the case of a prior which is normally distributed.)



[1] The logistic distribution is leptokurtic. That is, it has fatter tails than the normal distribution, i.e., more outliers.

The above plot represents a prior for the number of kilometres John ran given by a logistic distribution with a mean of 4. Thus to obtain $P_{L_0}(w \mid u)$ for any u of the form 'John ran x kilometres', one should crop this distribution at x (on the left, so only the right part remains) and renormalise.

3.1 Literal information gain

The instrumental concept underlying our information-theoretic reformulation is the Kullback–Leibler (K-L) divergence, which measures the information gained when updating a prior belief taking a distribution P to a posterior belief taking distribution Q. It is defined as follows:[2]

$$D_{\mathrm{KL}}(Q \parallel P) = -\sum_{x \in \mathcal{X}} Q(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

In terms of this definition, we may compute the *literal* information gain provided by an utterance u as the K-L divergence between the prior on worlds P, and the posterior $Q_u(w) = P_{L_0}(w \mid u)$ computed by L_0 (which takes u literally):

$$Q_u(w) \propto l(u, w) \times P(w)$$
$$G_{L_0}(u) = D_{\mathrm{KL}}(Q_u \parallel P)$$

Because $l(u, w)$ takes 0 or 1 values, the following reformulation of G_{L_0} is possible, by Theorem 1 (given in Appendix A):

$$G_{L_0}(u) = -\log \sum_{w \in \mathcal{W}} l(u, w) \times P(w) \quad (4)$$

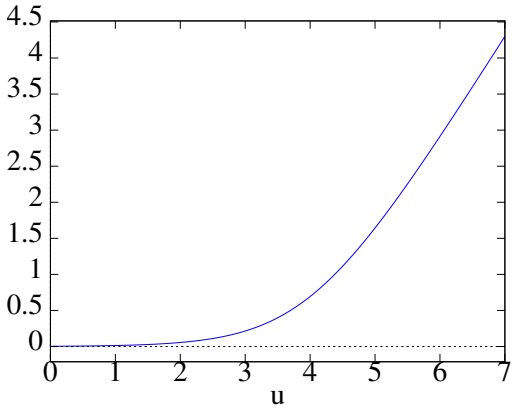
For an alternative, more compact presentation of G_{L_0}, one may first define the following prior over utterances associated with the literal listener:

$$P_{L_0}(u) = \sum_{w \in \mathcal{W}} l(u, w) \times P(w) \quad (5)$$

That is, P_{L_0} is the probability associated with u by L_0, given the prior over world states. G_{L_0} may then, instead, be rendered as follows:

$$G_{L_0}(u) = -\log P_{L_0}(u) \quad (6)$$

In our running example, which uses a logistic prior, we then have the following information gain for the literal listener (G_{L_0}) for utterances of the form 'John ran x kilometres':



The flat regime toward the left of the plot is explained by the fact that utterances of 'John ran x kilometres', where x is lower than the mean of the prior, do not provide much information: they are compatible with most world states. The steadily increasing gain after the mean is explained by the converse: these utterances are *incompatible* with most world states. Furthermore, in this part of the plot, a given increase in the value of x leads to a roughly constant increase in information gain. Thus the information gain here increases in a roughly linear relationship with utterance strength. Such an increase, in turn, rules out a roughly constant *proportion* of the remaining possible world states, given the log scale associated with information gain.

3.2 The reformulated speaker model

With the above notion of information gain in mind, we can make the normalisation factor in P_{L_0} of Eq. (1) explicit, thus turning the proportionality relation into an equality:

$$\begin{aligned} P_{L_0}(w \mid u) &= \frac{l(u, w) \times P(w)}{\sum_{w_1 \in \mathcal{W}} l(u, w_1) \times P(w_1)} \\ &= \frac{l(u, w) \times P(w)}{P_{L_0}(u)} \\ &= l(u, w) \times P(w) \times e^{G_{L_0}(u)} \end{aligned}$$

One can now substitute $P_{L_0}(w \mid u)$ by $l(u, w) \times P(w) \times e^{G_{L_0}(u)}$ in the definition of P_{S_1} (Eq. (2)), and simplify the result. Note that making the normalisation factor explicit was necessary to carry out such a substitution, which is only valid for strictly

[2] The notation '$P(w)$' normally suggests that the described distribution is discrete, in which case weighted averages are computed with a sum. We will use these notations throughout, since they are easier to present than density functions and integrals, a choice we make even though our running example uses a continuous variable.

equal (not just proportional) terms.

$$
\begin{aligned}
P_{S_1}(u \mid w) &\propto \frac{P_{L_0}(w \mid u)^{\alpha}}{e^{\alpha \times C(u)}} \\
&\propto \frac{\left(l(u,w) \times P(w) \times e^{G_{L_0}(u)}\right)^{\alpha}}{e^{\alpha \times C(u)}} \\
&\propto l(u,w)^{\alpha} \times P(w)^{\alpha} \times e^{\alpha \times (G_{L_0}(u) - C(u))} \\
&\propto l(u,w) \times e^{\alpha \times (G_{L_0}(u) - C(u))} \qquad (7)
\end{aligned}
$$

The last step of rewriting (7) is justified because (i) the exponent α of $l(u,w)$ has no effect, since $l(u,w) \in \{0,1\}$, and (ii) the term $P(w)$ does not depend on u, and thus does not affect the proportionality relation.

At this point, we may introduce our reformulation of the speaker model as a filter on a *pragmatic prior*. We do so by first defining this prior; in particular, we may view the second factor in (7) as proportional to a speaker's *pragmatic prior over utterances*:[3]

$$P_{S_1}(u) \propto e^{\alpha \times (G_{L_0}(u) - C(u))} \qquad (8)$$

Note that this reformulation shows that the speaker *a priori* favours utterances whose information gains are larger than their costs, a preference which is exacerbated by high values of α.[4] We may now formulate $P_{S_1}(u \mid w)$ as a *filter* on the above prior, provided by $l(u,w)$:

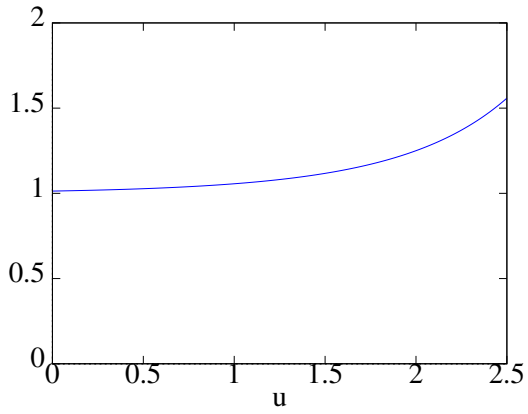
$$P_{S_1}(u \mid w) \propto l(u,w) \times P_{S_1}(u) \qquad (9)$$

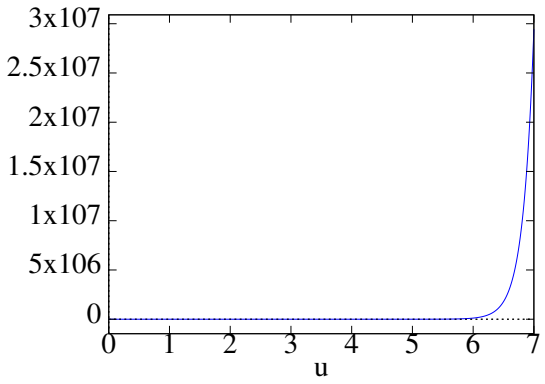
As an algorithmic model of the speaker's reasoning, the above proportionality relation can be seen as implying that a "table" of utterances and their relative degrees of preferredness (according to informativeness and cost) has been constructed *a priori*. Upon choosing a world state w to communicate, the speaker may then filter out those utterances incompatible with w, in order to then select an utterance among those that remain. In other words, the utility of an utterance ($U(u) = G_{L_0}(u) - C(u)$) is *independent of the world state w that the speaker wishes to communicate*.

[3] Given an infinite set of possible alternative utterances (and, indeed, in our running example), $P_{S_1}(u)$ need not define a probability distribution. This is not a problem in practice, as the speaker and listener posteriors, $P_{S_1}(u \mid w)$ and $P_{L_1}(w \mid u)$, will nevertheless be proper probability distributions.

[4] One can additionally choose cost to be proportional to utterance length, following Lassiter and Goodman (2013). Given such a definition of cost, the speaker will prefer utterances with a high (literal) information density.

In our running example, $P_{S_1}(u \mid w)$ is thus obtained by cropping $P_{S_1}(u)$ on the right (and then renormalising). The following plots exemplify $P_{S_1}(u \mid w)$ for two different values of w, prior to normalisation: 2.5, and 7 (where $\alpha = 4$ in each case).





As can be seen, $P_{S_1}(u|w)$ is nearly constant for low values of u, but it shoots up exponentially once u exceeds the mean of the prior over world states. In other words, if choosing an utterance whose strength exceeds the mean of the prior is at all possible, then the speaker will most definitely do so. If only utterances whose strength is below the mean are possible, then the speaker will still be biased towards stronger utterances, but not to the same degree.

3.3 Normalisation via information gain

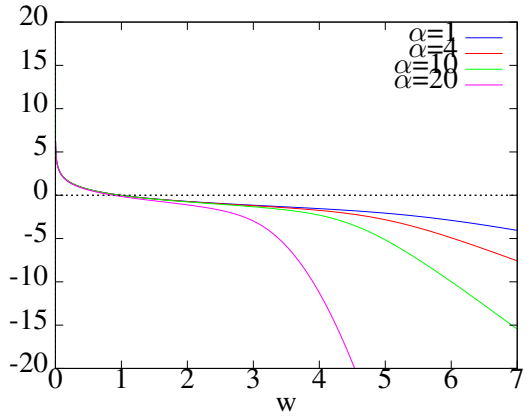
The above formulation of P_{S_1} provides only a proportionality relation, which needs to be normalised, in order to obtain a full definition. To do so, we may apply the same information-theoretic treatment to the speaker model as we did to the literal-listener model. First, we make explicit the normalisation factor in P_{S_1}; then, we encode this factor in terms

of information gain, relying again on Theorem 1:[5]

$$P_{S_1}(u|w) = \frac{l(u,w) \times P_{S_1}(u)}{\sum_{u_1} l(u_1,w) \times P_{S_1}(u_1)} = l(u,w) \times P_{S_1}(u) \times e^{G_{S_1}(w)} \quad (10)$$

Here, $G_{S_1}(w)$ is the information gain on the distribution P_{S_1} provided by w. This gain is high if $l(u,w)$ allows the speaker to discard many utterances u, where $P_{S_1}(u)$ is high. We refer to $G_{S_1}(w)$ as the *specificity* of w; in general, the function G_{S_1} is determined by the pragmatic interpretation problem $(\mathcal{U}, \mathcal{W}, P, C, l)$, together with the model temperature α.

In our running example, we obtain the following contours of specificity for various values of α:[6]



These curves can be analysed as sequences of three different regimes. First, there is asymptotic behaviour around 0: values near 0 are nearly impossible by virtue of excluding most *a priori* possible utterances, and thus they provide an information gain tending to infinity. The transition to the next regime happens very quickly, around 0.2. The middle regime is a small slope with a roughly flat decrease, which lasts up to around the mean of the prior distribution. Above the mean of the prior, the third regime kicks in: there is another roughly flat decrease, but, this time, with a much larger slope. The difference in slope is explained by the following two facts: (i) that the literal information gains associated with utterances increase drastically above the mean of the prior (see the plot of (Eq. (6))), and (ii) that these information gains enter into the calculation of specificity for world states above the mean, as these world states become compatible with more utterances. Moreover, for large values of α, this slope is more pronounced.

The reader may find it odd that there are negative specificities in these plots, given that the distributions that these specificities come from are obtained as *filters* of the pragmatic prior over utterances. Negative values appear because P_{S_1} is not, strictly speaking, a *probability distribution* over utterances (see Footnote 3). Fortunately, negative specificities do not pose a problem in practice. For example, once we get to the pragmatic listener model, they may be seen as having been shifted by a positive constant during normalisation (given that the posterior itself will be multiplied by a constant).

3.4 The reformulated pragmatic listener model

If we now substitute the definition of the speaker model of Eq. (10) into the definition of the pragmatic listener model of Eq. (3), we may obtain the following new definition of the latter:

$$\begin{aligned} P_{L_1}(w \mid u) &\propto P_{S_1}(u \mid w) \times P(w) \\ &\propto l(u,w) \times P_{S_1}(u) \times e^{G_{S_1}(w)} \times P(w) \\ &\propto l(u,w) \times e^{G_{S_1}(w)} \times P(w) \quad (11) \end{aligned}$$

The justification for removing the term $P_{S_1}(u)$ in the fourth line is the fact that u is fixed, and thus the proportionality relation does not depend on it. Now note that we may define the following *pragmatic prior* for the listener model:

$$P_{L_1}(w) = e^{G_{S_1}(w)} \times P(w) \quad (12)$$

Given Eq. (11), the pragmatic listener model may therefore instead be presented as a *filter* on this prior:

$$P_{L_1}(w|u) \propto l(u,w) \times P_{L_1}(w) \quad (13)$$

Equation (13) constitutes the fully reformulated RSA model. Reading it out, we see that L_1 chooses the distribution over world states in a way very

[5] Note that in practice, P_{S_1} in (10) may be substituted by the right-hand side of the proportionality relation in (8), since the relevant normalisation factor is cancelled out.

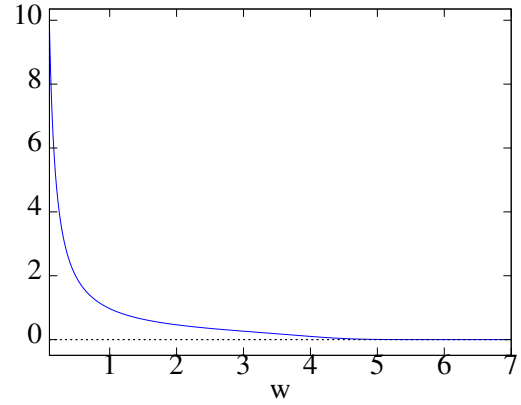
[6] Computed by taking

$$\begin{aligned} & G_{S_1}(w) \\ = & -\log(\sum_{u_1} l(u_1,w) \times P_{S_1}(u_1)) \\ = & -\log(\sum_{u_1} l(u_1,w) \times \frac{e^{\alpha \times (G_{L_0}(u_1) - C(u_1))}}{k}) \quad \text{(by (8))} \\ = & -\log(\sum_{u_1} l(u_1,w) \times e^{\alpha \times (G_{L_0}(u_1) - C(u_1))}) + \log(k) \end{aligned}$$

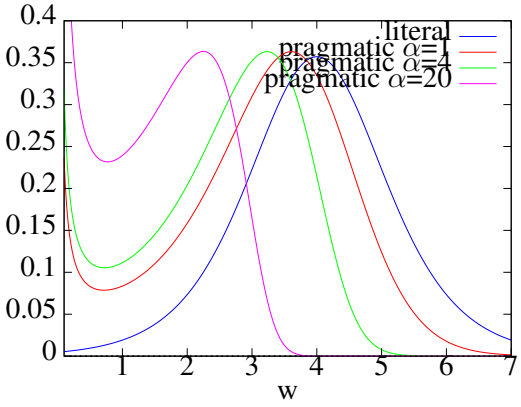
and choosing $\log(k) = 0$ (or $k = 1$). Note that the resulting contours are therefore independent of the normalisation factor k.

similar to L_0. Namely, both merely apply a filter to some prior. In the case of L_0, the relevant prior over world states w is $P(w)$, i.e., the "literal" prior; L_1, instead, uses the pragmatic prior $P_{L_1}(w)$, which multiplies the literal prior by a measure of specificity, defined as the exponentiated information gain associated with the pragmatic speaker. In sum, the RSA model has it that all pragmatic effects are attributable to the relative specificity of world states, relative to the set of possible utterances.

In our running example, the factor $e^{G_{S_1}(w)}$ contributing specificity has the following shape:

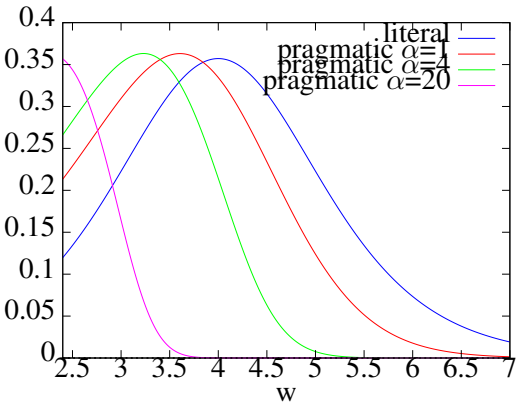


The following plot of P_{L_1} illustrates the effect that this factor has on the prior $P(w)$ over world states w (for various values of α):



The different regimes of specificity can thus be seen to have the following effects on P_{L_1}. Toward the left of the plot, there is an asymptote greatly favouring tiny values of w. Meanwhile, the right of the plot is essentially zeroed out, with a smooth transition, and the peak has also shifted leftward. Thus the bulk of the distribution is shifted to the left, in comparison to the prior P. This shift is larger when α is large; indeed, we would expect that for large enough values of α, the peak will "merge" with the asymptotic behaviour around zero. Unfortunately, our numeric tool cannot handle very large values of α, so we are not able to produce the corresponding plot.

Analogous to the literal listener, one can obtain $P_{L_1}(w \mid u)$ for any u by cropping the pragmatic listener's prior distribution on the left of the plot and renormalising. Consider, for instance, the utterance 'John ran 2.4 kilometres'. In this case, we crop the prior distribution at 2.4:



We indeed observe the effect of a Gricean implicature when $\alpha = 20$, insofar as the mode of the resulting distribution is the value uttered. This effect occurs due to the fact that the pragmatic prior distribution for this value of α happens to have a sharp drop after the uttered value. For lower values of α, however, the observed pragmatic effect is merely to shift the distribution to the left, relative to the literal prior. What results does not, in any obvious way, reflect a Gricean implicature.

4 Discussion

4.1 Algorithmic plausibility

At first glance, the psychological plausibility of RSA as an algorithmic model (in the sense of Marr (1982)) seems highly suspect; for example, it requires the pragmatic listener to consider all compatible combinations of world states and utterances on each occasion of utterance interpretation (though see, e.g., White et al. (2020); Zaslavsky et al. (2021) for recent attempts to address the psychological principles grounding RSA models). In principle, the space of world states includes all those literally compatible with the observed utterance, requiring the listener to deal with a very large space of possibilities, in order to interpret a single utterance. Because our reformulation of RSA as a mere filter on a prior is *functionally equivalent* to RSA as traditionally conceived, it provides a new lens into the issue of algorithmic plausibility. Neither the computation of the literal listener model, nor

the computation of the pragmatic speaker model need directly enter into the pragmatic listener's computation of a posterior distribution. Instead, the contributions of the literal listener and pragmatic speaker in the original formulation of RSA are now *repackaged* as part of the prior. As a result, these contributions may be learned and then "memorised";[7] utterance interpretation, meanwhile, becomes a process of merely eliminating alternatives from this re-conceived prior. (Likewise for the pragmatic speaker, whose prior over utterances need not depend on the world state that it wishes to communicate.)

A consequence of this fact is that literal interpretations and pragmatic interpretations (as according to RSA) may be viewed as updates of the same kind: given an utterance and the right prior, both styles of interpretation involve the elimination of world states incompatible with the utterance, followed by a renormalisation step. From this perspective, the RSA framework is not committed to a particular algorithmic implementation of the pragmatic interpretation process beyond what would be required for literal interpretation.

Relatedly, we do not require 3-d (or density) plots with axes representing utterance strength and world state, respectively, in order to illustrate the effect obtained by a pragmatic listener from sequential renormalisation steps. Our theoretical result may thus be framed as the observation that, in such a 3-d plot, and for a semantics of the sort $u \leq w$, any 2-d *slice* acquired by fixing a value for the utterance u is just like the slices associated with weaker utterances, but for a step of cropping and renormalisation.

4.2 Implicature

As mentioned in Section 3.4, the implicature expected based on Grice's Cooperative Principle (in particular, Quantity) is not obtained by the pragmatic listener model in our running example. The expected implicature is an "exactly" interpretation associated with the numeral occurring in the utterance, while what the model obtains is merely a decrease in the mode of the posterior, in comparison to the prior. (This result persists even for large values of α.)

Nevertheless, we can show that the expected implicature occurs when α tends to infinity.[8] This is because the utterance 'John ran $x + \varepsilon$ kilometers' is always, if only slightly, more informative than 'John ran x kilometers'; thus it will always be preferred by the pragmatic speaker for legitimate values of α, if only by a small amount. As a result, only $w = u$ will be admissible by a pragmatic listener, in the limit, where probabilistic choice becomes categorical.

Intriguingly, the theoretical result that the implicatures expected are not always generated may, in fact, reflect some aspects of real human behaviour in certain settings. For instance, Sikos et al. (2021a) found that, even in the non-interactive one-shot games against which RSA models have been most extensively tested, consistency with human performance was driven by cases in which non-Gricean behaviour was, in fact, predicted by the model. In such cases, the RSA model's prior overrode the pragmatic effects associated with specificity. Thus perhaps ironically, RSA's failure to predict Gricean implicatures may sometimes contribute to its empirical successes. As Sikos et al. note, however, RSA does not reflect human behaviour better than a literal semantic model does on such tasks, making it difficult to consider this property a boon.

5 Conclusions and future Work

RSA models can be critiqued on a number of fronts, both theoretical and empirical. In the present work, we have focused on the algorithmic nature of the model, in order to show that it may be reformulated in a manner which appears relatively attractive from a psychological perspective (and which is easier to compute in simulations). We have also shown that doing so brings to the fore the model's unexpected (i.e., non-Gricean) behaviour when it is faced with certain priors: a single plot illustrates the relationship between utterance and posterior in the pragmatic listener model, revealing limits on the conditions under which expected implicatures are actually generated.

On the empirical side, critics of RSA have emphasised the artificial and non-interactive nature of the tasks used to verify the model's predictions, pointing out, for example, that in less constrained contexts, people often produce non-optimal utter-

[7] A reviewer points out that our reformulation of RSA in terms of a pragmatic prior generates questions about how such a prior might be learned in the first place. While we won't provide an account of semantic learning here, we note that the two terms in (12) representing information gain and the prior over worlds suggest that they may be learned independently of one another.

[8] That is, as α tends to infinity, $P_{S_1}(u \mid w) = \delta_{u-w}$, where δ is the Dirac δ function.

ances, e.g., by over-specification (Gatt et al., 2013). Sikos et al. (2021a) have provided evidence suggesting that, even in the restricted reference game domain, people's judgements do not always accord with rational choice, as defined by the model. These authors show that, while speakers behave as the model would predict, listeners do not, and a baseline literal listener model outperforms RSA. It has also been argued that, even where RSA provides a good fit to human data, it does so when other parameters, such as utterance cost, are implausible (Wilcox and Spector, 2019). Moreover, simulation models have called into question whether or not reasoning over an interlocutor's intentions is generally necessary, if, for example, a repair mechanism is available (Van Arkel et al., 2020).[9]

By providing a functionally equivalent reformulation of RSA, we have shown that, for both the pragmatic listener and speaker models, the *merit* of a given world state or utterance can be expressed and evaluated in its own terms, making both models analogous to the literal listener model. Succinctly, RSA models are just *filters* of some chosen prior, in which merit is predetermined. It is thus straightforward to imagine a generalisation of our reformulation in which information gains are not always computed "rationally". Rather, according to such hypothetical alternative models, a speaker might compute a more approximate information gain for each utterance and act accordingly. Similarly, a listener might compute a more approximate notion of specificity with respect to a set of possible utterances, which may then be used to tweak the prior. Such "approximately rational" priors might then be refined over time as more pragmatic problems are encountered.[10] We leave this possibility for future work.

Acknowledgements

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics and Theory of Science at the University of Gothenburg. We thank the anonymous reviewers for their useful comments on an earlier draft of the paper.

[9] Such repair mechanisms are well-studied and highly prevalent in natural dialogue (Schegloff et al., 1977; Dingemanse et al., 2015).

[10] Such a generalised model is supported by findings in human-human reference games where "familiarity with the communicative setting can influence the degree of rationality that listeners realise" (Sikos et al., 2021b, p.1471).

References

Guanyi Chen, CJ van Deemter, and Chenghua Lin. 2018. Modelling pro-drop with the rational speech acts model. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66. Association for Computational Linguistics (ACL).

Mark Dingemanse, Seán G Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S Gisladottir, Kobin H Kendrick, Stephen C Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal principles in the repair of communication problems. *PloS one*, 10(9):e0136100.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336:998 – 998.

Michael C. Frank and Noah D. Goodman. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75:80–96.

Albert Gatt, Roger PG van Gompel, Kees van Deemter, and Emiel Krahmer. 2013. Are we Bayesian referring expression generators? In *Proceedings of the Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

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

Daniel Lassiter and Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory*, 23(0):587–610. Number: 0.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, Cambridge.

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

Gregory Scontras and Noah D. Goodman. 2017. Resolving uncertainty in plural predication. *Cognition*, 168:294–311.

Gregory Scontras, Michael Henry Tessler, and Michael Franke. A practical introduction to the rational speech act modeling framework. manuscript.

Les Sikos, Noortje J Venhuizen, Heiner Drenhaus, and Matthew W Crocker. 2021a. Reevaluating pragmatic reasoning in language games. *PloS one*, 16(3):e0248388.

Les Sikos, Noortje J Venhuizen, Heiner Drenhaus, and Matthew W Crocker. 2021b. Speak before you listen: Pragmatic reasoning in multi-trial language games.

In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Jacqueline Van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. 2020. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *the 24th (Virtual) Conference on Computational Natural Language Learning (CoNLL 2020)*, pages 177–194. ACL.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. ArXiv:2006.00418 [cs].

Ethan Wilcox and Benjamin Spector. 2019. The role of prior beliefs in the rational speech act model of pragmatics: Exhaustivity as a case study. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 3099–3105.

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021. A Rate–Distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online. Association for Computational Linguistics.

A Proofs

Theorem 1. *If $P(x) \propto f(x) \times Q(x)$, and the codomain of f is $\{0, 1\}$, then*

$$D_{KL}(P \parallel Q) = -\log\left(\sum_x f(x) \times Q(x)\right)$$

Proof. Let $P(x) = \alpha f(x) \times Q(x)$, with α constant.

First, we have

$$\alpha = \frac{1}{\sum_x f(x) \times Q(x)} \tag{14}$$

Indeed, P is a distribution, and we have

$$\sum_x P(x) = 1$$
$$\sum_x \alpha f(x) \times Q(x) = 1$$
$$\alpha \sum_x f(x) \times Q(x) = 1$$

Second, we have

$$\alpha f(x) \times \log(\alpha f(x)) = \alpha f(x) \times \log(\alpha) \tag{15}$$

This can be seen by case analysis.

- If $f(x) = 0$, then
$$\begin{aligned} & \alpha f(x) \times \log(\alpha f(x)) \\ = \; & 0 \\ = \; & \alpha f(x) \times \log(\alpha) \end{aligned}$$
The first equality follows from the fact that, in general, $\lim_{a \to 0}(a \times log(a)) = 0$.
- If $f(x) = 1$, then
$$\begin{aligned} & \alpha f(x) \times \log(\alpha f(x)) \\ = \; & \alpha \log(\alpha) \\ = \; & \alpha f(x) \times \log(\alpha) \end{aligned}$$

Using the above two facts, we can compute:

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_x P(x) \times \log(\frac{P(x)}{Q(x)}) \\ &= \sum_x Q(x) \times \alpha f(x) \times \log(\alpha f(x))) && \text{by def of } P \\ &= \sum_x Q(x) \times \alpha f(x) \times \log(\alpha) && \text{by Eq. (15)} \\ &= \alpha \log(\alpha) \times \sum_x f(x) \times Q(x) \\ &= \alpha \log(\alpha) \times \alpha^{-1} && \text{by Eq. (14)} \\ &= \log(\alpha) \\ &= \log(\frac{1}{\sum_x f(x) \times Q(x)}) && \text{by Eq. (14)} \\ &= -\log(\sum_x f(x) \times Q(x)) \end{aligned}$$

□

Relationality is Not Enough: The Organization of Dynamic Structures

Maximilian Zachrau
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Box 200, SE40530 Sweden
maximilian.zachrau@gu.se

Abstract

The interactivist model (Bickhard, 2009b) posits action and, importantly *inter*action to be the key notions based on which a wide array of phenomena are understood better than traditional models of representation, cognition, perception etc. The metaphysical foundations on which it rests comprise both dynamic process philosophy and a strong relational framework. The paper intends to demonstrate that these two pillars are not accidentally meeting each other on this fundament. Interactivism requires that processes demand relationality, and relational structures need a dynamic interpretation. This latter conceptualisation of structures as dynamic, labelled here a metaphysics of dynamic structures, has only recently gotten some traction. I explore some programmatic ideas and consequences while calling for further investigation into these dynamic structures.

1 Shift to Relationality

It is a relatively uncontroversial claim that physics throughout the 20th century has puzzled and even troubled our ordinary thinking of how the world is. The nature of space, time, entanglement, objective probability, and the rise of field theory all challenge fundamental ideas of traditional physics and our understanding of it.

Somewhat more controversial (but still almost entirely agreed upon) is the observation that many of the new theories and ideas shift away from individual and independent particles to relations, systems, and structures of those.[1] Entities formerly thought to be self-sufficient ultimate "building blocks", like the atom, space or particles, could either be dissolved further or are now understood only in the system they are embedded in and often relative to further constraints.

Just to sketch two examples: In the good-old Newtonian world with absolute space and time, there was a sense in which an object, a particle, is moving - independent of any other thing. But the relativity of space makes it necessary for at least one other object to exist. It is only with respect to this other object that the former can change position relative to and therefore move. The spatial position and movement, formerly conceived of as absolute notions, now involve a relation to other entities essentially. Things become even more intricate with the introduction of spacetime.

The other example concerns quantum mechanics: The mathematical framework in quantum mechanics yields the wave function as the description of how the system develops over time. The classical analogy for such a function is a function that describes the behaviour of the constituting particles. In the classic-mechanical framework, it makes sense to think that the particles are the fundamental objects and the function is the derivative description of their behaviour. The particles are in a certain state, and the function describing the system merely sums up the individual states into a system description.

However, in the quantum mechanical setting, it is difficult to uphold that individual particles constitute the system in every scenario.[2] Some authors argue that instead, the wave function is what is truly real.[3] Again, the supposedly independent,

[1] Structures and systems are understood to denote the totality of entities involved, including the connecting relation. This mathematics-inspired definition means to be innocuous and broadly applicable to all types of examples and sciences. Following the definition, every structure involves a relation, and every relation induces a structure. For this reason, the "shift to relationality" is synonymous with a "shift to structures" and a "shift to systems". Furthermore, I will use the terms "relationality", "structural", and "systematic" interchangeably.

[2] This has partly to do with the problem that the traditional means of individuating them are failing. In his article Muller (2015) echoes the undercurrent of a major debate in the philosophy of quantum mechanics and physics in general, namely that the strong sense of individuality, and absolute discernibility, cannot be upheld in all scenarios of modern physics. His solution will be that there are distinct particles involved but that they have to be re-characterised as relationals - repeating the alleged general shift here.

[3] Compare David Albert's summary of the development of

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

self-sufficient particles on the new picture can *only* be understood in their position within the system as arguments in the wave function. They are essentially related to one another concerning this system description.

I would like to label the supposedly very general shift, from constituents to the structures thereof, as the shift to relationality. The shift to relationality will be short for the general advance of the idea that the former basic, independent entities now are taken to be related so strongly that they essentially involve the structure, the system they stand in.[4]

2 Internality of Relations

Never mind the accurateness of the history of science thesis. Maybe relations have always played an important role, or perhaps the alleged shift is not as pervasive as suggested. Regardless, in the promising framework of Interactivism (Bickhard, 2009a) and akin ideas like the enactive approach (McGann et al., 2013) relationality features prominently. Interactivism evolved as a model of representation (cf., Bickhard (2009b, p.548)) where the core or "minimal model of representation"(Bickhard (2009b, p.570)) form anticipatory activities which have truth-value (cf., (Bickhard, 2009b, p.570) and (Bickhard, 2009a)). On the basis of "interactive goal-directed systems"(Bickhard (1998a, p.212)) doing those anticipatory activities emerge the multi-level phenomenon of representation (cf., (Bickhard, 1998a) and (Bickhard, 1998b, p.6)).

> In this model, representation emerges naturally in the problem of the selection of actions and interactions by agents – it is an interactive model of representation.(Bickhard, 1998b, p.3)

The model not only meets a meta-epistemological criterion as well as the normative criterion, which are hardly even addressed by competing accounts (Bickhard (1999, p.435)). It also fulfills the crucial desiderata of a model of representation (Bickhard (2009b, p.569)). While representation is where Interactivism started, the model developed into a much more encompassing model, including related phenomena such as cognition, language and normative biological functions (Bickhard (2009b, p.548)). To Interactivism, the metaphysical foundation is of crucial importance. It subscribes to a more general shift towards processes metaphysics throughout the sciences (cf., (Bickhard, 2019, p.228)) and contributes to it by aligning representation and cognition with this general shift.

Interactivism also features relationality, where relationality is more than the mere acceptance of relations. Arguably, relations in some way or other do play a role in systems that do not stress relationality as much. Billiard balls, mathematical points and qualities considered "entirely loose and separate"[5] do feature in the respective systems as relata, e.g., as causal relata. One billiard ball and its motion can cause the other billiard to roll; two electrons are in a specific distance relation (say $5,2 \times 10^{-11}$m apart). Crucially, when considered "loose and separate", they could also not be related in that way or even not at all. Their being does not necessarily involve any connection.

Conversely, in a relational framework such as Interactivism, entities involve (at least some) relations essentially (Bickhard (2009a) and Bickhard (2019, p.230)). Relations of that kind are "internal" relations. They are "intrinsic to the nature of one or more of the relata. They are a kind of essential relation, rather than an essential property." (Bickhard, 2003, p.1)

the GRW interpretation: "And this new approach very naturally brought with it a new and more straightforward and more flat-footed and more traditionally scientific way of thinking about the wave-function *itself*. This *new* way of thinking turns everything about the foregoing tradition elegantly inside out: The wave-function is not an abstract mathematical representation of the states of concrete physical systems, but (rather) the unique fundamental concrete physical stuff of the world *itself*. First-quantised non-relativistic quantum mechanics is not a theory of the 3-dimensional motions of *particles*, but (rather) of the 3 N-dimensional *undulations* of a concrete physical field – which is nothing other than the wave-function itself – where N is a very large number that corresponds, on the *old* way of thinking, to the number of elementary particles in the universe. And once this new picture is fully taken in, there are no longer any such metaphysical conundrums in the world as indeterminacy or superposition or non-separability:[...]" (Albert, 2019, p.92-93).

[4] Of course, recursion is allowed. The systems themselves may be integrated into even larger systems, and supposing that the interdependency among the systems is strong enough, one may go on to put that even larger system at the basic level.

[5] It is no coincidence that a quote from Hume (2007, p.58) enters the picture here. Nowhere but in Humean metaphysics we find the opposition to relationality expressed as strongly. "Humean supervenience is named in honor of the greater denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another.[...]. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated." (Lewis, 1986, p.ix-x).

Circumventing some worries about "intrinsic" [6], the notion can be cast into a semi-logical formula:

(Internality) R is an internal relation $=_{df}$ $(\forall x_1) \ldots (\forall x_n)$ if $Rx_1 \ldots x_n$ then necessarily $((x_1$ exists$\leftrightarrow Rx_1 \ldots x_n)$ & $\ldots$ & $(x_n$ exists$\leftrightarrow Rx_1 \ldots x_n))$(Schaffer, 2010a, p.349).

An internal relation allows to infer from the existence of any of the relata the holding of the relation. In that sense, the relation "flows" from the nature of their relata. But a word of caution is advised. Internality should not be confused with the idea that relations reduce to monadic properties.[7]

An example helps to illustrate that point. If there are two mountains, each being 5000m high, they stand in the relation of "being of equal height". The relation flows from their nature in that it is reducible to the monadic properties. In virtue of having the properties, the relation holds. Yet, this relation is not internal! We cannot deduce that the relation holds from the existence of one of the mountains. The other mountain may be of a different height or not exist at all.

The example demonstrates an important consequence of internal relations too. Internal relatedness leads to interdependence.[8] From the existence of one of the relata in an internal relation, we can infer the relation holding, from which, in turn, we can infer the existence of the other relata. And vice versa. Neither relata can exist without the other because of their strong structural connection. Interdependency is an immediate consequence of having relations in the nature of the relata.

To Bickhard, who makes a very strong point, that process metaphysics is the proper foundation for Interactivism, processes are such relata. Processes are related to one another in terms of Organization (cf., Bickhard (2009b, p.554).) and "[a] process, however, has whatever properties it has, including causal properties, in virtue (in part) of its organisation: new organisations may generate new (causal) properties [...]. "(Bickhard, 2011, p.7)

The burning of a candle is organised with the inflow of fresh oxygen, melting of wax and the oxidation of the wick. If these other processes were different, the burning of the candle would be different. Suppose the oxygen inflow was to speed up by some variant of a chimney effect. Consequently, the burning of the candle would drastically alter its characteristics, getting much hotter and brighter and causing the melting of wax and oxidation of the wick to speed up. The burning of the candle involves those other processes essentially, making it and its characteristics dependent on them.

Relationality, i.e., the internal relatedness, demands a perspective where the structure (the organisation or network) the entities are embedded in is of critical importance. Entities are partly constituted by their relations and interconnections promoting the overall structure. Where formerly structure, or the whole, really was an abstract conglomerate of "local matters" [9] these "local matters" now depend on the global network in which they are. Metaphysically speaking, the structure is no longer derivative to the individual particulars and their intrinsic characteristics. From the relational view, the structure is at least as fundamental as the structured content itself.

There is a worry to be addressed here with the formulation of the fundamentality of structure. Starting with processes and emphasising their interconnection, one need not necessarily end up with a picture where a structure is considered a separate entity. Therefore, the identification of relationality with the thesis claiming that structure is fundamental is ill-conceived. That is because there may not be "a structure" on the list of beings which then gets awarded with fundamentality. While that is true, it does not change much regarding the upcoming argument. Relationality claims that something about the relata (the processes) is such that you cannot understand them without embedding them into the relations with other relata. At the very least, that suggests that there is a part of each of the relata which is essentially connected to other relata. Structure can then be understood to be short for all these parts. By relationality, we know that such parts are not derivative to other (intrinsic) aspects of the relata. The fundamentality of structure need not amount to more than this.

[6] Schaffer makes the valid point that instead of intrinsic, the proper internal notion should be in terms of essentiality, cf., Schaffer (2010a, p.348-349).

[7] The reducibility of relations to monadic properties is sometimes confused with internality. Yates in (Yates, 2016) clearly distinguishes between the two. Moreover, in the course of this paper, it will become clear how internal relations may just as well be fundamental, say in the case of the discussed Ontic Structural Realism (OSR).

[8] Again compare Yates (2016) and Schaffer (2010a).

[9] Compare footnote 5.

3 Static Relationality, Static Structures

At least two prominent positions have addressed relationality in the same way as it has been treated here so far: Priority Monism and Ontic Structural Realism (OSR).

Priority Monism

According to Priority Monism the relational character of the interdependent particulars makes it necessary to see the whole as more fundamental than its parts (Schaffer, 2010a). While any part is a dependent entity, the whole, the sum of all parts and thereby including structure, is not dependent in the same way. All the dependencies are "resolved" within the whole, all the parts depend upon the whole but not the other way around.

> The monist holds that the whole is prior to its parts, and thus views the cosmos as fundamental, with metaphysical explanation dangling downward from the One. (Schaffer, 2010b)

Arguably, one can also imagine that there are levels in the Priority Monism picture. Physical particles may depend upon one another, and the larger whole could be protons, and neutrons, so physical particles of higher complexity. Those higher complexity particles may again depend upon another and jointly be parts of the larger whole, specific atoms, which again depend upon another to make up molecules etc. Given such a chain, only the most extensive whole is fundamental, even if for the dependency to end, such a largest whole could be the entire universe consisting of literally everything there is.

Bradley famously thought so, as according to him, everything was related internally to everything else and recently, Jonathan Schaffer has defended priority monism on similar grounds (Bradley (1897), Schaffer (2010a), and Schaffer (2010b)).

Ontic Structural Realism (OSR)

Ontic Structural Realism (OSR) assumes too that relata can only be understood in terms of the relations and thus their connection to other relata. Yet, the position is slightly more radical. Not only are some aspects of the relata derivative to structure, but everything about them. The understanding of any aspect of a relatum requires the reference to relations. In consequence the entities are nothing but relata, points in a web of relations.[10]

> Ontic Structural Realism (OSR) is the view that the world has an objective modal structure that is ontologically fundamental, in the sense of not supervening on the intrinsic properties of a set of individuals. According to OSR, even the identity and individuality of objects depend on the relational structure of the world. Hence, a first approximation to our metaphysics is: 'There are no things. Structure is all there is.' (Ladyman et al., 2007, p.130)

Structuralism traditionally was advocated mainly in the fields of philosophy of language and mathematics (cf., Shapiro (1997)). But it has had a career as an interesting realist position in philosophy of science as well.[11]

Relationality is the idea that structure is non-derivative - it is at least as fundamental as its relata. On Priority Monism and OSR that idea is taken even further, namely that structure or the whole is even more fundamental than the relata. Here, the relata are derivative to the network. Nevertheless, there is no denying that both Priority Monism and Structuralism incorporate relationality.

The central argument of the paper at hand is that relationality can both be developed in a static form and a dynamic form. The next section will tackle the dynamic form of relationality and explore the notion further. Here, however, it is argued that the relationality provided by Priority Monism and OSR *can* be understood as static. Neither Priority Monism nor OSR is a Process Philosophy.[12] On the contrary, they share many, albeit not all, presuppositions with the opposing substance paradigm.[13] That is not to say that they cannot have processes as their relata. Instead, the crucial point is that structure itself is not dynamic.[14]

[10] Compare as examples, Ladyman et al. (2007), French (2014) and Muller (2015).

[11] See footnote 10.

[12] To be more precise, the dominant variants of these theories are not Process Philosophies. As this article intends to show, there is a way of reading structure dynamically and thereby integrating the structural priority with the fundamentality of processes. Some deviant forms of structuralism have also noted this connection, cf., (Ferrari, 2021).

[13] For a characterisation of that paradigm see Seibt (1990, Appendix).

[14] Arguably, Structuralism and Priority Monism are well within their right to also integrate dynamic structure, rendering the verdict of static against them empty. Part of the argument of this article was to point out that relationality does not by itself lead to a dynamic process view. For that, it suffices that Priority Monism and OSR are at least compatible with a static

The traditional conception of structure is reflected in the foundations of classical logic and set theory. Structure and relations are in essence sets of a specific form. Regardless of the exact items within the set, A, the relation is but a mere subset of $A \times A$. The usual attributes of the relation, e.g., reflexivity, transitivity and symmetry, are then but specific demands on which pairs in $A \times A$ have to be included or must not be included in the subset. Due to the definition by extension, if any of the pairs were added or subtracted from the relation, it yields a different relation. Relations cannot "change" or be brought about. It is only the items that can undergo change and thereby exemplify new relations or structures.

Coming from mathematics and language, structuralist use this abstract and static conception of structure to interpret the physical world. Only recently, some non-classical branches in logic and alternatives to set-theory in mathematics are making way to explore on an abstract level the dynamicity of structure(cf.,(Baltag and Smets, 2011)). Notwithstanding, the predominant conception of structure is one without the possibility of a dynamic interpretation.

And since structure on these relational views is fundamental, fundamentality points away from dynamicity. On the ultimate fundamental level, there are no dynamic but static entities, namely the structure or the whole.

Consequently, relationality leads to a systematic/structural view compatible with a static substance paradigm. Even more, many systematic views derive their appeal within the substance paradigm by providing even more stability than their particle-view competitors. The structure is deemed to be even more stable and permanent than its residents (cf., French (2014)), which is why many substance metaphysicians could cast their metaphysical quest as the inquiry into the "most fundamental structure of reality "(Lowe et al., 1998, p.1).

Furthermore, this compatibility of systematic views with static paradigms holds, even if the entities within the systems and structures are processes. Because of the relational character, an advocate of the substance paradigm may uphold that structure understood as something static is fundamental, whereas the processes involved in those structures may be dynamic, but *derivative* entities.

substance view, which is how they are usually perceived. The argument did not intend to demonstrate that any specific view is not able to move to a more dynamic view on structures.

4 Dynamic Structures

Interactivism, on the other hand, strikes me as a fundamentally dynamic view. Not only are important features, like emergence supposedly dependent on a process metaphysics[15], but also the view in itself puts activity (change, dynamicity) before stability, objects and substances.[16] The same goes for the enactive approach to cognitive science, where "the mind is seen not as inhering in the individual, but as emerging, *existing dynamically* in the relationship between organisms and their surroundings (including other agents)" (McGann et al., 2013, p.203, my italics). To dynamic views of that kind change and processes feature on the fundamental level of reality, they are not to be reduced to states, properties and substance.

However, in the previous section, it was argued that one way to understand structure is in terms of a static system of relations. On such a static account, the fundamentality of structure, deriving from the relational character of processes, is in stark tension with the dynamicity of the view. After all, now something static underlies the character of the processes defining their being essentially. Following the static conception of structure, we run into a conflict between the "inter" (relationality) and the "activism" (dynamicity). Accordingly, the solution is to pursue a non-static, dynamic conception of structure, where not only the relata are dynamic but relating is a process.[17] Thereby, the superficially supposed conflict between "inter" and "activism" is dispersed.

To Bickhard, there is no tension between dynamicity and relationality, but rather the shift to relationality is a strong argument in favour of dy-

[15] An argument by Bickhard made in several articles, compare for instance Bickhard (2019) and Bickhard (2009b).

[16] "For a substance metaphysics, stasis or inertness is the default. Change requires explanation. In contrast, process is inherently and always changing - a return to Heraclitus, if you will. Change is the default. In such a view, any stability of organisation or pattern of process requires explanation - and we will find that the kinds of these explanations can be of fundamental importance." (Bickhard, 2011, p.6).

[17] I echo here something very similar to the point De Jaegher, Peräkylä and Stevanovic make in distinguishing coordination in interactional sociology from coordination in enactivism. "Unlike interactional sociology, which highlights the structures that facilitate coordination, enaction describes interactional organisation in terms of dynamic, emergent processes of coordination." (De Jaegher et al., 2016, p.4) However, I believe that Luhmann, in contrast to Goffman, may have had this dynamic aspect of organisation more in mind.

namicity.

> "As mentioned above, it is this relationality that I will be arguing is most important in the shift to a process framework." (Bickhard, 2011, p.13)

The two "pillars" of interactivism are not an arbitrary selection but a natural fit. Yet, this line of reasoning from relationality to the process framework only holds with the dynamic conception of structure and relations working in the background. It is this assumption, the dynamicity of structure, that the paper at hand intends to draw attention to. Without the assumption, relationality and priority of dynamicity are in tension instead of overlapping. However, even with the assumption there is the threat of circularity.

The inference from relationality to the process framework only holds with the assumption of dynamic structures which already seems influenced by the priority of dynamicity. I can see two possible replies: (a) One is to embrace the critique and drop the argument from relationality to the process framework. Instead, one could defend relationality by itself and the process framework on other grounds. Then the dynamic account of structures would drop out as a consequence of the combination of these two pillars. (b) The other option is to challenge the critique. The auspicious exploration of dynamic accounts of structures may find that this conception of relation is fruitful and promising in its own right without presupposing the priority of dynamicity. It is but a somewhat recent trend that dynamic logic with a focus on action instead of propositional descriptions draws attention (cf., (Baltag and Smets, 2011, p.287)). Again some of the driving factors for such a trend-shift are coming from the many problems that the application of "static" views brings in many areas of modern science (cf., Baltag and Smets (2011)). While Process Philosophers and Interactivists alike should welcome such dynamic shifts, there is still a lot of work to do, and dynamicity itself has to be spelt out further so as not to become an empty phrase. The shift to relationality plus independent reasons for conceiving relationality as dynamic leads to a process framework. Either way, the dynamic account of structures and relations requires further investigation.

Organization

While the overarching paradigm still is Process Philosophy, I want to label such special positions "metaphysics of dynamic structures". On such views, processes are not cast into preshaped moulds but rather woven into a fabric where thread and fabric are coming into being as processes.[18] Aligning with the Interactivist-picture we can call such a fabric-process: “organization” (Bickhard, 2009b, p.554). The term resonates nicely with Whitehead’s "philosophy of organism" (Whitehead et al., 1978) as well as Luhmann’s view on Systemtheory (Luhmann et al., 2013)[19], the latter of which gives an insightful example to the view:

> The structures can only be built through the system’s own operations. It is a circular process: structures can be built only through the system’s own operations because the system’s own structures in turn determine operations. This is obvious in the case of the biochemical cell structure, for the operations simultaneously contribute to the build-up of the programs – in this case, the enzymes – in accordance with which the cell regenerates structures as well as operations.(Luhmann et al., 2013, p.76)[20]

Within the cell, the processes (the "operations") relate internally via the programs which organise them. However, the programs stemming from this organisation are "*simultaneously*" built up by the processes. The "building" metaphor is slightly misleading, one could think of the structure to be a static product of a dynamic construction process. That is not what is meant here. The structure is inseparable from its constituting process, namely the dynamic bringing about of structure, *is* the structure. The example further shows that organisation need not necessarily be a separate process from the processes organised. It could be an aspect of

[18] In Sellar’s words: the world is "the ongoing tissue of goings-on"(Sellars, 1981, p.81).

[19] That interactional sociology is coming from a very similar perspective, has not gone unnoticed. In (De Jaegher et al., 2016) some of the parallels are worked out.

[20] Identifying the operations with the related processes makes the apparent parallel to Luhmann striking. Never mind the idea that the system by itself must produce both those operations and structures, which is a consequence of the idea of closed systems. The important point to notice here is that both structure and operation must be produced and that they mutually determine and create another. Luhmann too calls the production of structure "organisation" (and in the theory of closed systems, therefore "self-organisation" (Luhmann et al., 2013, p.70-71)).

such processes in some cases, and then that aspect can bear the name organisation. For current purposes, the important idea is merely that relat*ing* is understood dynamically.

Complexity - General Process Theory

One immediate consequence of the idea is that (at least some) processes are complex. In so far as they constitute the weaving of other processes into the fabric, they need to be sufficiently intricate. Process Philosophy and, more specifically, the General Process Theory (GPT) by Johanna Seibt has been aware of that special process.

> GPT is a mono-categoreal domain theory whose one basic category is called 'general process' or (for expositional purposes also) 'dynamics.' This category is defined in terms of a new configuration of familiar category features: dynamics are concrete, non-particular, non-countable (in the traditional sense of countability that implies necessary uniqueness, i.e., particularity, yet countable in the way in which we count kinds), more or less indeterminate or determinable, independent, dynamic individuals. The core claim of GPT is that whatever we reason about in common sense and science can be described as a type of dynamics.(Seibt, 2018, p.138)

The non-particular nature allows for the processes ("dynamics" in the quote) to be multiply occurrent. They not only "stretch" across a certain spacetime region, but the region they exist in could be disconnected. Such processes are well suited to bring about structure, as this allows for the processes to relate without requiring immediate vicinity. Processes can literally recur, allowing structuring processes to be present throughout their relata.

In GPT structure of processes and within processes play an important role throughout. Dynamics have a mereological signature, a participant structure, dynamic composition, dynamic shape and context (cf., (Seibt, 2018, p.141)). All these spell out different structural notions, and as they are aspects of dynamics, different structur*ing* notions.

It is no coincidence that Seibt's framework seems particularly fit to model the new "troublesome" (from the viewpoint of traditional particle/substance metaphysics) physical entities, e.g., those arising from Quantum-Field-Theory (QFT). Echoing earlier sections of the paper, that too is a domain where the "loose and separate "-ness makes way in favour of a more systematic approach. Cutting some more detailed and sensitive analysis short, GPT (back then under the name of APT (Axiomatic Process Theory)) offers complex processes, "the interaction of component processes [...]. [...] a dynamic "mixture" of dynamic "stuffs" [...] as assistance to the interpretation of QFT" (cf., Seibt (2002)). Thereby the interconnectedness and relationality are paid tribute to, yet the structure itself is understood to be dynamic, a complex process.

In a text on Sellars, Seibt addresses that structure in a process ontology should be understood dynamically: "Since pure processes are occurring-suchly's or *modes* of spatiotemporal occurrence, in a process ontology the mode or configuration in which processes occur is itself a process." (Seibt, 2016, p.196) And further in the footnote on that sentence: "E.g., a vortex is the mode in which certain other processes (movements of water molecules) occur, photosynthesis is the mode (configuration) in which other processes occur, and so forth. [...] the dynamic organisation of processes can count as a process itself." (Seibt, 2016, Footote 20)

The enormous task of developing a complete metaphysics of dynamic structures still lies ahead. This small paper can but draw attention to the need for further work and the promise such work holds. Confined in analysing one well-developed process metaphysical application, Interactivism, a programmatic emphasis was put on the conceptualisation of structure as dynamic. With its many variants and applications, process philosophy should welcome such insights - insights stemming from one of its applications and then inducing feedback for the overall view. Not only Interactivism, but process philosophy, in general, can only gain by further developing an account of dynamic structures.

At this point, the departure from "static" conceptions of structure allows for two minor remarks about their place in this new framework.

The first remark is that arguably not every relation needs to be understood processual. The alleged shift to relationality and the Interactivist perspectives feature central constitutive relations. Those are the relations essential to the entities, and those, I argued, require a dynamic interpretation within the Interactivist model. Still, other relations may enter as surplus structures but not play such consti-

tutive roles. Those relations need not necessarily be understood dynamically.

Other entities may derive from the "ongoing tissue of goings-on" that is, the fundamental processes and structure. Simons has an account of processes according to which static objects can be thought to be derivative patterns of processes, thereby accomplishing everything static substance metaphysics does while still having processes at rock bottom (cf., Simons (2018)). In such a case, relations and structure holding between the derivative static objects too will be derivative. Those relations and structures need not necessarily be conceptualised dynamically. Because these are derivative to the "Final Realities" (Whitehead et al., 1978, p.22) which are dynamic in nature, their static character poses no threat to the fundamentality of dynamicity.

Some relations were argued to be essential to processes, those constituting parts of the actual concrete world of becomings. The other relations owe their existence to higher degrees of abstractness, where we move away from the fundamental concrete constituents to broader patterns, kinds and relationships between those. So, again, there is the option to include these in the picture but not have them as residents on the ground floor.

The other place (or rather time) for static structures and relations is in the past. While the world is "the *ongoing* tissue of goings-on "(Sellars, 1981, p.81, my italics) one could imagine that this dynamic happening leaves a trail. The activities ongoing now pass into the past and are no longer ongoing, but rather *have gone on*. Retrospectively processes are still activities with temporal extension, but they are no longer dynamic in the sense of being currently ongoing. From the time of finishing onwards, they will always remain exactly what and where they are. I believe that those with sympathies for the recent revival of the Growing-Block view of time (cf., Correia and Rosenkranz (2018)) may be interested in that application of the dynamic structure view outlined here.

The growing block view can be summed up by the principle that what the widest quantifier ranges over always increases. The totality of existence forms a four-dimensional block, where entities are located at their respective space and time. Yet, on the "edge" of the block (defined by those entities which do not have anything later than them), new slices are coming into being, increasing the block in the direction of time. This growth constitutes the block as all parts of it once were created due to that process of growing. Combining it with the dynamic structure view here, one could say that processes are considered ongoing at the edge, and their activity is fundamental. As these processes were argued to stand in some internal relations, these relations too are processes ongoing at the edge of being. In contrast, the processes that *have gone on* are now located in the past. They may still be related, but the involved relations are not something ongoing and need not be further conceived of as dynamic. The residue of the ongoing process of becoming are entities that have become and no longer need the dynamicity of that process.

Both the integration of abstract static relations and the combination with a growing-block view of time are mere options the dynamic structure perspective offers for accommodating their static counterparts. By no means is one obliged to follow these tracks but coming from the discussion of static structures, it may present a valuable perk to the view to be able to integrate them.

Conclusion

Relationality features prominently in both scientific theories and philosophical interpretations and approaches. When taken seriously, entities can no longer be studied separately and individually but must be seen in the nexus from which they were taken. Some have drawn radical consequences from this interrelatedness, like OSR and Priority Monism. This paper argues that for dynamic views, which put processes before substances, an account of relations and structure must be spelt out, reflecting said dynamicity. Relations need to be understood as processes of relating, lest they challenge the fundamentality of processes, creating a hiatus between "inter" (representing the relationality) and "activism" (the undercurrents of Process Philosophy). I called attempts of such nature "metaphysics of dynamic structure" and explored some accounts and consequences.

References

David Albert. 2019. Preliminary considerations on the emergence of space and time. In *Philosophers Look at Quantum Mechanics*, pages 87–95. Springer.

Alexanu Baltag and Sonja Smets. 2011. Quantum logic as a dynamic logic. *Synthese (Dordrecht)*, 179(2):285–306.

Mark H. Bickhard. 1998a. Levels of representationality. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(2):179–215.

Mark H. Bickhard. 1998b. A process model of the emergence of representation. In G. L. Farre and T. Oksala, editors, *Emergence, Complexity, Hierarchy, Organization, Selected and Edited Papers From the Echo Iii Conference*, pages 3–7. Acta Polytechnica Scandinavica.

Mark H. Bickhard. 1999. Interaction and representation. *Theory & Psychology*, 9(4):435–458.

Mark H Bickhard. 2003. Some notes on internal and external relations and representation. *Consciousness & Emotion*, 4(1):101–110.

Mark H Bickhard. 2009a. Interactivism: A manifesto. *New Ideas in Psychology*, 27(1):85–95.

Mark H Bickhard. 2009b. The interactivist model. *Synthese*, 166(3):547–591.

Mark H Bickhard. 2011. Some consequences (and enablings) of process metaphysics. *Axiomathes*, 21(1):3–32.

Mark H Bickhard. 2019. Dynamics is not enough: An interactivist perspective. *Human Development*, 63(3-4):227–244.

Francis Herbert Bradley. 1897. *Appearance and reality: a metaphysical essay*. Oxford: Clarendon Press.

Fabrice Correia and Sven Rosenkranz. 2018. *Nothing to Come: A Defence of the Growing Block Theory of Time*. Cham, Switzerland: Springer Verlag.

Hanne De Jaegher, Anssi Peräkylä, and Melisa Stevanovic. 2016. The co-creation of meaningful action: bridging enaction and interactional sociology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150378.

Francesco Ferrari. 2021. Process-based entities are relational structures. from whitehead to structuralism. *Manuscrito*, 44:149–207.

Steven French. 2014. *The structure of the world: Metaphysics and representation*. Oxford University Press.

David Hume. 2007. *An Enquiry concerning Human Understanding*. Oxford World's Classics. Oxford University Press, UK, Oxford.

James Ladyman, Don Ross, Don Spurrett, David Spurrett, John Collier, John Gordon Collier, et al. 2007. *Every thing must go: Metaphysics naturalized*. Oxford University Press on Demand.

David Lewis. 1986. *Philosophical Papers: Volume 2*. Oxford University Press, Oxford.

E Jonathan Lowe, Jonathan Lowe, et al. 1998. *The possibility of metaphysics: Substance, identity, and time*, volume 181. Clarendon Press Oxford.

Niklas Luhmann, Dirk Baecker (edited by), and Peter Gilgen (translated by). 2013. *Introduction to systems theory*. Polity Cambridge.

Marek McGann, Hanne De Jaegher, and Ezequiel Di Paolo. 2013. Enaction and psychology. *Review of General Psychology*, 17(2):203–209.

Fred A Muller. 2015. The rise of relationals. *Mind*, 124(493):201–237.

Jonathan Schaffer. 2010a. The internal relatedness of all things. *Mind*, 119(474):341–376.

Jonathan Schaffer. 2010b. Monism: The priority of the whole. *The Philosophical Review*, 119(1):31–76.

Johanna Seibt. 1990. *Towards process ontology: A critical study in substance-ontological premises*. Ph.D. thesis, University of Pittsburgh.

Johanna Seibt. 2002. "quanta," tropes, or processes: Ontologies for qft beyond the myth of substance. In Meinard Kuhlmann, Holger Lyre, and Andrew Wayne, editors, *Ontological Aspects of Quantum Field Theory*, chapter 3, pages 53–97. World Scientific Publishing Co Pte Ltd.

Johanna Seibt. 2016. How to naturalize sensory consciousness and intentionality within a process monism with normativity gradient. In James R. O'Shea, editor, *Sellars and his Legacy*, pages 187–223. Oxford University Press, Oxford.

Johanna Seibt. 2018. What is a process? modes of occurrence and forms of dynamicity in general process theory. In Roland Stout, editor, *Process, Action, and Experience*, pages 120–148. Oxford: Oxford University Press.

Wilfrid Sellars. 1981. Foundations for a metaphysics of pure process: The carus lectures of wilfried sellars. *The Monist*, 64(1):3–90.

Stewart Shapiro. 1997. *Philosophy of mathematics: Structure and ontology*. Oxford University Press on Demand.

Peter Simons. 2018. Processes and precipitates. In Daniel J. Nicholson and John Dupré, editors, *Everything Flows: Towards a Processual Philosophy of Biology*. Oxford University Press.

Alfred North Whitehead, (Ed.) David Ray Griffin, and (Ed.) Donald W. Sherburne. 1978. *Process and reality*. New York: Free Press.

David Yates. 2016. Internal and external relations. In Anna Marmodoro and David Yates, editors, *The Metaphysics of Relations*, chapter Introduction, pages 7–14. Oxford University Press Oxford.

Interactive and Cooperative Delivery of Referring Expressions: A Comparison of Three Algorithms

Jana Götze **Karla Friedrichs** **David Schlangen**
Computational Linguistics, Department Linguistics, University of Potsdam, Germany
{jana.goetze,friedrichs,david.schlangen}@uni-potsdam.de

Abstract

In interaction, the establishment of reference is a collaborative process involving the main speaker and the addressee. Current work on visual natural language generation however minimizes interactivity and concentrates on the complexity of the input. Here, we return to some classical rule-based NLG algorithms, and extend them minimally to achieve incremental referring behavior guided by the listener's non-verbal feedback in a visual domain. We run a human evaluation study and show that these algorithms create behavior that is effective, though not judged as human-like. An additional, even simpler algorithm that generates finer-grained instructions is shown to be even more effective in ambiguous settings. We speculate that such simple algorithms can act as teachers that can help neural models take a step towards interactivity.

1 Introduction

In interactive settings, the establishment of reference to objects is a collaborative process, shaped by the referrer as well as the addressee. Even though this is by no means a new insight (e.g., Clark and Wilkes-Gibbs (1986); Heeman and Hirst (1995)), it is one that has moved outside the focus of much current work on visual natural language generation, which concentrates on the complexity of the input (e.g., raw image data instead of symbolic representations of the visual context) and minimizes interactivity, even if the chosen name of the task, e.g., Das et al. (2017)'s "visual dialog" or Savva et al. (2019)'s "embodied AI", might suggest otherwise (Benotti and Blackburn, 2021).

In task-oriented dialog, subdialogs emerge when an instruction follower (IF) asks for clarification in case they are unsure. Even when the IF does not interact verbally, the instruction giver (IG) collaboratively guides the IF after giving an initial instruction by iteratively providing feedback and additional information (Striegnitz et al., 2012). Figure 1 shows an example from a human-human data collection.

Figure 1: Example for a task-oriented interaction in shared visual space; cf. (Zarrieß and Schlangen, 2018).

Here, we investigate whether such non-verbal user behavior can be used in combination with classical rule-based Referring Expression Generation (REG) algorithms — the Incremental Algorithm of Dale and Reiter (1995); and Denis (2010)' Reference Domain Theory — for continuously providing feedback to the IF in an object identification task. We evaluate the resulting interactive algorithms in human evaluations, and show that they create behavior that leads to high task success. Humans evaluate none of the algorithms as human-like but accept them as reasonably likeable, friendly and competent. An additional, even simpler algorithm that generates finer-grained instructions is shown to be even more effective in ambiguous settings and is slightly favored overall by participants.

We close with speculations on how such rule-based systems that use symbolic input could be used as data generators for more flexible learning-based systems that combine robustness on the input side with more natural grounding behavior.

2 Related Work

Incorporating non-verbal listener feedback into REG systems has been the subject of previous studies. Especially eye gaze has been interesting to investigate in this context as studies in psycholinguistics have shown that listeners attend to objects

in their visual environment as they are being referred to (Tanenhaus et al., 1995).

In task-oriented interaction data in the context of the GIVE challenge (Byron et al., 2007), which provided a 3D environment in which instruction followers (IF) moved around, eye gaze has been found to be a good predictor of what object the listener resolved a referring expression (RE) to (Engonopoulos et al., 2013; Koleva et al., 2015; Staudte et al., 2012). Koller et al. (2012) have integrated eye gaze and movement information directly into their REG algorithm to produce positive and negative feedback and found that eye gaze improves referential success most but that also movement information was useful compared to giving no feedback at all. Both feedback systems reach high task success rates but required interaction data from humans for training. In our work, we want to investigate the suitability of existing rule-based algorithms that require no previous training data. Instead of eye gaze data, we rely on positional information of the listeners' movement which has also improved task success in Koller et al. (2012)'s experiments.

We will investigate two rule-based algorithms: the Incremental Algorithm (IA) (Dale and Reiter, 1995) and an algorithm based on Reference Domain Theory (RDT) (Denis, 2010). While the IA assumes the full context of objects to be available when generating a RE, RDT includes a notion of focus on subsets of objects and makes it suitable for environments in which the listener's view changes. We carry out our experiments in a 2D environment where the listener has access to the full set of objects and the IG – the REG algorithm – can "see" what the IF is doing.

Models that account for the explicit collaborativeness of reference have been proposed as well. For example, Heeman and Hirst (1995) use a planning-based approach that accounts for clarification requests as modifications of the plan and allows each partner to modify the plan – the referring expression – directly. We leave this extension to future work.

More recently, researchers have attempted to generate instructions and descriptions based on images, bypassing the need to create a symbolic representation of the domain, and thus being able to leverage the capabilities of modern neural network models (Das et al., 2017; Savva et al., 2019). However, these efforts do not account for the collaborative nature of reference even if the task names may suggest otherwise (Benotti and Blackburn, 2021). They instead separate the generation and evaluation of reference from one another without allowing for a collaborative modification of the generated RE. Instructions however need to go beyond correctness in that the description attempts to elicit the desired behavior in the listener. In order to obtain suitable data for training a neural network model, we therefore need to make sure that the input language data is both correct and suitable for the task. We investigate whether rule-based algorithms are a possible data generation mechanism by testing their generated output in human evaluation in a domain that we can access in both symbolic and continuous format.

3 Rule-based collaborative instructions

For this research, we adapt and extend the Incremental Algorithm (IA) (Dale and Reiter, 1995) and the Reference Domain Theory (RDT) (Denis, 2010), and set up an additional algorithm called Supervised Exploration (SE), which we explain in this section. All algorithms generate an initial referring expression based on the current visual context and then continuously monitor user behavior to provide continuous feedback to the IF. All three implementations are available at `https://github.com/kfriedrichs/golm/tree/ba`.

In order to select an object from the set, the instruction follower moves towards the object via a *gripper* – a cursor that can be controlled using the keyboard. This movement constitutes the user behavior that each algorithm bases its feedback on. After the initial RE, each system monitors gripper movement and generates either positive or negative feedback (YNFEEDBACK), or, when no movement has happened for a certain time, an adjusted RE (a new GENERATERE event) depending on the specific algorithm.

Each system monitors the gripper movement as well as the time to detect idle times. When the gripper has moved three grid units, YNFEEDBACK is generated based on the movement with respect to the target object. When the gripper has moved less than three units in 10 seconds or the participant has gripped an incorrect object, a new GENERATERE instruction is produced.

Algorithm 1 shows pseudocode for the general procedure that was used to instantiate each specific algorithm and the following feedback mechanism.

Algorithm 1 Event-driven feedback mechanism.
In the experiments, we set: Timeout=10sec, Threshold=3 units on grid, MaxTries=3

```
1:  procedure ON NEWTASKEVENT
2:      GENERATERE                                  // IA, RDT or SE
3:  end procedure
4:  procedure ON GRIPPERUPDATEEVENT
5:      if TargetGripped or MaxTriesReached then    // Skip to the next configuration when the
6:          NEWTASKEVENT                            ... correct object was picked or after 3 tries.
7:      else if IncorrectSelection then
8:          THATWASINCORRECT
9:          GENERATERE                              // repeat/rephrase
10:     else if MovedPastThreshold then             // When the gripper has moved in one direction
11:         YNFEEDBACK                              ... for a certain distance, give feedback.
12:     end if
13: end procedure
14: procedure ON TIMEOUTEVENT                       // If nothing has happened for too long
15:     if Gripper moved since Timeout/2 then       // If the gripper has moved recently
16:         YNFEEDBACK                              ... give feedback
17:     else
18:         GENERATERE                              // repeat/rephrase
19:     end if
20: end procedure
```

3.1 Incremental Algorithm

We implement the Incremental Algorithm (IA) as described in (Dale and Reiter, 1995) and extend it by the feedback loop as described above. The IA assumes a preference order of available properties that is known to influence the performance of the algorithm (van Deemter et al., 2012). We set the order to *color–shape–location* based on human RE from existing corpora in the same domain (Zarrieß et al., 2016). The IA will repeat its initial instruction in the case of a new GENERATERE decision.

The algorithm works as follows. It starts with all entities except the target as the *contrast set* and iterates through the given preference order of attributes. Each property that the current target has and that rules out some competing entity is immediately added to the RE, reflecting the *greedy* strategy. Ruled out entities are removed from the contrast set. The expression is complete and returned as soon as all distractors have been eliminated and the set is empty.

The YNFEEDBACK function is implemented here as a random selection from a fixed set; negative feedback is one of ["Not this direction", "Not there", "No"], positive feedback is one of ["Yes, this direction", "Yes", "Yeah", "Yes, this way"].

3.2 Reference Domain Theory

We implement a version of the algorithm based on Reference Domain Theory (RDT) as described in (Denis, 2010). RDT dynamically creates *reference domains* from the available object properties and accounts for discourse *salience* once a RE has been introduced as well as listener *focus* once the IF starts moving. This allows the algorithm to produce underspecified expressions like one-anaphora. We use the gripper movement to account for the listener's focus. The order of properties is set to be the same as for the Incremental Algorithm. Note that RDT uses an additional notion of location in its feedback generation: aside from the regular *location* property describing an object's global position, feedback may specify the position relative to the IF's gripper. Table 1 shows an example for how focus is used to dynamically generate underspecified RE with the RDT algorithm.

3.3 Supervised Exploration

The *Supervised Exploration Algorithm* (SE) generates instructions that are underspecified. It only verbalizes location information and then relies on guiding the IF using continuous feedback without verbalizing any further properties. Since this algorithm never produces a full RE, it continuously

Algorithm 2 Pseudocode of the feedback loop of the Supervised Exploration algorithm. The general feedback behavior in Algorithm 1 is extended by an additional check for whether the gripper is close to the target.

```
1:  if InXRange(TARGET) and InYRange(TARGET) then  // If the gripper is close to the target
2:      return TAKE(TARGET)                       ... issue an instruction
3:  else if MovingInRightDirection then           // If the gripper is moving
4:      return POSFEEDBACK                        ... issue positive and negative
5:  else if MovingInWrongDirection then           ... feedback based on its direction
6:      return NEGFEEDBACK
7:  else if IDLE and InXRange(TARGET) then        // If the gripper is still but close to
8:      return MOVE(y)                            ... the target on one axis, issue
9:  else if IDLE and InYRange(TARGET) then        ... an instruction for the other axis
10:     return MOVE(x)
11: end if
```

Task 2:	*There are multiple blue "W"s on the board. The leftmost is the target piece.*
Agent:	Select a blue W in the bottom.
User:	*moves the gripper towards the incorrect objects on the right*
Agent:	**Not these ones**. Get the blue W in the bottom of the board and left of the gripper.

Table 1: Possible interaction between the RDT agent and a user. The user's gripper movement is used to model their focus, enabling the algorithm to generate the **bold-faced underspecified instruction**. At the time of the second message, only non-target objects matching the initial ambiguous instruction are in the user's focus, therefore *"these ones"* suffices as a description of the negated objects.

checks whether the gripper has already reached its correct position along one of the axes to generate an additional STOP message.

The method is motivated by observations of human-human interactions that achieve object identification without the use of full REs. In some instances, the IF took a trial-and-error approach, continuously trying to guess the next referent or action and consequently receiving feedback from the IG. With this new algorithm, we explore a feedback-only reference strategy for an artificial instruction giver.

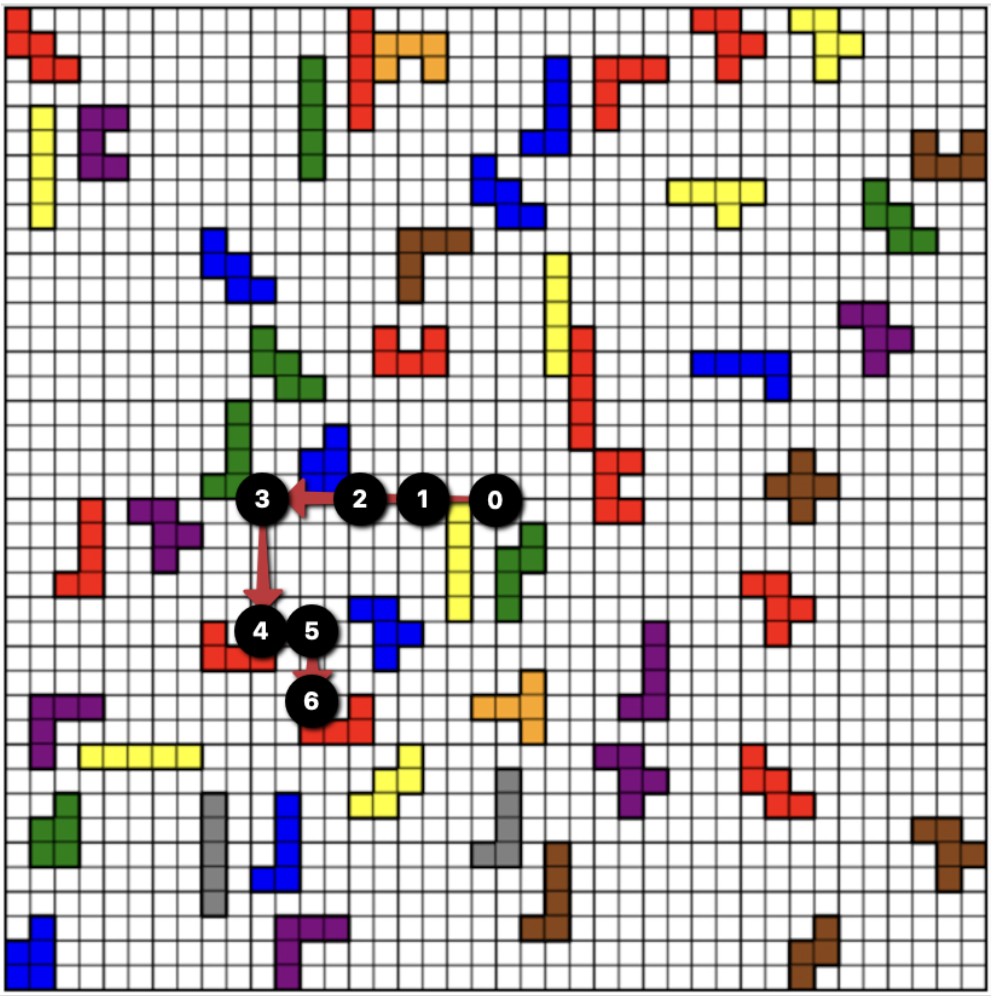
SE solely uses *location* as an attribute. Initially, it generates an instruction to move in one direction, starting with the x-axis. During the feedback loop, moving towards the target is supported by positive feedback, moving away is encountered by negative feedback, using the same fixed phrases as IA. Once one coordinate is in the target's range, *"Stop"* is output, followed either by a direction for the remaining axis or by the instruction to select the object. Since *stopping* the gripper is time-sensitive in order to not move past the object, SE uses an additional feedback trigger activated by the gripper entering the target range on one axis. Pseudocode for the algorithm's feedback loop is shown in Algorithm 2.

3.4 Example and comparison of the algorithms

Our domain is an online puzzle game in which Pentomino shapes have to be selected from a board of many pieces. Figure 2 shows an episode in which the three algorithms vary in their initial instructions.

The difference between IA and RDT is subtle and arises because the available attributes (shape, color and location) do not suffice for a discriminating description – two Pentomino pieces match the phrase "*red U in the bottom left*". Following (Dale and Reiter, 1995), this situation causes a failure of IA. In our experiment, we still used the final expression, which includes all features, for an instruction. Since IA assumes a discriminating RE, a definite article is used. RDT on the other hand acknowledges that the description is ambiguous by inserting the indefinite article to create an explicitly underspecified RE.

The feedback behavior of each algorithm is showcased in the same Figure 2. RDT clearly provides the most detailed feedback: at step 1, all algorithms reinforce the IF's moving direction, but RDT also provides a RE with the additional *location-relative-to-gripper* attribute. Still at 1, a



	IA	RDT	SE
0	— — — — — — *Initial state* — — — — — —		
	Take the red U in the bottom left.	Take a red U in the bottom left.	Go a bit left.
1	Yes.	Yeah. A red U in the bottom left of the board and below left of the gripper.	Yes.
	— — — — — — *No movement for 10 seconds.* — — — — — —		
	Take the red U in the bottom left.	Look for a red U in the bottom left.	Go a bit left.
2			Stop. Go a bit down.
3	Not there.	Not this one. Look for another one.	No. Go a bit right.
	— — — — — — *The user selects an incorrect object.* — — — — — —		
4	That was incorrect. Take the red U in the bottom left.	That was incorrect. Look for a red U in the bottom left.	That was incorrect. Go a bit right.
5			Stop. Go a bit down.
6			Take this object.
	— — — — — — *The user selects the correct target object.* — — — — — —		

Figure 2: Example episode including initial instructions (0) and the feedback (starting at 1) given by each algorithm. The gripper was moved along the arrows. At *1*, the gripper was halted until the feedback timeout triggered. At *4*, an incorrect object was selected. The SE algorithm continuously monitors whether the gripper gets close to the target to issue a STOP message.

new GENERATERE action features a full instruction for all algorithms by definition. Step 3 demonstrates RDT's strength: using one-anaphora, the algorithm acknowledges the presence of identical objects and tries to disambiguate them based on the IF's focus. At step 4, all algorithms output "That was incorrect" followed by a GENERATERE action as before.

The example also shows the increased feedback frequency of SE. At steps 2, 5, and 6, the gripper gets close to the target on at least one axis, triggering an instant *"Stop"* or *"Take this object"* response of the agent.

Note that there might be more feedback messages from each algorithm depending on the movement speed of the gripper, typically at least another YNFEEDBACK between 4 and 6.[1]

4 Experiments

4.1 Method and Procedure

We designed an interactive object identification task in which participants see a playing board online in their browser. The board contains 50 Pentomino puzzle pieces on a 40x40 grid. The pieces can have one of 12 different shapes and 8 different colors and can be rotated and mirrored. Figure 2 shows an example. We design 12 different episodes, one for each of the 12 different Pentomino shapes as target piece. Each participant sees all 12 episodes in the same order.

In order to select pieces, participants use the arrow keys on their keyboard to move a *gripper* depicted by a cross. The gripper is initially positioned in the center of the board for each new episode and can be moved in steps of 0.5 units of the visible grid. In order to select a piece when the gripper touches it, participants use their *space* or *enter* key.

We create 6 *hard* and 6 *easy* configurations. In the easy configurations, the target piece has at least one unique property. In the hard configurations, more than one piece will match the initial instruction, even when all attributes are specified. We achieve this by placing copies (or rotated copies, as rotation is not used as an attribute here) of the target next to the target piece (cf. Figure 2 for an example hard episode). We generate instructions in English, using each of the three algorithms described in Section 3. The instructions are generated offline for each configuration and synthesized using the Amazon Polly TTS standard Matthew voice.[2]

Each participant was randomly assigned one of the algorithms. The data collection starts with an audio test in which the participant is asked to transcribe a phrase that they hear in order to ensure that they could play audio in their browser. Participants are then presented a trial episode in which they could familiarize themselves with the interface.[3]

We log each gripper movement and instruction event in a json format. Timestamped logs are sent to our self-hosted server at the end of the interaction. We use the logged information to derive the following metrics:

Number of incorrect attempts for each episode. The maximum number of trials in each episode is 3, which the participants were informed about during the training episode. After the third trial, the participant sees the next configuration regardless of success. A value of 2 or fewer incorrect attempts reflects task success.

Time to solve an episode in seconds, starting at the end of the initial spoken instruction until the correct grip or third grip.

Number of feedback messages for each episode, i.e. how many times the algorithm verbally reacted to a participant's behavior.

Subjective ratings using 7-point Likert scales in a post-task questionnaire to measure participants' perception of the agent. Throughout the data collection, the voice was referred to as "Matthew" in order to give the agent an identity.

4.2 Results and Discussion

We collect a convenience sample as part of a student project, recruiting primarily university students via email. Participants were unaware of the specific research question. They did not receive reimbursement, but participated in order to support the project. Participation was anonymous.

91 subjects participated. Data from 1 participant was removed because they did not pass the audio test. Of the remaining 90 participants, 43 were female, 41 male, 2 non-binary and 4 did not report,

[1] Apart from the user's speed in moving the gripper, the movement speed also depends on the fire rate of keyboard events. Since the setting did not allow us to control the system setup of each participant, we acknowledge there might have been some variance.

[2] https://aws.amazon.com/polly/

[3] The data collection interface is available at https://github.com/clp-research/golmi.

Algorithm	Success Rate			# Failed attempts			# Feedbacks			Task length			# Episodes		
	all	easy	hard	all	easy	hard	all	easy	hard	all	easy	hard	all	easy	hard
IA	0.94	**0.96**	0.92	0.52	**0.19**	0.85	3.46	**2.56**	4.35	**11.59**	**10.07**	**13.10**	331	164	167
RDT	0.93	0.95	0.90	0.58	0.27	0.89	**3.07**	2.73	**3.41**	12.91	11.25	14.57	351	174	177
SE	**0.95**	0.95	**0.95**	**0.29***	0.33	**0.24***	5.93*	6.18*	5.68*	14.98*	15.58*	14.37	384	192	192

Table 2: Summary of results by episode type. The task ends when the correct piece is gripped or after 3 attempts. The task is successful if there were 2 or fewer failed attempts. Task length is reported in seconds. *indicates a statistically significant difference between the result for SE and both IA and RDT (ind.t-test, p<0.001).

Dimension	IA	RDT	SE
machine-like – human-like	2.43	2.67	**2.75**
incompetent – competent	4.29	3.40	**4.81**
dislike – like	4.25	3.73	**4.28**
unfriendly – friendly	4.46	4.40	**4.97**
unpleasant – pleasant	**4.29**	3.63	4.25

Table 3: Results from the post-task questionnaire. All scales ranged from 1 to 7 in the order of the specified adjectives. $N_{IA} = 28$, $N_{RDT} = 30$, $N_{SE} = 32$.

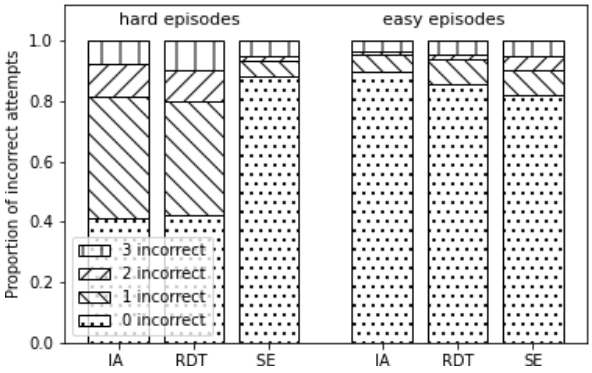


Figure 3: Incorrect choices made for each algorithm.

the mean age was 29.95 (5 did not report). 28 runs were collected for IA, 30 for RDT and 32 for SE. Each run consists of 12 episodes as described in the previous section. We removed single episodes from the data when the gripper stood still for 20 seconds or longer, assuming that the participant had abandoned it, resulting in a total of 331 episodes for IA, 351 episodes for RDT and 384 episodes for SE.

The results are summarized in Tables 2 and 3. All three algorithms achieve similarly high success rates overall as well as for the *easy* episodes. In the *hard* episodes, SE performs best. Participants interacting with IA were fastest in all settings and overall slowest with SE. Figure 3 additionally shows a more detailed visualization of failed attempts for each algorithm and setting.

The most striking differences between the three methods can be seen when looking at the number of failed attempts and the number of feedback messages participants received. Participants had a maximum of 3 attempts to identify the target object before they would see the next, unrelated episode. Overall, participants needed about half as many attempts with SE (0.29) compared with RDT (0.58) and clearly fewer than IA (0.52). However, this differed distinctly when separating easy and hard episodes. In easy episodes, participants with IA needed fewer attempts than both RDT and SE. In hard episodes, participants needed more than three times as many attempts with both IA and RDT as with SE. Unsurprisingly, participants received many more feedback messages with SE in all settings since SE evaluates the position of the gripper continuously rather than only after a certain distance (cf. Section 3.3).

Each participant interacted with only one algorithm, which makes direct comparison impossible. Instead, we asked for their subjective ratings as summarized in Table 3. All algorithms' output was rated as rather machine-like. Despite the underspecified instructions and high number of feedback messages, SE was rated as most competent, likeable, and friendly, with IA close in scores. None of the score averages surpass 5 on a scale up to 7, so a lot of room for improvement exists. Note that many factors can influence this rating, including the particular voice and feedback verbalization.

Based on these results, movement and timing information are appropriate indicators of the IF's reference resolution in this particular domain. Even without testing different timing settings, the success rate is high for all settings, showing that the instructions and feedback are interpretable. The results also give an indication on the level of performance we can expect in easy vs. hard settings and serve as a baseline for comparison when training a model using raw image data as input. The success rate when considering only the participants' first guesses differs greatly between these two settings for the IA and RDT algorithms; both achieve about twice the success rate in easy vs. hard configurations.

5 Conclusion

We have shown how rule-based REG algorithms can be enhanced with timing- and movement-based feedback to increase referential success, especially in ambiguous configurations, and without having to generate spatial relations between objects. The success rates give us a baseline for generating such RE based on raw image data, without access to absolute property values. As seen in Figure 3, these baselines differ for the three algorithms depending on the particular configuration of objects. For unambiguous settings, instructions given by all algorithms were picked on first try in most of the cases, while the success rate dropped visibly for ambiguous settings when IA and RDT gave an instruction. This is important for using these instructions as input for other learning mechanisms.

Our tool (mentioned in Section 4.1) lets us easily convert the symbolic visual game boards into images, making it suitable to compare the exact same settings with neural network models and generating the necessary amount of unbiased object configurations as training data. Instead of letting human annotators formulate instructions that potentially vary significantly in their verbalizations, we will use the rule-based algorithms to generate the training instructions we have tested with users in this paper.

6 Limitations

We acknowledge that the sample of participants is small and there is no guarantee that participants have focussed on the task at all times. We have removed outliers where the gripper stayed idle for a long time as explained in Section 4 but participants carried out the interaction in the environment of their choice rather than in the lab where they could have been supervised. Only 5% of participants reported to be native speakers of English. The remainder self-reported a mean fluency of 5.26 on a scale from 1 ("limited fluency") to 7 ("full fluency").

Acknowledgements

This work was partially funded by DFG project 423217434 ("recolage").

References

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a Collaborative Process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating Instructions in Virtual Environments (GIVE): A Challenge and an Evaluation Testbed for NLG. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

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

Robert Dale and Ehud Reiter. 1995. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19(2):233–263.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Alexandre Denis. 2010. Generating Referring Expressions with Reference Domain Theory. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Nikos Engonopoulos, Martin Villalba, Ivan Titov, and Alexander Koller. 2013. Predicting the Resolution of Referring Expressions from User Behavior. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1359. Association for Computational Linguistics.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on Referring Expressions. *Computational Linguistics*, 21(3):351–382.

Nikolina Koleva, Martin Villalba, Maria Staudte, and Alexander Koller. 2015. The Impact of Listener Gaze on Predicting Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 812–817, Beijing, China. Association for Computational Linguistics.

Alexander Koller, Maria Staudte, Konstantina Garoufi, and Matthew Crocker. 2012. Enhancing Referential Success by Tracking Hearer Gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pages 30–39, Seoul, South Korea. Association for Computational Linguistics.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019.

Habitat: A Platform for Embodied AI Research. *arXiv:1904.01201 [cs]*.

Maria Staudte, Alexander Koller, Konstantina Garoufi, and Matthew Crocker. 2012. Using listener gaze to augment speech generation in a virtual 3D environment. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34.

Kristina Striegnitz, Hendrik Buschmeier, and Stefan Kopp. 2012. Referring in Installments: A Corpus Study of Spoken Object References in an Interactive Virtual Environment. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 12–16, Utica, IL. Association for Computational Linguistics.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N.Y.)*, 268(5217):1632–1634.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012. Generation of Referring Expressions: Assessing the Incremental Algorithm. *Cognitive Science*, 36(5):799–836.

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

Sina Zarrieß and David Schlangen. 2018. Being data-driven is not enough: Revisiting interactive instruction giving as a challenge for NLG. In *Proceedings of the Workshop on NLG for Human–Robot Interaction*, pages 27–31, Tilburg, The Netherlands. Association for Computational Linguistics.

A Material

Figures 4 and 5 shows example episodes. Figure 6 shows the initial screen that participants saw when starting the data collection interface. Demographic questions in the post-task questionnaire were voluntary.

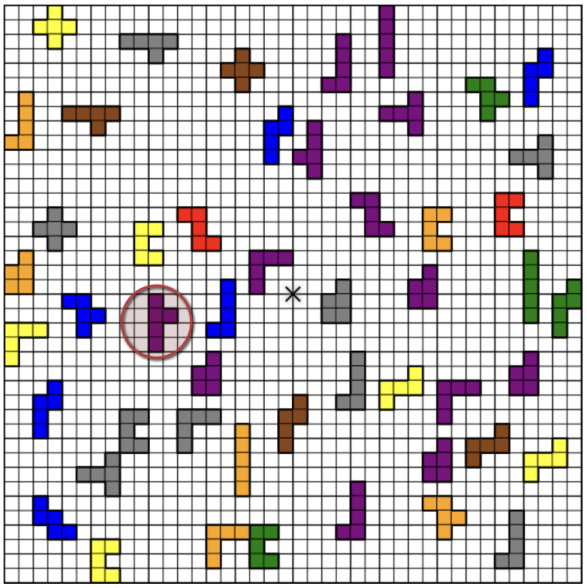


Figure 4: An example *easy* episode. The target object is circled, the *gripper* is positioned in the center.

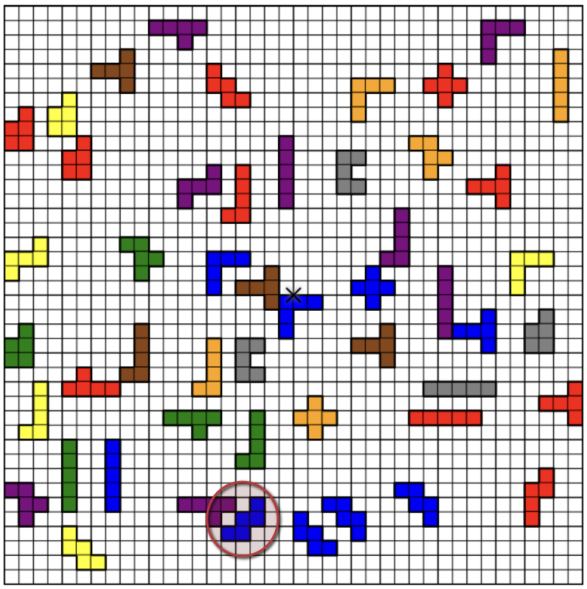


Figure 5: Example of a *hard* episode. The target object is circled, the *gripper* is positioned in the center.

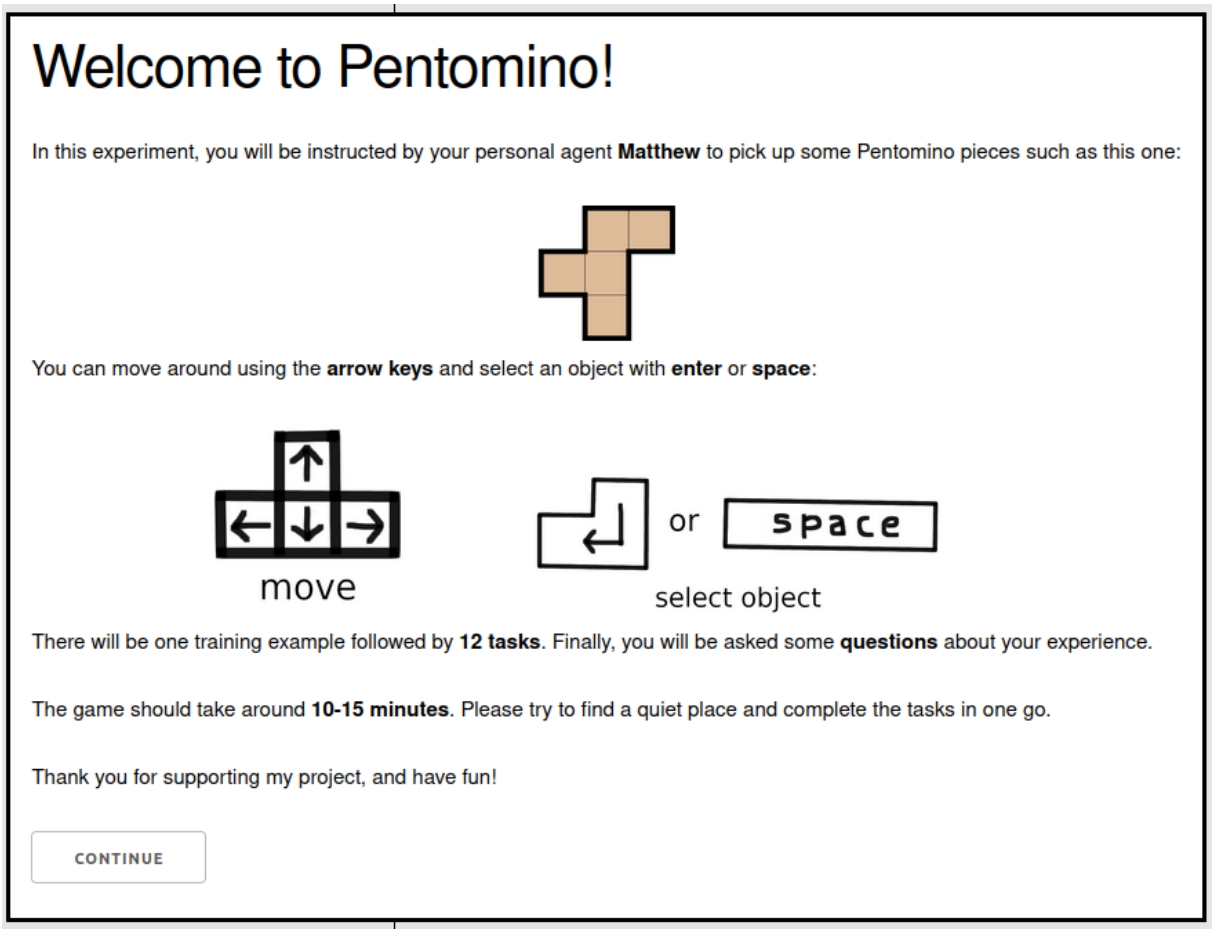


Figure 6: The welcome screen of the data collection.

Grounding Novel Utterances in Visual Dialogue

Mert İnan and **Malihe Alikhani**
Computer Science Department,
School of Computing and Information,
University of Pittsburgh, Pittsburgh, USA
{mert.inan, malihe}@pitt.edu

Abstract

Interlocutors use sufficiently salient yet creative and dynamic meaning pairs to communicate and coordinate in dialogue (Lewis, 2008). In this work, we focus on novel utterances in visual dialogue. We survey different types of lexical innovations discussed in the cognitive science and computer science literature and study how and when the transformer-based language models fail to probe context and process novel referring expressions. We annotate around 300 utterances that include novel utterances from the Photobook dataset (Haber et al., 2019) and present a data-driven study of lexical innovation and micro language in task-oriented dialogue. We then propose an algorithm that ranks the importance of the local context history according to the content of novel utterances. Based on this ranking, we create a model that can process and ground these novel utterances in context. We conclude with a discussion on how lexical innovations may change across conversations and how interlocutors can converge on shorter referring expressions about 52% of the time over the course of the interaction.

1 Introduction

Communication is inherently creative. Interlocutors produce utterances that include novel expression–meaning pairs to successfully communicate (Clark and Clark, 1979). Listeners understand these *lexical innovations* and uncover the intended meaning effortlessly. We build on Armstrong (2016)'s argument and present empirical evidence that shows that semantic conventions that influence language production in dialogue are dynamically determined by coordination between the engaged listener and speaker. These local conventions, which Clark (1998) refers to as *micro-languages* are suited for the needs of subgroups and may not be utilized by other subgroups or even the same speaker or listener in future interactions.

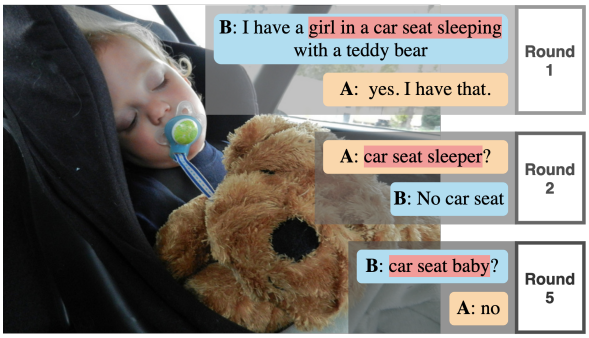


Figure 1: This is an example image from the Photobook Dataset, with its relevant dialogue history from different rounds of the game. The highlighted segments in the text correspond to the lexical innovations and their referents from the previous context. "car seat sleeper" is a novel utterance with noun-noun-noun compounding. Participants can converge to shorter lexical innovations over time. Converged lexical innovations can also change slightly, i.e, "car seat sleeper" to "car seat baby".

Krahmer and van Deemter (2012) discuss several ways in which the production of referring expressions can be creative and addressee–dependent i.e. the use of abbreviations and certain redundancies. In this work, we focus on visual dialogue and study ways that our cognitive capabilities and conventions might influence the production of lexical innovation. In particular, we evaluate the performance of language understanding models when conversations involve novel utterances in a visual dialogue task.

Resolving novel referring expressions in visual dialogue requires understanding the images and the context of the conversation. Agarwal et al. (2020) show that transformer models fail to understand and use the context in visual dialogue. They fail to generalize well in the presence of lexical innovations. We present a case study using the Photobook dataset (Haber et al., 2019) and examine the different types of lexical innovations that the dataset presents. Figure 1 presents an example conversation with lexical innovations and ways that the in-

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

terlocutors coordinate to understand each other and achieve a common ground. In what follows we (1) survey different definitions and classes of lexical innovations in the cognitive science and computer science literature; (2) present a data-driven analysis of lexical innovations in the Photobook dataset; (3) propose a novelty coefficient calculation algorithm which uses Part-of-Speech tagging to rank the novelty of utterances in a sentence; (4) develop a multimodal language understanding model that can detect and quantify novelties in the utterance, which then ranks local conversational context for better grounding of novelties; and, (5) show that lexical innovations are dynamic structures that are influenced by local conventions and may or may not be used in different dialogue rounds even by the same speakers.

2 Background and Definitions

Lexical Innovation Lexical innovation is different from conversational implicature (Bach, 1994), it is also not necessarily a deep metaphor. A deep metaphor is one, as defined in Armstrong (2016) and Lepore et al. (2010), where the meaning of the metaphor is linked to conventions that cannot be localized to that specific sentence. Furthermore, novel uses of utterances (most probably denominal verbs) that disrespect the semantic conventions are also not lexical innovations, such as "she foodified the ingredients", or "paper-outted the digital copy".

"Lexical innovation compositionally interacts with the other expressions in the sentences they occur", such as conditionals, negations and quantifiers. Armstrong (2016) argues that lexical innovation can happen in two ways:

- *Lexical expansion*: $L \rightarrow L'$ where L' is a lexicon with a new lexical expression that is not present in L.
- *Lexical shift*: $L \rightarrow L'$ where L' is a lexicon with a new lexical expression that is not present in L. Meaning of some expression in lexicon L' is different from that of in L.

In the fields of developmental psychology, and cognitive psychology, there are similar ways of defining strategies of lexical innovations. Clark (1980) and Bryant (2014) give the following strategies:

- Compounding: Two or more bases are combined and form a single word, for example, *bubble-hair* (a person with curly hair). Examples from Table 1 are *suit guy*, *truck thing*.
- Conversion: A word is adapted to a new word class without any changes. Examples from the Photobook dataset are *paving truck*, where paving is used as a noun instead of a verb or adjective.
- Affixation: A prefix or suffix is added to modify a base semantically or grammatically, for example, *sworder* (swordsman), *unfilled* (empty). Examples from the Photobook dataset are *pinkish*.
- Compounding with affixation: This is a subcategory of compounding and a combination of affixation. Examples from Table 1 are *the stripey cake*.
- Onomatopoeia: Words that sound like an action or an object. Examples from the Photobook dataset are *chuck*, *fluff*, *clip*, *pat*, *puff*.
- Child-talk: Such as that found in children's books (e.g., "Do you know what shlom is?"). Annotated section of the Photobook dataset does not contain child-talk due to its domain.

People are capable of producing novel utterances dynamically during a conversation. Yet, language understanding models fail to understand them (Testoni et al., 2022).

As an initial exploration, we focus on compounding and conversion—which are the major two categories of innovation that are automatically detectable—by using a detection algorithm that we propose. Using these strategies, we quantify the novel segments in grounded collaborative dialogues in a multimodal setting. We combine our understanding of lexical innovation from the already-present literature of linguistics, philosophy and cognitive psychology, and explore the Photobook dataset with the tools and understanding from these fields.

Collaborative Reference Grounding vague and ambiguous utterances have been addressed before in the context of grounding color terms. McMahan and Stone (2015); Monroe et al. (2017); Winn and Muresan (2018), and Fried et al. (2021) have all explored modelling color perception and comparative color descriptions using Bayesian models and reinforcement learning (Khalid et al., 2020a,b),

		
Utterance Chain	**Round 3:** **A:** my last one is the atari person with socks **Round 4:** **A:** okay, atari guy again **Round 5:** **A:** atari with socks	**Round 3:** **A:** I have the two men cutting a cake with red and white stripes **Round 4:** **B:** I have the two men with the stripey cake
Qualitative Observation	Across multiple rounds, utterance length becomes shorter, as the previous dialogue history context gets longer for both agents, the use of compounding to create novel segments increases. This example signifies looking up from previous history and increasing the attention to previous round utterances.	Across multiple rounds, different users can refer to the same object and only one of them may contain lexical innovation. While probing for context, both agents' previous rounds and previous turns should be used to find the necessary context for the specific novel segment.
		
Utterance Chain	**Round 1:** **B:** white guy with a orange vase looking at a truck thing **Round 3:** **B:** man in orange vase looking at a truck	**Round 3:** **B:** yes, I have the suit guy, on the bench again too **Round 4:** **B:** do you have the sloucher on the bench on his phone? **Round 5:** **A:** sloucher dude?
Qualitative Observation	Grammatical errors are not necessarily considered as lexical innovation. The lexical innovation detection module needs to be robust for these cases. Here, "vase" is not lexical innovation, but "truck thing" is.	The lexical innovation segment in the last round exists in the previous rounds as a part of the whole utterance. Hence, the segment needs to be extracted and then a coefficient needs to be calculated for the whole utterance to probe the previous context.

Table 1: This figure illustrates different dialogue examples based on the images above them. It also gives qualitative observations on how the lexical innovation plays a role in understanding certain segments of the utterances by the agents. Underlined portions denote the novel segments and their previous references. Most of them are noun-noun compounding lexical innovations.

whereas in our work we study Transformer-based models. While these papers are applied specifically to color terms and mainly work on resolving ambiguities, we are looking at grounding novel combinations of nouns that are not necessarily vague.

Resolving ambiguous novel utterances have also been studied in robotics and situated dialogue. It is still an open investigation area which has been mentioned in the recent survey for spoken interactions with robots by Marge et al. (2022). Liu et al. (2013) study novel referring expressions, where a graph mapping between a robot's visual context and the dialogue utterances is established for novel objects in the environment. In this line of research, a resolution of ambiguity of "novel" utterances have been addressed using cognitive processes. Our work is also inspired by the categories that cognitive scientists have proposed but we mainly focus on dynamically-formed novel utterances or micro language in visual dialogue.

Different corpora exist for the problem of visual collaborative reference: task-oriented visual dialogue such as *VisDial* (Das et al., 2017), *TalkTheWalk* where participants describe locations as they are walking, (de Vries et al., 2018), *MeetUp!* which is about dialogues that contain referring to locations and objects, (Ilinykh et al., 2019), *CoDraw* which has referring to objects and figures in drawings, (Kim et al., 2019), *Photobook* that has rich referring expressions to objects in a synchronous image matching game (Haber et al., 2019), *TEACh* where a commander directs a robot to complete tasks (Padmakumar et al., 2021), and SIMMC 2.0 (Kottur et al., 2021) where an agent resolves ambiguities when a human refers to objects in a shopping setting. While in all of these works there is

an exploration of resolving ambiguous referring utterances, none of the baseline models in these works address lexical innovations, and do not generalize well to out-of-domain corpora (Kim et al., 2020). Grounding and the problem of collaborative reference in dialogue is analyzed even more in the surveys by Schlangen (2019); Agarwal et al. (2020). Overall most models focus on a plethora of tasks and specific domains, but we are focusing on grounding creative utterances when people are referring to objects dynamically.

3 Data creation and annotation

In this work, we use the Photobook Task and its related datasets[1], which are components of a dialogue-based image-identification game (Haber et al., 2019).

In the original Photobook task, two participants are each shown 6 images selected from the MS COCO Dataset (Lin et al., 2014) on a randomized grid with some shared images. The primary task of the game is for each participant to select if any of the highlighted images is common or different by communicating with each other over a dialogue interface. The task is symmetric, as both participants can ask questions and provide answers. When the participants finalize a selection about the common or different images, then one round of the game ends, and another round begins with a newly randomized set of images. This new set may contain some of the same images from the previous rounds providing a history for participants to refer back to across rounds. A single game consists of five rounds, each of which contains three highlighted target images to label as common or different. This multi-round structure of the game allows an analysis of novel expressions that are getting created across different rounds by same or different participants, letting us observe the dynamics of lexical innovation. See Table 1 for a few examples.

Full dialogues of the Photobook dataset contain a total of 2,506 human–human conversations, and a total of 164,615 utterances. Because it is more straightforward to find novel utterances in the reference chains, we used that instead of the full dialogues. These chains are extracted from the full dialogues and for each MS COCO image in the game there is a chain. They are composed of multiple utterances taken from different rounds and different games referring to the same image. Each of these utterances contain a description about their corresponding image target from the dialogues. This Photobook utterance-based reference chain dataset is accessed through this link[2]. The total number of utterance chains is 16,525, which contain a total of 41,340 referring utterances. These are split into train, validation and test sets originally in the data with 11540, 2503, and 2482 utterances in each split, respectively.

As shown in Table 1 we observe various novel referring expressions such as "atari guy", and "the stripey cake" in the utterance chains. We formalize different classes of these type of novel referring expressions in Section 2, then annotate a portion of the utterance chains by identifying novel utterance segments and their classes. Then we use the Part-of-Speech tag patterns to detect these lexical innovations in Section 3.1.

3.1 Lexical Innovation Statistics in the Photobook Dataset

We observe that lexical innovation happens following semi-structured patterns of part-of-speech for the compounding and non-structured patterns for conversion classes. These patterns are as follows: for compounding, multiple `NOUN` classes are used consecutively; for conversion, an `ADJ` class or a `VERB` class is used in front of multiple consecutive `NOUN` classes. This is an empirical observation made on the available data, and it is assumed that these patterns are generalizable across datasets from different domains.

Counts for lexical innovation that we have identified in the Photobook dataset are presented in Table 2. This table shows multiple characteristics of the Photobook dataset in terms of lexical innovation. It shows that the most common way of creating novel words is by compounding. All the noun compoundings are the most common among all lexical innovation types. `ADJ-NOUN` and `VERB-NOUN` compoundings are assumed to be corresponding to the *conversion* type of lexical innovation.

We annotate a small subset of the training data (277 samples) with the lexical innovation types, by two human experts. In this data, we identify the presence of the segment inside the utterance that the lexical innovation is corresponding to. To calculate the Cohen's κ inter-rater agreement, we

[1] https://dmg-photobook.github.io/

[2] https://github.com/dmg-photobook/ref-gen-photobook/blob/main/dataset/v2.zip?raw=true

Lexical Innovation Type	Train	Test	Annotated
2-noun	4708	950	78
3-noun	1072	200	32
4-noun	202	37	2
5-noun	47	9	0
6-noun	99	29	0
adj-noun-noun	2981	662	46
verb-noun-noun	4471	121	14
onomatopoeia	1849	-	62
child-talk	-	-	0
affixation	-	-	0
Total count	34903	7450	277

Table 2: Numbers of different compounding types from the Photobook utterance chains. Annotated set is from the training set of the corpus. Dashes mean that POS-tag rules were not found to detect lexical innovation automatically in the data.

select 30 utterances randomly and assign them to two annotators. The Kappa coefficient is κ = 0.76 which indicates a substantial agreement (Viera et al., 2005).

After this annotation is complete, we run the automatic POS-tagging on this small subset. Here we observe that majority of the lexical innovations exist within the noun-noun compounding type. We observe that the distribution of the 2-noun, 3-noun, 4-noun, 5-noun and 6-noun compoundings follow a similar pattern for the train, test and the annotated subset. This shows that our POS-tagging strategy is a fast and feasible approximation of detecting lexical innovations similar to human annotations. We detect the onomatopoeia using a dictionary extracted from the Oxford English Dictionary by (Sugahara, 2011). After identifying these POS statistics, we try to find a way to quantify the novelty of these specific lexical innovation segments.

4 Model

Here we describe a listener model for collaborative reference grounding in the presence of novel utterances (see Figure 2). The inputs to our model are six images, one current utterance, and a history of reference-chain utterances referring to each of the six images, while the output is a single image chosen out of the six images. We measure our task success using accuracy and mean reciprocal rank (MRR) measures for image retrieval. We also present an algorithm for lexical innovation detection and coefficient calculation. Our code is publicly available[3].

Our model contains a modified listener module of the Reference Resolution Model as proposed by (Takmaz et al., 2020) [4]. In the original model, when the hypothesis utterance, u_t is received by the listener, BERT embeddings, $BERT(u_t)$, are extracted for each utterance using uncased base BERT (Devlin et al., 2019; Wolf et al., 2020) and they are concatenated with ResNet-152 embeddings, $RESNET(I_i)$ (He et al., 2016) of each of the images for multimodal representation. In our model, when the hypothesis utterance is received by the listener, we first identify whether there is a probable lexical innovation in the utterance. If there is no lexical innovation, then we run the original listener model. If there is, then we use a separate mechanism to rank previous rounds' utterances and increase the visibility of utterances that have less novel segments to the model.

When a lexical innovation is detected in the utterance, u_t, that refers to an image, i, then the model first fetches all the utterances that refer to i from the same game but previous rounds, which can be represented by $[u_{t-1}, u_{t-2}, ..., u_{t-k}]$, where t represents the current round number, and k is the number of maximum possible history of rounds for that specific image, i. Then we use our lexical novelty coefficient calculation algorithm to measure how novel each utterance in the history, u_{t-m}, is, where m is an arbitrary number less than k.

The novelty coefficient calculation is given in Algorithm 1. Given an utterance, we first run a POS-tagger on each word of the utterance, then find the segment of the utterance where it has POS tags corresponding to the segment of lexical innovation. We query Google n-gram Book database (Breder Birkenes et al., 2015; Lin et al., 2012) for that segment concatenated with its POS-tags (i.e. `umbrella_NOUN cat_NOUN lady_NOUN`) with insensitive case-matching and zero smoothing, resulting in a ratio, in the range of $(0, 1)$. If there was no lexical innovation segment found then we assign it a value of 1. We then multiply the ratio with the number of total entries in the Google Books n-gram database (around 10^{14} entries) to get an estimate count of the occur-

[3] https://github.com/Merterm/lexical-innovation

[4] model code is retrieved through: https://github.com/dmg-photobook/ref-gen-photobook/tree/main/models/listener

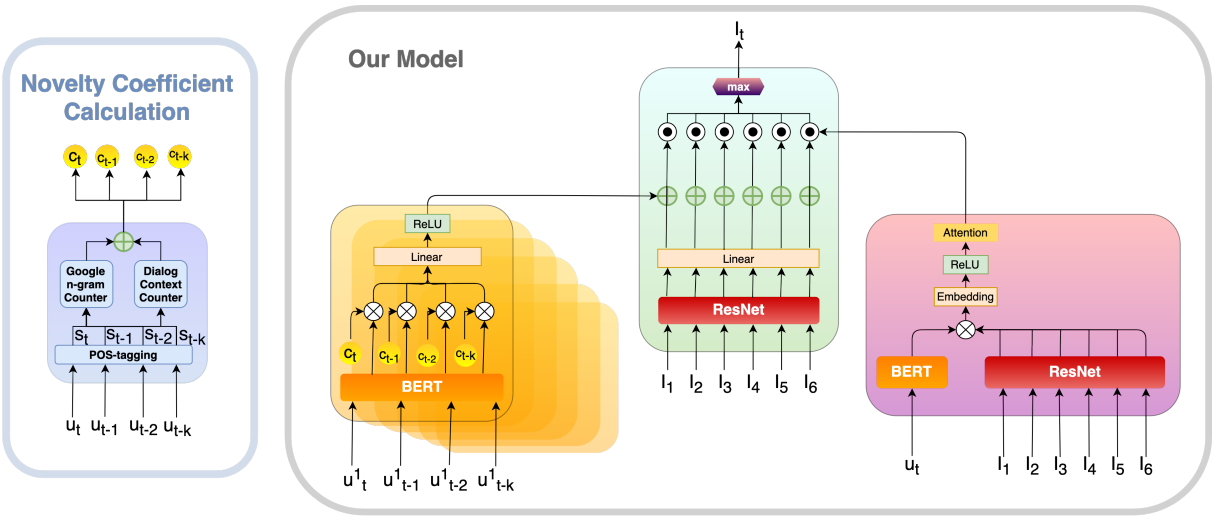


Figure 2: This is a simplified depiction of our multimodal language understanding model. Novelty weighing mechanism changes the weight of the history of utterances based on their lexical novelty during the dot product multiplication with each of the images. On the left, we have a lexical innovation coefficient calculator, which takes the dialogue history, $u^i_{t...t-k}$, and the current utterance, u_t, as input. It outputs coefficients, $c_{t...t-k}$, that are used in weighing the multimodal context in the model. In our model, there are 6 history modules per image in orange color, one utterance module in pink color, and one image picker module in green color. Inputs to the model are the utterance, u_t, history of utterances, $u^i_{t...t-k}$, for each image i, and six images $I_{1...6}$. The output is the chosen image, I_t, out of the 6 images. Intuitively, our lexical innovation module probes for the least innovative context and increases the weight of it in companion with the novel utterance.

rences of the segment in the English book corpora. This represents our game-independent external context coefficient with a range of $(0, 10^{14})$. Then we calculate the number of total occurrences of the segment in the given game and round, which represents our game-only local context coefficient. We finally calculate the novelty coefficient by adding both external and local context. The higher the coefficient, the less novel the utterance is.

Algorithm 1: Lexical Novelty Coefficient Calculation

Data: u_i, belonging to game g_i and round r_i
LI-RULES $\leftarrow$ {2−NOUN, 3−NOUN, 4−NOUN, 5−NOUN, 6−NOUN, ADJ−NOUN−NOUN, VERB−NOUN−NOUN}
for *each word,* w_i*, in* u_i **do**
  $tagged[w_i] \leftarrow$ pos_tagger(w_i);
end
if ***any*** *LI-RULES* ***in*** *tagged[i...i+n]* **then**
  $segment \leftarrow w_{i...i+n}$
end
$r_g \leftarrow 1$;
$q \leftarrow$ query_ngram($tagged$, $segment$);
if $q > 0$ **then**
  $r_g \leftarrow q$
end
$c_E \leftarrow r_g \times 10^{14}$;
$c_L \leftarrow$ count($segment$, g_i, r_i);
return $c \leftarrow c_E + c_L$;

In the original model, to pick one image out of 6 candidate images, a dot product is taken between the multimodal representation of u_t with the multimodal representation of each image $I_{1...6}$. The multimodal representation of u_t is the concatenation of ResNet features of $I_1...I_6$ with $BERT(u_t)$. The multimodal representation of each image, $I_i, i = 1..6$, is the summation of average of the history of the utterances, $BERT(u^i_{t-1}) + ... + BERT(u^i_{t-k})/k$, with $RESNET(I_i)$.

In our model, in order to pick one image out of 6 candidate images, we take a dot product between the multimodal representation of u_t with the multimodal representation of each image $I_{1...6}$, weighed by the lexical novelty coefficient. The multimodal representation of u_t is again the concatenation of $RESNET(I_1)...RESNET(I_6)$ with $BERT(u_t)$. However, in our case, the novelty-weighed multimodal representation of each image is the summation of the novelty-weighed history of the utterances, $c^i_{t-1} * BERT(u^i_{t-1}) + ... + c^i_{t-1} * BERT(u^i_{t-k})$, with $RESNET(I_i)$, where c^i_{t-1} represents the lexical novelty coefficient corresponding to u^i_{t-1}. Our model is depicted in Figure 2.

The main motivation for coefficient multiplication is to weigh the representations of different utterances from different rounds of the game. For

instance, if u_3^2–which is the utterance corresponding to round 3 of image 2–is "I have the two men cutting a cake with red and white stripe", and u_4^2 is "I have the two men with the stripey cake", then $c_3^2 > c_4^2$ as the u_3^2 does not contain any lexical novelty. We are giving highest weight to the least novel utterance because it is assumed that the least novel utterance representation is already grounded by the model compared to the most novel, which can guide the dot product towards grounding the novel utterance, as well.

	Full Photobook				**Only Novel**	
	Train		Test		Test	
	ACC	MRR	ACC	MRR	ACC	MRR
ReRef	95.2	97.3	85.3	91.2	82.5	89.5
Ours	97.6	98.7	85.4	92.1	**85.3**	**91.1**

Table 3: This table shows the performance of the Re-Ref model and our model on the train and test sets of the Full Photobook Corpus and Lexical Innovation-Only dataset. ACC corresponds to Accuracy and MRR corresponds to mean reciprocal rank. Re-Ref model performs worse on the lexical innovation extracted subset of the data than the full data. Our model improves on this giving more weight to less innovative utterances from the history of the conversation.

5 Results

We show that our model that is aware of lexical innovation improves on the accuracy and mean reciprocal rank (MRR) in the image retrieval task of choosing the target image from 6 candidate images (see Table 3).

Here we compare our model to the Re-Ref model introduced by (Takmaz et al., 2020). They show that their model performs with 85.32% and 91.20% accuracy in the test set of the full Photobook corpus. But we identify that their model's performance is slightly worse for the specific subset of lexical-innovation-only samples. As explained in Section 3, we select the samples using the automatic POS-tagger algorithm which contain segments that have lexical innovations in them according to our definition in Section 2. We show that ReRef model has an accuracy of 82.46% and an MRR of 89.49%, which are 3% and 2% less than the full dataset results, respectively.

Our listener model improves on the training data with around 2% in accuracy and MRR compared to the ReRef Baseline. More so, our model is able to bring up the test results for the lexical-innovation-only subset of the corpus to the full corpus performance levels. It improves the results by 2.8% for the accuracy, and 1.6% for the MRR compared to the ReRef baseline. In order to further investigate the performance of our model and investigate the dynamics of lexical innovation, we present qualitative and quantitative analyses in the following subsections.

5.1 Qualitative Error Analysis

In this section, the authors of the paper qualitatively observe the outputs of the novelty calculation. We see that the majority of the time, lexical innovation coefficient calculation successfully detects the novel utterances even in complex cases of 6-noun compounding. It is also able to detect non-novel utterances majority of the time as well.

We give more specific analysis of different types of qualitative phenomena we observe in Table 4. We can also see where the coefficient calculator does not perform as expected. For instance, one can observe that even though the sentences contain novel segments, the POS-tagging may select the non-novel segment such as "black bowl" instead of "orangy food", resulting in a false segmentation but correct coefficient calculation. This is still valuable for the listener model because the coefficient corresponds to all of the utterance instead of just the segment.

In certain cases, not novel segments can falsely get low coefficients (i.e. very novel), such as "hot dogs". This may be because Google n-gram database does not contain daily dialogues, and words that are not novel in daily communication may be absent on a book dataset, giving it a high coefficient even though it is not externally novel.

5.2 Do Lexical Innovations Change Across Games and Participants?

Lexical innovations can dynamically change during different rounds in a single game, during different games, and across different participants. Based on these three levels we ask three questions: how do novel words get modified across different rounds, how do they change across different games without considering rounds, and how do they change across participants regardless of the games? According to Armstrong (2016), lexical innovations exist dynamically, hence it can be hypothesized that after the game is over or even across different rounds, lexical innovation segments may get altered. To test this

	Utterance	Segment	Novelty Coefficient
Novel & Mis-segmented	do you have black bowl with orangy food, bowl with white rice, 3 part tray with food?	black bowl	1
	do you have a salad in a white bowl; salad looks like twigs with a red thing at the top...	white bowl	1
Novel & Correctly Segmented	green leafy salad with maybe red or orange item at top?	green leafy salad	1
	halloween cat?	halloween cat	1
	yes, pink rice, cat, tree, moon.	pink rice	1
	i have the red orange one	red orange	1
Not Novel with Low Coefficient	do you have a photo of fries and 3 hot dogs?	hot dogs	1
	black cat?	black cat	1
Novel with Medium Coefficient	do you have a dish on a square plate that has broccoli and white fluffy stuff ?	white fluffy stuff	4.03×10^6
	do you have broccoli with the white stuff again	white stuff	6.61×10^8
Not Novel with High Coefficient	salad with glass of grape juice or wine	-	10^{14}
	i have a picture with fries and three subs	-	10^{14}
Both Novel & Non-Novel with Different Scores	bowl of red vegetable next to loaf of bread on kitchen table?	-	10^{14}
	bowl of red veg next to loaf of bread	red veg	1
	a lunch box with 4 different colored comparents	colored comparents	1
	i have the lunchbox with the four compartments	-	10^{14}
Typo	largew hite square plate, with broccoli and rice etc	largew hite square	1

Table 4: This table shows different utterance examples and how the lexical innovation calculator module scores them for error analysis purposes. There are several classes of scoring and utterance pairs. First rows show differences in segmentation performance and how it affects the scoring. Next rows show how the novelty affects the score and finally an example with a typo is given. Here, higher score means less novel, as the novelty coefficient corresponds to a count of the word in the Google n-gram database and the previous dialogue context. Minimum score is 1, and the maximum score is 10^{14}.

hypothesis, we both qualitatively and quantitatively analyze the data. We list the lexical innovation segments that are found in our annotated data, then we cross-check the exact segment in our full dataset of utterances.

We find that lexical innovations re-occur in other games 22.2% of the time (267 different game re-occurrences out of 1203 lexical innovation re-occurences in the annotated dataset). This shows that same lexical innovation can be used multiple times across rounds and games. In Table 5, we present lexical innovation segment examples to observe their dynamic behavior across rounds and games. For instance, "white lap" re-occurs in different rounds of game number 744, 10.6% times out of all its re-occurrences. This shows that across different rounds, participants come back to the exact same lexical innovation segment. This is statistically significant with p = 0.0008 and t = 9.1259. We measure the significance using one sample t-test between the hypothetical uniform distribution mean of 1.52 of and the actual distributions across the games.

Lexical innovations from the annotated set re-occur 13.9% (167 same photo re-occurrences out of all lexical innovations re-occurrences) times when the picture is the same. Hence, different participants looking at the same picture can come up with the same lexical innovation even across different games. As an example, if we look at the same-photo re-occurrence probability of "choc cake", we see that 50% of its re-occurrence happens in games with the same photo, but with different participants.

On the contrary, participants can also converge to different lexical innovations when the game changes or after different rounds. For instance, in game 1140, participants can converge to "wii lap showing feet guy", then converge to "point of view wii remote" in another game. In another game, in round 3 participants converge to "feet up gaming" while in round 5 they re-converge to "close up wii remote guy". This shows that the durability of

Lexical Innovation	Game ID	In-Game Re-Occurrence Probability	Same-Photo Re-Occurrence Probability
choc cake	702	0.333‡	0.167‡
	635	0.111‡	0.278‡
	1900	0.167‡	0.500‡
	1903	0.111‡	
white lap	2433	0.091†	0.409‡
	1716	0.076†	0.136‡
	1346	0.061†	0.061‡
	2484	0.091†	0.242‡
	744	0.106†	0.061‡
salvation army truck	1502	0.081†	0.973
	1520	0.054†	
	1799	0.081†	
	2092	0.081†	
weird looking 5 wheeled black bike	1339	1.000	1.000
wii lap showing feet guy	1140	1.000	1.000

Table 5: This table shows the number of re-occurrences of some lexical innovation examples that were identified during annotation. In-game re-occurrence probability is the count of lexical innovation in the game with the given ID, divided by the number of total re-occurrences in all the annotated data. Same-photo re-occurrence probability is the count of the lexical innovation segment referring to the same photo divided by the number of total re-occurrences in all the annotated data. (†: statistically significant results with the power of $p \leq 0.001$, ‡: significant results with the power of $p \leq 0.1$)

novel utterances is dynamic, as some lexical innovations are easy for people to converge to and stay attached to even across rounds and games while some lexical innovations can dynamically vanish once the image or game is gone.

5.3 Do Participants Converge to Shorter Lexical Innovations?

People tend to converge to lexical innovations over the course of the dialogue in two different ways: either long and complex compoundings, or short and simple compoundings. Here we explore how these are distributed in our data. We observe that complex or lengthy lexical innovations that are 4 to 6-noun compounding do not re-occur in the data at all. These type of complex lexical innovations happen 46.9% of the time (130 out of 277 utterances) in the annotated dataset. We explain this phenomenon further with examples from Table 5. "weird looking 5 wheeled black bike" has an in-game re-occurrence probability of 1, which means that it only occurs in game 1339 once and never again in the data. This is because it is long and specific. For "wii lap showing feet guy", the participants converge to that lexical innovation in round 5 of the game, but it never exists in any other game. This shows the ephemeral nature of long and specific lexical innovations.

We observe another phenomenon in which participants converge to simpler and shorter lexical innovations as they continue to future rounds. In the annotated dialogues with lexical innovation, 51.8% (28 games out of 54) of the games converge from more than 5-token description of the object to 2 or 3-noun compounding lexical innovation after multiple rounds. This shows that participants converge to shorter lexical innovations as an establishment of common ground.

6 Discussion & Conclusion

We introduce a language understanding model that is able to probe both previous dialogue context and the external context for grounding novel utterances. The proposed model performs better particulary on the subset of the data that includes lexical innovations. Due to the nature of the task, users tend to come up with similar "novel" segments. Hence in the end, task-specific models which just memorize the vocabulary can perform just as well as a lexical-innovation-aware model. Also, as is shown in Table 2, it is difficult to find POS rules for lexical innovations, and some lexical innovation types such as child-talk do not exist in our chosen multimodal dataset, which requires further data exploration. Exploring other multimodal dialogue corpora is left for future work.

7 Acknowledgements

We thank Katherine Atwell, Anthony Sicilia, Ece Takmaz, Vishakh Padmakumar, Joshua-Christian Wyatt and Sabit Hassan and the anonymous reviewers for the helpful comments and discussions.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Josh Armstrong. 2016. The problem of lexical innovation. *Linguistics and Philosophy*.

Kent Bach. 1994. Conversational impliciture. *Mind & Language*, 9(2):124–162.

Magnus Breder Birkenes, Lars G. Johnsen, Arne Martinus Lindstad, and Johanne Ostad. 2015. From digital library to n-grams: NB n-gram. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 293–295, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Judith Becker Bryant. 2014. Encyclopedia of language development. Lexical Innovations.

Eve V. Clark. 1980. Lexical innovations: How children learn to create new words. papers and reports on child language development, number 18. *null*.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*.

Herbert H Clark. 1998. 4 communal lexicons.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Fried, Justin Chiu, and Dan Klein. 2021. Reference-centric models for grounded collaborative dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment.

Baber Khalid, Malihe Alikhani, Michael Fellner, Brian McMahan, and Matthew Stone. 2020a. Discourse coherence, reference grounding and goal oriented dialogue. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Baber Khalid, Malihe Alikhani, and Matthew Stone. 2020b. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4417–4428, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hyounghun Kim, Hao Tan, and Mohit Bansal. 2020. Modality-balanced models for visual dialogue.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Ernest Lepore, Ernest Lepore, and Matthew Stone. 2010. Against metaphorical meaning. *Topoi-an International Review of Philosophy*.

David Lewis. 2008. *Convention: A philosophical study*. John Wiley & Sons.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco:

Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. 2013. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France. Association for Computational Linguistics.

Matthew Marge, Carol Y. Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gilmer L. Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, Mary P. Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David R. Traum, and Zhou Yu. 2022. Spoken language interaction with robots: Recommendations for future research. *Comput. Speech Lang.*, 71:101255.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*.

Will S. Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Teach: Task-driven embodied agents that chat.

David Schlangen. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings.

Takashi Sugahara. 2011. Onomatopoeia in spoken and written english: Corpus- and usage-based analysis.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Alberto Testoni, Claudio Greco, and Raffaella Bernardi. 2022. Artificial intelligence models do not ground negation, humans do. guesswhat?! dialogues as a case study. *Frontiers in Big Data*, 4.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

Olivia Winn and Smaranda Muresan. 2018. 'lighter' can still be dark: Modeling comparative color descriptions. *ACL*.

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

What to refer to and when? Reference and re-reference in two language-and-vision tasks

Simon Dobnik and **Nikolai Ilinykh** and **Aram Karimi**
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
name.surname@gu.se

Abstract

How do we refer to scene entities in interactive language-and-vision tasks? We explore reference and re-reference in two tasks, link them to a model of attention and discuss our findings in relation to modelling situated interaction.

1 Introduction

In this paper we examine how conversational partners refer to scene entities in two language-and-vision tasks. Knowing the strategies and models of referring is crucial for natural language processing tasks of situated interaction, both interpretation and generation of referring expression. In natural language generation, the step is crucial for content selection (Deemter, 2016): a visual scene may include several entities, their features and spatial relations between them but only some are selected and included in the expression to be generated. In natural language understanding, referring expressions have to be resolved to scene entities, their attributes and spatial relations between them: similarly, referring expressions are ambiguous and they may be resolved to several potential candidates. In situated interaction involving several conversational partners several aspects of referring are relevant (Byron, 2003). Firstly, elements in the scenes are described to and referred to in a particular order which is reflected in the discourse model (Grosz and Sidner, 1986; Ilinykh and Dobnik, 2020; Takmaz et al., 2020). The same discourse elements may be re-referred during the discourse which is described by co-reference (Stede, 2011; Poesio et al., 2018; Loáiciga et al., 2021). When referring to discourse entities conversational participants may also take different spatial perspectives (Maillat, 2003). Our hope is that this investigation will shed light on strategies that need to be taken into consideration in modelling situated discourse. This is particularly relevant for multi-modal neural networks as understanding the properties of visual interaction will help us to evaluate and study these models for such properties (Ilinykh and Dobnik, 2022).

The mechanisms driving linguistic reference, connecting words with the physical properties of the scene, are driven by the notion of *attention*. Attention can be of two different kinds: linguistic and perceptual (visual) attention. Objects attain linguistic salience (i) if they have been mentioned in the conversation before, and (ii) depending on how thematically they are relevant to the topic of conversation and the task that the participants are engaged in. Objects attain visual salience by attention on the visual properties of the scene such as colour, size, shape and geometric arrangement. In resolving the reference of objects both kind of attention interact. Furthermore, in dynamic environments as the conversation progresses the attention on objects changes based on object visibility and recency of it being added to the common ground (for discussion see (Kelleher and Dobnik, 2020)). In this paper we examine attention on objects by inspecting how they are referred to in two different tasks using two corpora: the Cups corpus (Dobnik et al., 2020) and the Tell-me-more corpus (Ilinykh et al., 2019).

2 Tasks and corpora

The Cups corpus contains longer English and Swedish dialogues where participants have to identify missing cups on a large table that are hidden to them but these are visible to their conversational partner and vice versa. The cups differ in features such as type, colour and location. Participants are located at the opposite sides of the table and they see each other as an avatar. Figure 1 shows a top-down view of the scene. Each participant sees the same table scene from their own point of view as shown in Figure 7 in Appendix. In addition, there is also a passive observer Katie on the side of the table. Participants are instructed to interact over a chat interface to find the cups each is missing. Beyond this information they are not specifically

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

told how they should approach the task, the aim is that they negotiate the strategies through their linguistic interaction and engage in a longer dialogue. Table 5 in Appendix shows the overall coverage of the dialogues. The data has been annotated to study different conversational phenomena including spatial perspective taking (Dobnik et al., 2020), dialogue games (Storckenfeldt, 2018; Dobnik and Storckenfeldt, 2018) and reference and coreference (Dobnik and Loáiciga, 2019; Silfversparre, 2021; Dobnik and Silfversparre, 2021). The results reported in this paper are based on these annotations.[1]

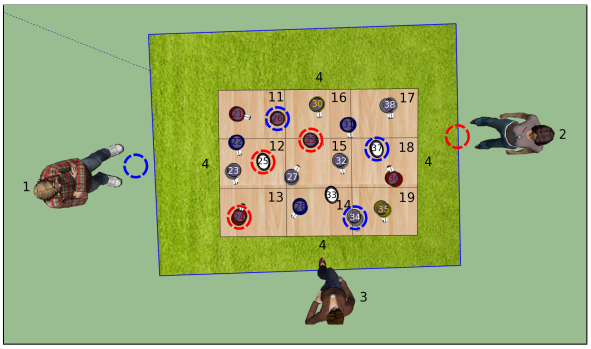


Figure 1: A top-down view of the scene with all objects included and their IDs. Objects marked with coloured circles cannot be seen by a participant marked with the same colour. P3 is a passive observer Katie.

The Tell-me-more corpus (Ilinykh et al., 2019) contains descriptions of images of house environments where participants (via crowd-sourcing) were encouraged to provide multi-sentence descriptions of them. The task can be considered as a simplified form of dialogue with fixed conversational roles of participants. It involves incremental updates of scene descriptions from a describer to an imaginary interactive partner requesting additional information over five turns. The goal of the task is to study incremental referring which is reflected in the discourse structure of the generated text.

We choose these datasets because they provide different scenarios for the study of attention patterns being relevant for the resolution of reference. The Cups scene is known and is identical for all the dialogues. It contains objects of restricted kind, namely the cups, but these vary in terms of their properties such as colour and location. This allows us to study referring over longer sequences of dialogue as well as how participants visually segment larger scenes into smaller regions and how such structuring of a task is reflected in their interaction. Both participants are human, they each have the same goal and by default they do not have pre-determined roles. Instead, these are negotiated between them as the conversation unfolds so that they both can complete the task. The Tell-me-more images are are real-world images different for each discourse where the view of the scene has been determined by the author of the photo. The conversational roles and the view are fixed and consequently interactions are short. However, in this fixed view a variety of scene entities are available that can be potentially referred to. Therefore, the Tell-me-more corpus allows us to study reference and re-reference at a thematic and scene-topological level whereas Cups allows us to study them at the interaction level. Since each involves a different task, a comparison of referring also sheds light on the effect of the task on referring.

While reference in Cups was annotated by human annotators, for Tell-me more we perform this by automatic linking of noun phrases from sequences of image descriptions to object descriptions detected by an object detector. We extract noun phrases from image descriptions using SpaCy (Honnibal et al., 2020). If the head of a noun phrase is not a noun, we consider it an incorrect detection and remove it. We also create a list of words describing types of rooms (e.g., "bedroom", "'attic") based on the (Zhou et al., 2017) hierarchy of images. Overall, we extract 51,953 noun phrases with 9.11 noun phrases per image description and 15,507 noun phrases describing rooms with 2.72 phrases per image description. For object detection we use the model by Anderson et al. (2018).[2] This takes an image and produces a list of detected objects with bounding boxes and object descriptions. The latter include labels (e.g., "chair") and their attributes (e.g., "black"). We limited the number of extracted objects per image to 36.

We explore different methods for linking noun phrases from textual descriptions and object descriptions of detected objects. (Ilinykh and Dobnik, 2022) demonstrate that a transformer-based model and cosine similarity between two phrases (Reimers and Gurevych, 2019) with a threshold 0.5 gives the best performance. For plural noun phrases we follow (Ilinykh et al., 2019) by taking their singular form and link them to objects that have the

[1] https://github.com/sdobnik/cups-corpus

[2] https://github.com/peteanderson80/bottom-up-attention

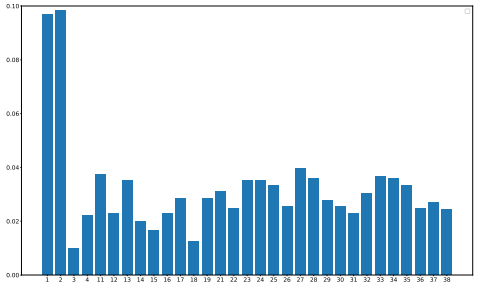
most similar word as head in their description.

We extend the linking method by controlling detection of objects based on the confidence scores of the predicted object labels (L) and attributes (A). We consider three conditions: (i) **AL**, (ii) **[A]L** and (iii) **[A][L]** where brackets indicate that the inclusion of a label or an attribute is conditioned by a confidence score threshold. The thresholds we use are approximations from (Anderson et al., 2018) and were 0.4 for attributes and 0.1 for labels. We evaluate each method manually, by randomly sampling 10 image-text pairs from the dataset and inspecting the correctness of the linking against the expected links, annotated by one of the authors. In 10 image-text pairs there were 102 noun phrases on which each method performed similarly, with the number of incorrect links not exceeding 30. Specifically, **[A][L]** made the fewest errors (25), while **AN** and **[A]N** followed with 28 and 30 errors respectively. One explanation why controlling for both attributes and labels performs best is that it filters out detections with low confidence scores and decreases hallucinations based on textual predictions. As objects with missing labels are removed, it also removes duplicate bounding boxes with low confidence scores. As linking is a highly complex semantic task, no doubt more work is required to improve and evaluate different methods.

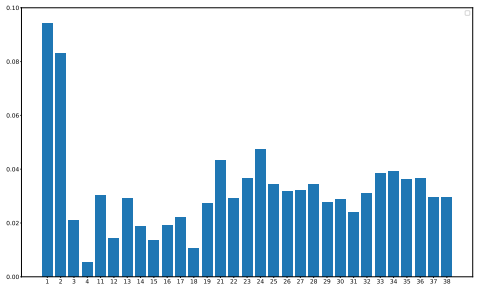
3 Reference in the Cups corpus

Table 1 shows reference to objects in both English and Swedish dialogues. The column *Objects* indicates the number of references to individual objects including both pre-annotated objects (see Figure 1) and objects identified by annotators while the column *Our objects* contains references to pre-annotated objects. Objects identified by annotators include references to parts of objects (e.g. handles of cups and lids) and regions that are created based on the topological arrangement of objects (rather than our pre-annotated regions) (Dobnik and Silfversparre, 2021). In the remainder of the paper we will be referring to these objects. The column *Referring expressions* lists the number of referring expressions in each dialogue. The counts in this column are lower than the counts from the previous column which means that referring expressions contain reference to more than one object, on average two objects per referring expression. However, when one examines the dialogues it can be seen that there is a considerable variation between the number of objects between referring expressions. Some are referring to uniquely identified objects while others are referring to groups of varies sizes. Since dialogues are of different lengths we normalise all three columns to average values per turns. There are differences between individual dialogues but no differences between English and Swedish dialogues. Overall, there are between 3 and 7 objects referred to per turn, when we exclude annotator created objects, between 2 and 5. There are between 1.5 and 3 referring expressions per turn.

3.1 Objects referred to



(a) English dialogues



(b) Swedish dialogues

Figure 2: Reference to entities for English and Swedish dialogues: 1–3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. To allow comparison all counts are normalised to the total number of references per language corpus, i.e. the columns sum to 1. See also Figure 1 for the representation of the scene objects.

Overall, there are differences in referring to objects between languages. A χ^2 test of independence found a significant relationship between language and reference to scene entities: χ^2(df=30, N=4344)=60.5756, p=0.0008.

Participants (1, 2) most frequently refer to themselves. In the English dialogues the reference to both participants is nearly equal but in the Swedish dialogues participant 1 is more frequently referred to than participant 2. Katie (3), a passive observer

Dialogue	*Length in turns*	*Objects referred to*	*per turn*	*Our objects referred to*	*per turn*	*Referring expressions*	*per turn*
en-1	157	530	3.376	478	3.045	282	1.783
en-2	441	1316	2.984	968	2.195	683	1.549
sv-1	118	407	3.445	261	2.212	177	1.5
sv-2	114	613	5.377	480	4.211	314	2.754
sv-4	75	513	6.84	369	4.92	251	3.347
sv-5	163	628	3.853	473	2.90	334	2.05
sv-6	248	786	3.17	604	2.435	408	1.645
sv-7	308	922	2.994	711	2.309	469	1.523

Table 1: Objects referred and the number of referring expressions in the Cups dialogues.

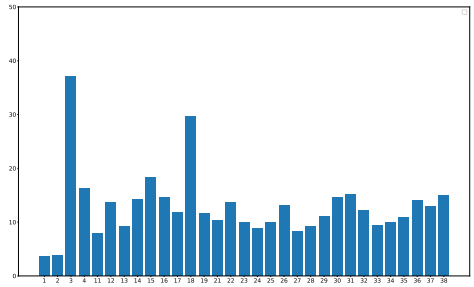
is rarely referred to, even less than objects or regions. This indicates the effect of the task on referring. Participants have a central role in the task (they have to find the missing cups each) as well as they are coordinating the task and the dialogue. Objects and regions are a part of the task. Katie, although animate and therefore potentially a salient landmark, is only a passive observer in this case and does not contribute to the task. The table (4) is more frequently referred to in the English than Swedish dialogues but overall it is among less frequently referred to entities, possibly serving as a landmark in descriptions involving top-view allocentric frame of reference. The next type of entities ranked by the increasing frequency are regions (11–19). Here we see that in both groups of dialogues regions 11, 13, 17 and 19 are most frequently referred to whereas region 18 is the least frequently referred to region. Figure 1 shows that these are the corner regions of the table, hence regions of the table that are closest to each participant and on their left and right. Regions that are between these regions receive less attention, most notably region 18 which is the central region closer to P2. Overall, objects are even more frequently referred to than regions. Here there is a slight difference between languages for example some most frequently objects referred to in English are 27, 28, 33, 34, 23, 24 and for Swedish 24, 21, 33, 34, 35, 36, 28. Examining the scene in Figure 1 we can see that are related to the missing cups 24, 25, 26, 29, 34, 37 either because they are the missing cup (e.g. 24, 34), they are a distractor object for the missing cup (i.e. the cup that could be referred to with the same description as the missing cup, e.g. 21 for 26 or 24, and 33 and 35 for 34.) For example, 28, on the other hand, is a cup proximal to two missing cups and therefore a good landmark to refer to to resolve the task. A considerable part of the dialogue involves resolving reference of these descriptions and there are sections of dialogue where a describer and interpreter (who later also becomes a describer) refer to different entities with the same description until a contradiction is detected and diverged commons grounds are reconciled (for example, en-1 turns 42–62). Object 31 is the least frequently referred to in dialogues of both languages. This is a blue cup close to the missing red cup 29 and the white cup 37. As such it is not a distractor object to either of them and therefore likely be used only as a landmark for reference to other cups in configurations where other landmarks are also possible candidates (the same holds for object 22). Overall, the results indicate that the task and the way the scene was structured through the introduction of the missing cups has an effect on the attention the objects receive through reference.

The proportion of objects therefore tells us about their perceptual and task related salience. Frequently referred to objects are those that are related to the task but also those that are perceptually salient either because they are visually similar to the target objects or because they are good landmarks that target objects can be described with, for example the corner regions of the table. It is also observable that both properties interact. For example, visually accessible regions on the lateral dimension of the scene are more perceptually accessible to participants than the front and back regions and therefore they are more frequently referred to.

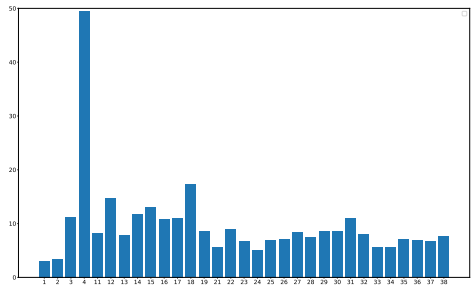
	en-1	en-2	sv-1	sv-2	sv-4	sv-5	sv-6	sv-7
en-1		ns	***	***	ns	ns	ns	**
en-2	ns		***	***	*	*	ns	***
sv-1	***	***		***	***	***	***	***
sv-2	***	***	***		ns	ns	ns	***
sv-4	ns	*	***	ns		ns	**	*
sv-5	ns	*	***	ns	ns		ns	***
sv-6	ns	ns	***	ns	**	ns		***
sv-7	**	***	***	***	*	***	***	

Table 2: χ^2 test of independence comparing reference to scene entities across dialogues. *** indicates $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ and ns indicates non-significant difference. For details see Table 6 in Appendix.

What is the variation of reference to objects between different conversational participant pairs in both languages? Table 2 shows the results of a χ^2 test of whether reference to scene entities differs between different dialogues. For English we only have two dialogues, en-1 and en2, for which the test indicates that they do not differ in reference to objects. The en-1 dialogue is more similar to the Swedish dialogues than en-2 which is an interesting observation since the speakers of en-1 are native speakers of Swedish (see Table 5 in Appendix). This suggest that there may be differences between languages in terms of referring. Among the Swedish dialogues sv-1 and sv-7 stand out as the reference there differs to reference in all other dialogues. The reference in the remaining Swedish dialogues, sv-2, sv-4, sv-5 and sv-6 is similar, except for dialogues sv-4 and sv-6 where reference is different but only when compared with each other. The results point that although different pairs of conversational participants structure the task freely and sometimes very differently, objects are still referred to in the same way. There is also an effect of language and possibly the way conversational participants approached to solve the task and their individual preferences.



(a) English dialogues



(b) Swedish dialogues

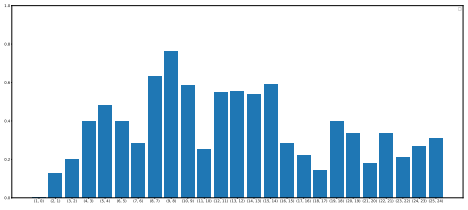
Figure 3: Mean distance between turns that repeat reference to entity for English and Swedish dialogues: 1–3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. See also Figure 1.

3.2 Re-reference to objects

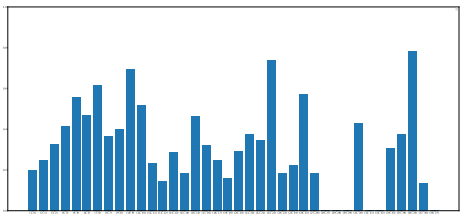
How likely is that an object will be re-referred in subsequent turns? Figure 3 shows for each scene entity the average distance (separation) between turns when this entity is re-referred in both English and Swedish dialogues. We estimate distance between each consecutive pair of turns when a particular entity has been referred to. Re-reference shows similar trends for both English and Swedish dialogues for individual scene entities. However, overall, the distance between turns over which they are re-referred is slightly greater in English than Swedish. This excludes object 4, the table. As expected, large distance of re-reference is associated with low frequency. Participants 1 and 2 are re-referred most recently but also most frequently (see Figure 2). Similarly, objects 24, 27, 28, 34, 34 for English and 24, 21, 34, 34 for Swedish. Katie (3) and table (4) are re-referred to a greater number of turns apart but also very infrequently. Similarly, objects 18, 15, 30, 31 for English and 18, 12, 15, 31 for Swedish. Overall, regions are re-referred after a greater number of turns than objects in both English and Swedish dialogues. Regions 11 and 13 are the most recently re-referred regions both in English and Swedish which is again associated with their high frequency. That regions are re-referred after greater number of turns than objects again confirms that they serve as landmarks for identifying objects when needed while objects are the main targets of descriptions identified by the task.

A non-uniform distribution in which objects and regions are re-referred indicates that these are not referred to randomly as the dialogue progresses. Work on dialogue interaction (Clark, 1996) and as well as previous work on the Cups conversations indicate that participants split the task, the scene and therefore conversations into sub-parts. The Swedish dialogues have been annotated for dialogue games (Kowtko et al., 1992; Carletta et al., 1997) with two kinds of tags, one indicating the scope of the games over turns and one indicating the type of the games (Storckenfeldt, 2018). Dialogue games can be nested, a typical example being a clarification game which is embedded in another game. In the next experiment we measure to what degree objects referred to in one dialogue game overlap with the objects referred to in other

dialogue games. As a measure of overlap we use Sørensen–Dice coefficient $DSC = \frac{2|A \cap B|}{|A|+|B|}$ which ranges between 0 (no overlap) and 1 (perfect overlap). Note that here we calculate overlap of sets which means that duplicate reference is counted only once. As individual conversations structured differently in terms of dialogues games and strategies to refer to objects we represent these for each dialogue game separately. When comparing pairs of adjacent games for objects they are referring to we obtain the mean values of DSC and their standard deviations as follows: sv-1: μ=60.7 σ=24.1, sv-2: μ=36.2 σ=18.6, sv-4: μ=31.7 σ=21.1, sv-5: μ=31 σ=21.2, sv-6: μ=37 σ=25.1 and sv-7: μ=29.5 σ=23.9. The results indicate that except for sv-1 where there is a high overlap of objects referred to across adjacent games (60.7), adjacent dialogue games overlap in reference in about a 1/3 according to DSC. However, notably there is a high standard deviation which indicates a high variability between individual games.



(a) sv-2



(b) sv-5

Figure 4: Dice-Sørensen coefficient of reference overlap between adjacent dialogue games.

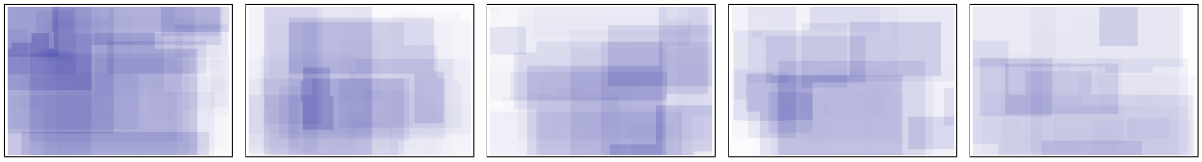
Figure 4 shows a comparison of two dialogues from which it can be seen that referring overlap is non-uniform and there are sections of dialogue where there is either a high or a low degree of reference overlap between adjacent games. In Figure 8 in Appendix we cross-tabulate DSC for all games (i.e. not just adjacent ones). The plots indicate that reference is local and restricted to a couple of adjacent dialogue games but a reference to the same objects might be made at a later stages of dialogue, again with a local scope. Overall, this indicates that reference to objects is highly dependent on how conversational partners negotiate and structure their task. Conversationally, structuring a large scene into local sub-parts has a referring advantage as expressions can be made more optimal and be less ambiguous (for example, by requiring less descriptive attributes) as attention is placed on a smaller number of distractor objects that are potential referents.

4 Reference in the Tell-me-more corpus

4.1 The location of objects referred to

Where (Landau and Jackendoff, 1993; Landau, 2016) in the image frame are these objects located? We track attention to objects in images by representing the overlap of the bounding boxes of objects referred to in each of the five (5) sentences that constitute a single image description. To demonstrate the effects of the discourse we represent attention maps collectively for all images for the first, second, ..., fifth sentence of the discourse. First, we take all images and re-scale them to $T \times T$ pixels, where $T = 2000$. Along with the images, we also resize bounding boxes accordingly to ensure that they correspond to the detected objects in size and location. For each sentence in a sequence, we draw a heat-map from bounding boxes of those objects that are mentioned in that sentence. In order to generate a single heat-map per sentence across all images, we use alpha blending (Blinn, 1994), a method that takes an image and maps another image on top of it. The mapping is controlled by two α values which determine the transparency of each of the two images. We set them to 0.9 and 0.1 for the background and foreground mapped images respectively. We also normalise the resulting heat-maps by the number of images in the dataset.

Figure 5 shows five attention heat-maps with darker areas indicating attention to objects being referred to. In general, the first sentence refers to the most of the scene, there is a high overlap between object bounding boxes, changing their attention on specific parts and objects later in the sequence. This finding shows the sequential nature of image description sequences and aligns with the idea that humans structure scene discourse and mention objects in some order (Grosz and Sidner, 1986). Note that there is also an impact of the type of the visual



(a) Sentence 1 (b) Sentence 2 (c) Sentence 3 (d) Sentence 4 (e) Sentence 5

Figure 5: Attention maps of bounding boxes of objects referred to determined through automatic linking. The results are averaged per sentences and across all images and discourses.

scenes as humans tend to describe apartment layouts through the "tour strategy", organising spatial descriptions of the house environments in a tree structure (Linde and Goguen, 1980).

We note that among sentences 2, 3 and 4 the attention shifts from one side of the image to another on the later dimension. This could be related to the fact that left-right is a prominent relation used in spatial inference along which target and landmark objects are related. In the last sentence the attention is generally weaker, indicating much fewer and smaller objects described. The number of objects linked on average per image is 3.94, 2.38, 2.02, 1.79, 1.60 for sentence from 1 to 5 respectively, showing that humans start with detailed descriptions of images and later focus on smaller parts of the scene, describing fewer objects. Overall, the results indicate that the attention on the image changes over the discourse: from several larger objects to fewer and smaller objects. There is also evidence of spatial inference in the left-right axis.

4.2 Thematic associations between objects

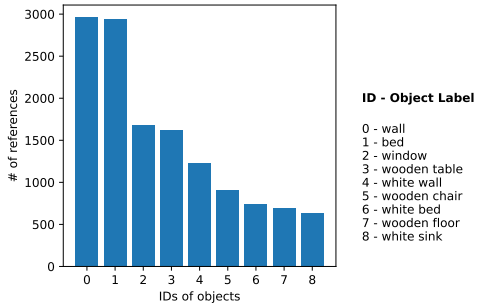


Figure 6: The frequency of top-*k* referred to objects across all images as determined by automatic linking.

What are the most frequently described objects across all image description sequences in the Tell-me-more corpus? Figure 6 shows the top frequencies of objects referred to as determined by automatic linking. These objects are typically objects that occur across many different room types: walls, chairs, floors, tables. Among 1,686 described object types, 560 of them (approximately 33%) are described only once. These least frequently described objects include, for example, "gray skirt", "orange stripe" and "beige light switch".

To investigate how thematic relations between objects drive attention as expressed through reference we examine and compare (i) how objects appear in situational contexts and (ii) how these objects are referred to in descriptive discourses. For this we build *vector space models* (Turney et al., 2010) of object co-occurrences where context counts are either determined by (i) scene co-occurrence and (ii) scene and referential co-occurrence. The *scene vector space* captures the number of times each object appears in the scene with other objects. We consider objects which have acceptable confidence scores for both attributes and labels as determined by the **[A][N]** linking method. Table 8 in Appendix shows a part of the scene co-occurrence vector space. The size of this space is $3,054 \times 3,054$. Note that this is greater than the number of objects labels in the model $(1,600)$, because object descriptions also contain attributes which introduces permutations of descriptions. The *referential vector space* captures for each object the frequency of it being mentioned together with other referentially linked objects. Table 7 in Appendix shows an excerpt from the constructed referential vector space. The size of the space is 1686×1686 which is smaller than the size of the scene vector space. This indicates that a large number (44.8%) of scene objects is not mentioned at all (including false negatives of our automatic linking method).

What are the differences between the two vector spaces? To examine the relationship between frequencies of objects in two vector spaces we compute a Spearman's rank correlation coefficient. We observe a positive correlation between the two variables, $r = 0.82$, $p = .000$, demonstrating that the frequencies of objects in images correlate with the frequencies of them being mentioned in image de-

scriptions, subject to the accuracy of linking. This is of no surprise since if an object is in an image, it has a certain probability to be mentioned in the image description.

For each vector space we rank the objects by their frequency of occurrence and then extract their ten most similar objects using the kd-tree (Maneewongvatana and Mount, 1999) which is an improved version of the k-nearest neighbour algorithm. Figure 9 and Figure 10 in Appendix show the most frequently and the least frequently occuring objects in both vector spaces and 10 of their most similar objects. The results indicate that the most frequently occuring objects are similar for both configurations. They include objects most commonly found in rooms such as wall, window, table and chair. However, there is a difference in what their most similar objects are. It appears that the similarity of objects from the referential vector space is based on the attributes and not just object co-occurrence, e.g. wooden table: wooden floor, white wall: white lamp, white window. This indicates that semantic distinctions captured are not only based on situational co-existence but other dimensions of meaning defined by the attribute: i.e. objects of the same colour or consisting of the same material (cf. the semantic distinction between sense and reference). For the least frequently occurring objects there is a high variation both in terms of what these objects are and their most similar objects in the corresponding vector spaces. This is expected because of their low frequency support.

Table 3 shows three objects and their most similar objects in both vector spaces. The referential vector space captures also *thematic* relations between objects: "stainless steel oven" is similar to "blender" and "silver coffee maker" which fit into a thematic cluster of kitchen appliances. On the other hand, the scene vector space captures similarities of co-occuring objects: it predicts "brown pot" and "white floor" similar to "stainless steel oven". Referential vector space therefore also encodes information about how humans group objects in scenes and describe them within a depiction of same event or a task. Other words show similar trends: "marble counter" is similar to bowls, knobs, food, bananas and hair dryers indicating other objects that interact with marble counters. On the other hand shelves, windows, refrigerators and ceiling predicted by the scene vector space are co-occuring objects in the same rooms. This shows that the task and subsequent human communicative intents are important factors of what gets included in a description: objects are not only described because they are there, but because they are thematically connected with each other at a higher task-related level.

4.3 Attention to objects through reference

Can we estimate this thematic attention for the objects referred to in the Tell-me more dataset? From objects appearing in a scene, what objects (i) are likely to be referred to, (ii) are likely to be re-referred in the same discourse, and (iii) are likely not to be referred to? To answer these questions, for each object w_n we compute *attention* as a ratio A_{w_n} between its vector in the reference vector space $\mathbf{V}^r$ and the scene vector space $\mathbf{V}^s$:

$$A_{w_n} = \frac{\sum \mathbf{v}^r_{w_n}}{\sum \mathbf{v}^s_{w_n}}, \quad (1)$$

where $\mathbf{v}^*_{w_n}$ is a word frequency vector in the corresponding vector space. An attention score 1 indicates that an object is referred every time when it occurs in an image. An attention score > 1 indicates that an object is likely to be re-referred in the same discourse and an attention score < 1 indicates that an object is referred to less frequently than it occurs. Attention scores close to 0 indicate that objects are nearly never referred to. Therefore, the resulting attention scores can be interpreted as *thematic salience* of objects in this domain.

Table 4 shows some of the most and the least attended objects in this corpus. First we note that 1,368 out of 3,054 objects are assigned an attention score 0.0 because they are never referred to (subject to the automatic linking method). Object names of the most attended objects often include attributes which refer to colour (e.g., "green stripe", "white artwork"). For example, "painted wall" is likely to be referred to (attention score 2.375) but "wall" has a score of 0.210848. While "black horse" is highly attended (ranked 13 among 1,686 objects), "black faucet" is ranked 1,672. This could be an artefact of using phrase similarity to match descriptions with object names containing attributes. It could be that the colour of the faucets is less likely to be described than the colour of horses and therefore an object label "black faucet" is less likely to be matched with a description "faucet". Similarly, "orange flower" and "white freezer" are unlikely to be referred to with these attributes while "blue flowers" are more likely with an attention score of

stainless steel oven		**silver refrigerator**		**marble counter**	
Ref space	*Scene space*	*Ref Space*	*Scene space*	*Ref Space*	*Scene space*
stainless steel oven	stainless steel oven	silver refrigerator	silver refrigerator	marble counter	marble counter
backsplash	white stove	silver microwave	white rug	black handle	mantle
tiled backsplash	red fruit	stainless steel dishwasher	marble counter	white bowl	shelf
white backsplash	yellow bottle	silver stove	clear wine glass	white hair dryer	brown window
white blender	brown pot	stainless steel oven	bowl	white knob	silver refrigerator
metal hood	white floor	stainless steel refrigerator	brown floor	green bananas	hanging chandelier
silver coffee maker	white table	black microwave	round table	gray towel	ceiling
pink bottle	pink cushion	oven	black printer	food	wooden floor
clear wine glass	white windows	brown cabinets	mantle	black light	rug
silver dishes	hanging light	food	shelf	stainless steel stove	black table
white lights	wooden wall	black oven	wooden chair	cabinets	white lamp

Table 3: The most similar objects for three target objects in referential and scene vector spaces. Objects are ordered from most (top) to least similar (bottom).

	Object	Attention score
0	green stripe	3.428571
1	white artwork	3.375000
2	red comforter	2.888889
3	decorative painting	2.823529
4	colorful couch	2.500000
5	painted wall	2.375000
6	white chicken	2.187500
7	seat	2.166667
8	yellow game	2.000000
9	black barrel	1.993243
10	pink sink	1.928571
11	silver drawers	1.800000
12	black horse	1.722222
13	gold headboard	1.684211
14	gold ceiling	1.666667
15	brown horse	1.664000
16	purple table	1.652174
17	leather recliner	1.642857
18	black machine	1.421687
19	white clothes	1.411765
...	...	...

	Object	Attention score
...	...	...
1666	hanging chain	0.015748
1667	red room	0.015504
1668	red shelf	0.014925
1669	white freezer	0.014787
1670	red rack	0.014706
1671	orange flower	0.014184
1672	white cup	0.013514
1673	yellow bottle	0.012121
1674	wooden entertainment center	0.011792
1675	black tire	0.011164
1676	red door	0.010870
1677	pot	0.010063
1678	black faucet	0.009740
1679	handle	0.008611
1680	outlet	0.007282
1681	yellow bowl	0.007067
1682	vent	0.006589
1683	parked car	0.006494
1684	metal pole	0.005952
1685	silver shower head	0.004907

Table 4: Attention scores for twenty most attended (left) and least attended objects (right).

0.524193. The attribute salience described here is common-sense thematic salience which is different from visual salience (Kelleher et al., 2005). It is important to note that both kinds of salience interact. For example, a "white freezer" is more likely to be referred to in the context of all black freezers.

5 Discussion and conclusion

Our comparison of reference in the Cups and Tell-me-more corpora reveals several factors that affect what objects are referred to and when. Referring is highly influenced by the nature of the conversational tasks which shapes the goals of participants and is reflected in conversational interaction. Participants in the Cups dialogues have identical conversational roles and are free to structure their interactions. On the other hand the task of referring in the Tell-me-more corpus and the roles of participants is highly restricted but so are the patterns of reference produced. Furthermore, we can observe differences in referring to scene entities as the discourse progresses. Therefore, it is *wrong* to assume that Tell-me-more and image captioning in general represent a task-neutral setting. Previously, referring expressions have been studied only within a particular corpus or a task but our findings indicate that this is by no means sufficient to understand referring. Further examination of the task structure which is reflected in discourse, for example in conversational games, might point to common referring patterns between tasks and make the notion of the task less elusive. We have also identified other factors relevant for referring: visual properties of the scene, geometric arrangements of scene objects and patterns of spatial reasoning. There are thematic relations between objects that go beyond the presence of objects in the scene and are related to description of coherent events.

Referring is a complex phenomenon that is hard for computational modelling. As it is context and task dependent this means that large corpora will have to be available to capture all the tasks, that involve referring. Focusing on simple tasks such as image captioning or dialogues with a single dialogue game is not enough. The task dependence has implications for transfer learning as this should be difficult between tasks that differ considerably. This could be the reason why using language-independent object detection in multi-modal NLP tasks with language-based transformers is better than utilising pre-trained visual embeddings which have been trained together with language. This way an interaction model can be trained separately and specifically for each task.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

J.F. Blinn. 1994. Compositing. 1. theory. *IEEE Computer Graphics and Applications*, 14(5):83–87.

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

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. The MIT Press, Cambridge, Massachusetts and London, England.

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

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

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of PotsDial - Semdial 2021: The 25th Workshop on the Semantics and Pragmatics of Dialogue*, Proceedings (SemDial), pages 50–60, Potsdam, Germany.

Simon Dobnik and Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue over spatial scenes. In *Proceedings of AixDial – Semdial 2018: The 22st Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, Aix-en-Provence, France.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.

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

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

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

Jacqueline C Kowtko, Stephen D Isard, and Gwyneth M Doherty. 1992. Conversational games within dialogue. HCRC research paper RP-31, University of Edinburgh.

Barbara Landau. 2016. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Charlotte Linde and J.A. Goguen. 1980. On the independence of discourse structure and semantic domain.

In *18th Annual Meeting of the Association for Computational Linguistics*, pages 35–37, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021. Annotating anaphoric phenomena in situated dialogue. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR) at IWCS 2021*, pages 78–88, Groningen, Netherlands (Online). Association for Computational Linguistics.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

Songrit Maneewongvatana and David M. Mount. 1999. Analysis of approximate nearest neighbor searching with clustered point sets. In *Data Structures, Near Neighbor Searches, and Methodology*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Vera Silfversparre. 2021. The red cup on your left: Reference, coreference and visual attention in visual dialogue. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Moa Ekbom.

Manfred Stede. 2011. Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

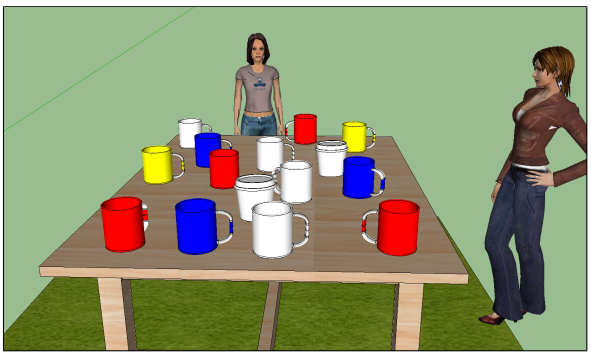
Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue referring to spatial scenes. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: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

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

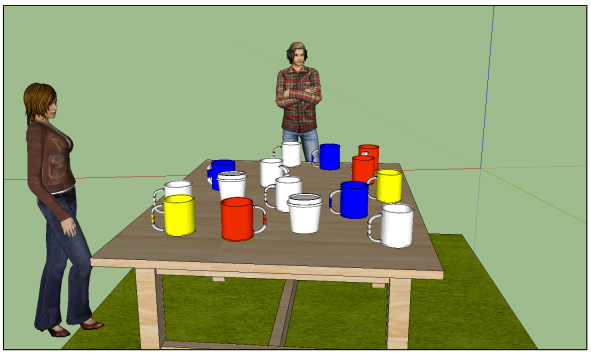
Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.

A Appendix



(a) The view of P1

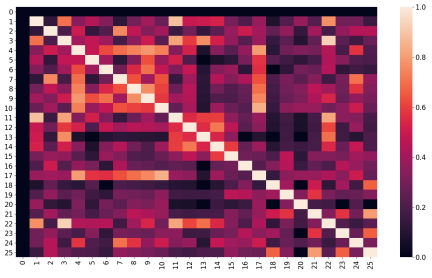


(b) The view of P2

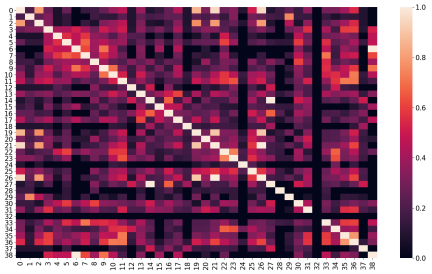
Figure 7: The scene as seen by P1 (a) and P2 (b). P3 is a passive observer Katie.

Dlg	*Language*	*Native*	*Duration (min)*	*Length (turns)*
en-1	English	Swedish	≈30	157
en-2	English	British	≈60	441
sv-1	Swedish	Swedish	≈80	118
sv-2	Swedish	Swedish	≈40	114
sv-4	Swedish	Swedish	≈30	75
sv-5	Swedish	Swedish	≈60	163
sv-6	Swedish	Swedish	≈60	248
sv-7	Swedish	Swedish	≈60	308

Table 5: The coverage of the Cups corpus per dialogues



(a) sv-2



(b) sv-5

Figure 8: Dice-Sørensen coefficient of reference overlap for all pairs of dialogue games.

Dialogue	χ^2	*N*	*p*	*dof*	sig
sv* vs en*	60.5756	4344	0.0008	30	***
en-1 vs en-2	29.0450	1446	0.5152	30	ns
en-1 vs sv-1	86.7180	739	2.05E-07	30	***
en-1 vs sv-2	71.0177	958	3.54E-05	30	***
en-1 vs sv-4	26.4639	847	0.6513	30	ns
en-1 vs sv-5	25.7953	951	0.6855	30	ns
en-1 vs sv-6	36.5352	1082	0.1912	30	ns
en-1 vs sv-7	57.3736	1189	0.0019	30	**
en-2 vs sv-1	130.6225	1229	1.61E-14	30	***
en-2 vs sv-2	80.8790	1448	1.48E-06	30	***
en-2 vs sv-4	48.0981	1337	0.0194	30	*
en-2 vs sv-5	47.6964	1441	0.0212	30	*
en-2 vs sv-6	35.2543	1572	0.2335	30	ns
en-2 vs sv-7	83.5087	1679	6.11E-07	30	***
sv-1 vs sv-2	99.9123	741	1.92E-09	30	***
sv-1 vs sv-4	89.3357	630	8.28E-08	30	***
sv-1 vs sv-5	130.1799	734	1.92E-14	30	***
sv-1 vs sv-6	117.1598	865	3.04E-12	30	***
sv-1 vs sv-7	84.5977	972	4.22E-07	30	***
sv-2 vs sv-4	32.3097	849	0.3533	30	ns
sv-2 vs sv-5	37.9412	953	0.1513	30	ns
sv-2 vs sv-6	38.9441	1084	0.1270	30	ns
sv-2 vs sv-7	70.8348	1191	375E-05	30	***
sv-4 vs sv-5	30.1364	842	0.4587	30	ns
sv-4 vs sv-6	52.2275	973	0.0072	30	**
sv-4 vs sv-7	48.9344	1080	0.01597	30	*
sv-5 vs sv-6	41.4435	1077	0.0798	30	ns
sv-5 vs sv-7	84.6866	1184	4.10E-07	30	***
sv-6 vs sv-7	83.4423	1315	6.25E-07	30	***

Table 6: χ^2 test of independence comparing reference to scene entities across different dialogues. *** indicates $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ and ns indicates no statistically significant association.

	bed	lit lamp	white ceiling	framed picture	room	wooden nightstand	yellow wall	white rug	···
bed	9030	57	239	593	256	209	166	0	···
lit lamp	57	6	5	6	5	5	6	0	···
white ceiling	239	5	206	66	67	3	40	3	···
framed picture	593	6	66	1144	58	8	29	1	···
room	256	5	67	58	278	7	34	0	···
···	···	···	···	···	···	···	···	···	···

Table 7: A sample of the referential vector space. The values represent the frequencies of objects being referred to together in the same discourse. For example, the object "bed" is referred to in the same discourse with the object "white ceiling" 239 times.

	wall	bed	wooden headboard	white wall	lit lamp	white ceiling	pillow	framed picture	
wall	8262	3167	4916	3458	3385	3991	2685	2207	···
bed	3167	7138	3074	2500	2961	2278	1657	1843	···
wooden headboard	4916	3074	6154	2482	2504	1686	2111	1699	···
white wall	3458	2500	2482	3554	2412	2137	3029	1199	···
lit lamp	3385	2961	2504	2412	1326	1930	1536	1442	···
···	···	···	···	···	···	···	···	···	···

Table 8: A sample of the situation vector space. The values represent the frequencies of objects occuring together with other objects in scenes. The rows and columns do not directly match the rows and columns in the referential vector space because not all objects are referred to in descriptions.

	Most similar objects (ranked from 0 to 9)								
	wall	bed	window	wooden table	white wall	wooden chair	white bed	wooden floor	white sink
0	wall	bed	window	wooden table	white wall	wooden chair	white bed	wooden floor	white sink
1	window	wall	white wall	window	white lamp	wooden table	white pillow	white ceiling	bathroom
2	wooden table	window	wooden floor	wooden floor	white window	wall	white lamp	room	mirror
3	framed picture	white pillow	lamp	white wall	lamp	window	white window	white window	white toilet
4	white wall	white wall	glass window	brown chair	room	wooden floor	white pillows	green plant	white bathtub
5	lamp	white lamp	white lamp	white ceiling	floor	white wall	room	light	white towel
6	wooden floor	lamp	floor	floor	white ceiling	black chair	white wall	floor	large mirror
7	white pillow	framed picture	framed picture	glass window	wooden floor	brown chair	brown pillow	large window	tiled floor
8	white lamp	white window	white ceiling	framed picture	wooden headboard	chair	white shade	ceiling	white tub
9	floor	wooden headboard	white window	lamp	white shade	glass window	lamp	white door	brown floor
	Least similar objects (ranked from 0 to 9)								
	patio	black hair dryer	blue drawers	blue plant	red plant	black windows	off television	wooden cross	brown ground
0	patio	black hair dryer	blue drawers	blue plant	red plant	black windows	off television	wooden cross	brown ground
1	large curtains	wooden doors	store	store	black toaster	glass dish	green bedspread	black shadow	red rack
2	gold lights	green bathroom	brown staircase	white dishes	open book	hanging mirror	blue recliner	hanging chain	wooden doors
3	open doorway	yellow frame	bird	yellow machine	blue recliner	toothbrush	yellow frame	black mantle	yellow frame
4	black pipe	bird	yellow frame	red balloon	wine glass	robe	green fence	orange light	green bathroom
5	gold rod	brown staircase	wooden doors	red kettle	yellow kite	gold hook	wooden doors	small toy	green fence
6	glass pitcher	green fence	red head	gold light switch	metal towel rack	white bucket	green bathroom	wooden light	bird
7	wooden cross	green bedspread	green bathroom	white star	beige sofa	dark window	brown staircase	black hair dryer	brown staircase
8	green drawer	yellow sink	green fence	red head	blue comforter	black holder	bird	wooden plate	pink basket
9	small toy	brick floor	yellow sink	white robe	plaid towel	orange soap	gray door	gray telephone	yellow sink

Table 9: Column names indicate either the most frequently occurring objects in images (the top section of the table) or the least frequently occurring ones (the bottom section of the table). Similarities are calculated on the **referential vector space** and objects are ordered from the most similar (0) to the least similar (9).

	Most similar objects (ranked from 0 to 9)								
0	wall	white wall	window	wooden table	room	framed picture	wooden chair	white ceiling	white window
1	wooden headboard	lit lamp	white wall	lit lamp	white wall	lit lamp	white wall	room	framed picture
2	white wall	white wall	lit lamp	pillow	white ceiling	white wall	lit lamp	wooden nightstand	wooden nightstand
3	bed	wooden headboard	wall	white ceiling	framed picture	framed picture	framed picture	white rug	white rug
4	white ceiling	white ceiling	bed	framed picture	room	wooden table	wooden nightstand	tiled floor	wooden table
5	lit lamp	room	white ceiling	room	tiled floor	tiled floor	room	yellow wall	tiled floor
6	pillow	framed picture	pillow	wooden headboard	wooden table	room	yellow wall	wooden table	green plant
7	framed picture	wall	framed picture	wooden nightstand	wooden nightstand	green plant	white rug	round table	white vent
8	room	pillow	room	tiled floor	green plant	white rug	round table	silver refrigerator	silver refrigerator
9	yellow wall	wooden table	wooden nightstand	wooden table	pillow	wooden nightstand	tiled floor	white vent	black printer
10	wooden nightstand	wooden nightstand	yellow wall	white rug	white rug	white vent	silver refrigerator	green plant	clear wine glass
	Least similar objects (ranked from 0 to 9)								
	gray skirt	black handlebars	black hair dryer	large statues	beige light switch	small shelf	pole	purple cabinet	orange stripe
0	purple stripe	stone column	silver toilet brush	brown umbrella	red booth	blade	blue stairway	striped floor	small sailboat
1	blue tablecloth	brown soap	red door frame	large statues	wooden bowl	pink shoes	pink shoes	glass refrigerator	brown soap
2	red boat	girl	egg	gold candle	blue mouse pad	long tie	glass shower door	metal chain	pink shoes
3	mountains	pink shoes	dark nightstand	blue mouse pad	pink shoes	purple table	orange counter	blue mouse pad	black bird
4	gold tree	blue mouse pad	palm plant	brown soap	blue doors	purple bird	blue mouse pad	pink shoes	girl
5	blue mouse pad	blue backsplash	gray skirt	girl	orange counter	black handles	wooden pen	decorative wall	black handles
6	metal chain	metal stand	blade	pink shoes	purple shade	white hole	black button	metal lock	silver heater
7	bucket	wooden pen	pink shoes	decorative wall	glass shower door	black person	decorative wall	circular mirror	blue mouse pad
8	blue radiator	decorative wall	trees	black button	blue keyboard	candles	girl	brown soap	plastic chair
9	silver wheel	black handles	gold light switch	tan surfboard	decorative wall	tan baseboard	black bird	black handles	wooden holder
10	gold bed	purple clothes	blue doors	wooden pole	wooden cabinet door	wooden tires	black handles	purple shade	red roof

Table 10: Column names indicate either the most frequently occurring objects in images (the top section of the table) or the least frequently occurring ones (the bottom section of the table). Similarities are calculated on the **scene vector space** and objects are ordered from the most similar (0) to the least similar (9).

Language and Cognition as Distributed Process Interactions

Eleni Gregoromichelaki
University of Gothenburg
eleni.gregoromichelaki@gu.se

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

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

Gregory J. Mills
University of Groningen
g.j.mills@rug.nl

Ruth Kempson
King's College London
ruth.kempson@kcl.ac.uk

Julian Hough
Queen Mary University of London
j.hough@qmul.ac.uk

Patrick G. T. Healey
Queen Mary University of London
p.healey@qmul.ac.uk

Matthew Purver
Queen Mary University of London
m.purver@qmul.ac.uk

Abstract

In this position paper we argue that a conception of linguistic competence and conversational abilities that would fulfil the aims of Artificial General Intelligence cannot remain characterised as a static system of patterns induced from disembodied textual data. Instead, it should be modelled as a continuous, active, and interactive learning process. This is in line with the metaphysical and cognitive assumptions of Interactivism regarding the fundamental status of processes, as well as distributed cognition perspectives which argue that language does not reside in individual minds, brains, or bodies but is "spread out", embedded, and distributed in the available multimodal interactions with the environment. We show the usefulness of the formalism of Dynamic Syntax with Type Theory with Records (DS-TTR) in modelling dialogue to this end.

1 Introduction

Until recently, internalistic and static accounts of cognition have been the mainstream position in cognitive science and philosophy. However, dynamic accounts are now on the rise (Noë, 2004; Bickhard, 2009; Seibt, 2018; Manzotti and Chella, 2018, a.o.) alongside a growing interest in process metaphysics, substantiating the intuitive phenomenal idea of a dynamic, ever-changing reality while further justification is provided by recent relational interpretations of quantum mechanics (Laudisa and Rovelli, 2021) and category-theoretic results in mathematics (e.g. the Yoneda lemma, see Bradley et al., 2021). Moving away from the computational theory of mind with brain-internal representations and computations, current theories also argue that body–world interactions is what should be taken to constitute cognition (see, e.g. Hutchins, 1995).

In contrast, the idea of human language knowledge as an *abstract* and *static* system still underpins much work in theoretical linguistics, as well as language model architectures underlying recent impressive advances in NLP and AI (such as BERT (Devlin et al.), GPT3 (Brown et al., 2020) and their multimodal analogues e.g. ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), Imagen (Saharia, 2022), DALL-E 2 (Ramesh et al., 2022), Gato (Reed et al, 2022) a.o.). The same view has been taken in computational dialogue modelling across the board, which retains the idea of human language knowledge as an autonomous and static system state. This system reconstructs human thought and communication as underpinned by module-internal rules and representations of a grammar and a lexicon enriched by some theory of mind module to explain performance. It is then natural to suggest that the system can be learned from static, disembodied textual data, and used for various downstream tasks after suitable fine-tuning.

Models implemented under this perspective have achieved great success in tasks that depend on reproducing patterns of very limited interactions with their environment (e.g., predicting upcoming input text), which allows learning of patterns of relationships among words. However, most researchers now concede that we have reached a point of diminishing returns (Bender and Koller, 2020). The constant increase of scale in amounts of data, computational resources, and parameters that are now required for minimal progress is un-

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

sustainable both environmentally and due to the complexity of "black box" 'foundational models' (Bommasani et al., 2021). This results in lack of trust and confidence by users and the public due to the inscrutability and unexpected behaviours of current systems (see, e.g., Molnar, 2022).

As an alternative, we argue that language that fulfils the aims of AI and full artificial general intelligence (AGI) cannot be characterised as a static system of patterns induced from data as the result of learning but remaining unchanged during moment-to-moment (incremental) interaction with the user. Instead, language needs to be itself characterised as a *continuous, active,* and *interactive learning process*. This means that constant change and adaptation is what sustains any stable organisation we might detect during snapshot observations. This is in line with distributed cognition perspectives which argue that language is a system *property* that does not reside in individual minds, brains or bodies but is "spread out", embedded and distributed in the available multimodal interactions with the sociomaterial environment (cf also the Vygotskian robotics perspective e.g. Mirolli and Parisi, 2011).

One particularly acute symptom of the misconception of language as a static, representational system is that progress in creating natural interactions in conversational AI (aka 'dialogue systems') has plateaued. End-users of such systems have expectations of naturalness, intelligence, flexibility, and robustness to error, regularly leading to disappointment and frustration (Moore, 2017; Clark et al., 2019; Chaves and Gerosa, 2021; Luger and Sellen, 2016; Fischer et al., 2019). Large-scale end-to-end neural architectures (e.g. Wolf et al., 2019) display impressive capacities in terms of producing fluent immediate responses, but do not adequately capture human capacities in learning *appropriately adaptive* incremental conversational behaviours. Often such systems neglect the overall coherence of a situated dialogue setting thus lacking consistency with respect to the longer history of the dialogue and its future prospects with respect to achieving some goal (see e.g. Li et al., 2020; Vinyals and Le, 2015; Shang et al., 2015; Sordoni et al., 2015). As a consequence, today's conversational AI systems do not possess the strategic and embodied skills to negotiate the ambiguity, vagueness, and nuances of human-human conversation, and thus cannot learn and adapt to new people, tasks, and situations.

In this respect, critics of deep learning and current AI constantly point out that what is missing from such models is some notion of "semantics" to be articulated independently from the level of "forms", which is what is supposedly captured by such models (see, e.g. Bender and Koller, 2020; Bender et al., 2021). However, this criticism is only valid if it is taken for granted that there is such an objectively defined separation, i.e., form vs meaning, and, moreover, that AI systems of whatever variety are all meant to operate independently as autonomous cognitive agents. Alternatively, from the perspective of seeing language as a constructivist sociocultural process, form and function do not have to be distinguished but both of them can be seen as human abstractions of the epiphenomenal effects of underlying processes. Process organisation is what constitutes 'form' but such organisations are inherently functional. Given that processes interact and self-organise with emergent results at various levels (Bickhard, 2021), the autonomy of AI and NLP systems does not have to be taken as an all-or-nothing issue but as gradations of autonomy and independence depending on the purposes of use and the abilities of the agent. Unlike Piagetian constructivist views of human development, which arguably resemble the construals of current foundational models' learning regimes, Vygotskian cognitive robotics approaches to higher-level cognitive skills emphasise the 'internalisation' of social processes within individual minds transforming interpersonal processes to intrapersonal operations (e.g. Mirolli and Parisi, 2011, cf. Bruineberg and Rietveld 2019). This approach retains the primacy of the organism's interaction with the sociomaterial environment as the unifying factor of the relevant processual (self-)organisation while also accounting for autonomous performance. From this perspective, a language model that is taken to solipsistically receive and process inputs similarly to an isolated "brain-in-a-vat" does not provide an adequate basis for expecting human-level performance. However, text-to-image systems like DALL-E 2 and Imagen or generalist systems like GATO (Reed et al, 2022) that connect language with another modality like vision and operate across various tasks are a first demonstration that convincing linguistic performance is not due to an autonomous knowledge system performing "linguistic" tasks in isolation. Instead, the processes that constitute the *linguistic* organisation of a system, whether human-

human, or 'human-in-the-loop', comprise a mode of perception/action that structures the phenomenal world for other modalities deriving the social co-constructive nature of cognition. Thus moving towards more realistically embedded language models, implemented through artificial agents that interact more and more autonomously but under the normative forces imposed by the sociomaterial environment, sustains the possibility of eventually developing artificial general intelligence (AGI).

In this position paper, we set out the challenge of language as process (Gregoromichelaki, 2018; Gregoromichelaki et al., 2019, 2020b,a), rejecting the separation between form and meaning, syntax and semantics/pragmatics, or structure and function. We then reflect on the effects of incorporating the process of establishing coordination in social interactions into the core of the model itself.

2 The inadequacy of code models and Gricean mechanisms

Human communication is often characterised under the 'code model', namely, as one agent encoding and transmitting a message (the 'sender') to be decoded by another agent (the 'receiver'). This is an instance of the 'encodingism problem' in cognitive science as identified by (Bickhard, 2009) a.o. Successful communication is characterised as the hearer correctly discovering some preformed message which the speaker intended to convey. This basic assumption underlies most psychological and pragmatic theories of interaction including the Interactive Alignment Model (Pickering and Garrod, 2004), Gricean pragmatics and Relevance Theory (Sperber and Wilson, 1995) which assume an underlying literal meaning enhanced by context-specific pragmatic inferences to uncover the speaker's intention. But this approach has failed spectacularly to account for the complexity and subtlety of sense-making in human interaction (see e.g., Rączaszek-Leonardi et al., 2014; Fowler and Hodges, 2016).

This failure is because the actions of participants in dialogue form a system of coupled components (see e.g., De Jaegher and Di Paolo, 2007) so that *feedback mechanisms*, like constant error indication and adjustment, are crucial for the stability, maintenance, and self-organisation of the system. Given the moment-by-moment possibility and precariousness of action coordination, participants do not need explicit representations of their own or others' mental states, and nor do they need to converge on a shared 'code' or criteria of success. Instead, their conceptions and contributions need to be complementary to sustain a social practice whose normative character is defined externally to their own private or explicit rationalisations of their behaviour.

Rethinking our conception of successful communication away from shared codes puts the flexibility and dynamism of natural language (NL) at the heart of communication. As Healey et al. (2018b) state "[i]nstead of thinking of effective communication as formulating a "perfect" message, it becomes about finding optimal ways to uncover and address misunderstandings" (see also Healey et al., 2018a). We go further, and do not characterise these practices as uncovering 'misunderstanding' or 'miscommunication', terms which suggest that they are in opposition to some common understanding or common ground. Instead, we characterise successful coordination (i.e. system self-organisation, rather than "communication") as the local, incremental accommodation of inevitable and necessary perturbations in the emergent formation of a complex dynamical system enabling people's contributions to larger social organisations that constitute their ecological niche ('form of life').

From a psychological perspective, the rapidity and high incrementality of turn-taking exchanges in dialogue (Levinson and Torreira, 2015; Sacks et al., 1974) shows that intractable exhaustive reasoning about some optimal local outcome is not what participants aim for (cf. Frank and Goodman, 2012). Instead, practices of navigating through, and local adjustment to, an incrementally evolving landscape of *affordances* (Rietveld et al., 2018) provided by the ecological niche and participants' own actions, enable the forms of distributed cognition observed in dialogue (e.g. Dingemanse, 2020).

Transferring this insight to the domain of language technology, this assumption partially explains the limited success of language models in mimicking many aspects of human performance in dialogue, especially when it comes to coordination and adaptation. We attribute the substantial current shortcomings of such models to the limited variety of data they are exposed to, lack of the ability to actively interact with the data (cf. Li et al., 2017; Lewis et al., 2017), lack of feedback, lack of physical embodiment (see e.g. Pustejovsky and Krishnaswamy, 2021), and lack of a system

of values (normativity) engendered through some moral framework (Hodges, 2022). We suggest that progress in modelling human dialogue and conversational AI requires a radical reconception of NLs as *mechanisms for (inter)action*.

Affordances and repair Under our interpretation, affordances are publicly available resources which trigger motivations for action within agents (*solicitations*, e.g. Dreyfus, 2013). Affordances are not, as standard, simply properties of the environment or agent-internal mechanisms (cf. Bickhard, 2009). Rather, they are relations between agent abilities and what the current sociomaterial environment makes available. This means that the shifting set of affordances in dialogue concerns the collective potential of the interactants, rather than individual perspectives whose meshing needs to be explicitly negotiated/represented. Interlocutors thus acquire a joint perspective as long as they operate as a system with autonomous self-organisation underpinned by prediction error minimisation (as modelled within the Free Energy Principle framework in its ecological/enactive interpretation, e.g., Bruineberg et al., 2018; Kiverstein et al., 2022). The local and shifting landscape of affordances and the state and abilities of the agents involved determine at each moment a demarcated 'field of affordances', i.e., a subset of the landscape of affordances that are perceived as relevant by the agents. This provides for a joint conceptualisation of the current action potential with minute adjustments at each subsentential stage resulting in the appearance of planned rational action at the macro-level. It also removes the need to define propositional structure substitutes to account for partial 'situation convention' transformations (Bickhard, 1980, forthcoming). Additionally, rather than modelling repair of intention recognition failures as phenomena in (1) and (2) are standardly characterised, this externalist and distributed perspective aims at modelling the strategically introduced public intention co-construction through the affordances of so-called 'repair mechanisms' (see also Haugh, 2008; Haugh and Obana, 2015; Arundale, 1999):

(1) (a) A: so …umm this afternoon …
(b) B: let's go watch a film
(c) A: yeah

(2) (a) A: I'm pretty sure that **the**
(b) B: **programmed visits**?
(c) A: programmed visits, yes, I think they'll have been debt inspections. [BNC KS1 789-791]

3 Form, meaning, and interaction

Looking at single individuals out of context, there are unlimited degrees of freedom available for realising action opportunities, which leads to intractability, especially in Gricean models where coordination is modelled as recursive mindreading. This limitation can be overcome by conceptualising conversational interaction as process organisation into a coherent system: when agents become coupled and subsumed under an emergent sociocognitive system, degrees of freedom are severely restricted due to the top-down constraints exercised on individuals to perform their particular role in the achievement of joint action (e.g. Deacon, 2011). This helps to locally constrain individual choices, without individuals having to necessarily conceptualise such choices or build matching models of reality inside their own heads (i.e. with the world taken to be its own "best model", (e.g. Brooks, 1990; Hutchins, 1995).

Mismatches in skills and information are necessary ingredients of such an emergent process of coordination and complementarity in action. While compatibilities between participants act as a channel for smooth, automatic navigation of aspects of a shared space of action opportunities (affordances), they also form the background for revealing divergences. These divergences constitute sources of scaffolded learning and thus require attention and work to sustain the interaction. The prerequisites and presuppositions of the interaction thus become "present-at-hand" (Heidegger in (Dreyfus, 1990)) and constitute sources of learning and development by "educating the attention" (Gibson, 1966) of agents allowing them to differentiate novel opportunities or threats in their joint environment. Divergences trigger 'solution probing' processes, where the interlocutors attempt to reorient the trajectory of the joint action towards its incrementally emerging joint goals. At these points, aspects of the interaction regarding what is "appropriate" in that particular sociocultural practice (social normativities) become available as experiences and training for the individual participants who are in this way enabled to learn and develop their skills through interactions scaffolded by the relevant practices and other agents' abilities and guidance (see, e.g. Steffensen et al., 2016).

Data from human-human dialogues, such as (3), provides evidence that participants can fluently interact, with emergent coordination, despite the fact

that conversational exchanges are superficially full of "fragments", non-linguistic signs, disfluencies, and non-verbal signals such as gestures and gaze:

(3)
1. **J:** Can you think of any catalysts?
2. **A:** Er is it potassium permanganate?
3. **J:** <unclear>
4. **A:** What
5. **J:** Pla <pause> a duck billed
6. **A:** Pardon?
7. **J:** A duck billed
8. **A:** Platypus.
9. **J:** And it's not platypus it's <pause> sounds like a type of pen.
10. **A:** Platinum.
11. **J:** Right, platinum. [BNC; FMR 728-737]

As seen here, units of meaning are co-created incrementally (Gregoromichelaki et al., 2013; Kempson et al., 2016) by multiple interlocutors using incomplete utterances (e.g. line 7 – Purver et al., 2011), with phenomena such as cross-person compound contributions (where one person continues another's utterance, as in lines 7 and 8 – Lerner, 1991; Howes, 2012), repairs (e.g. the clarification requests in lines 4 and 6 – Sacks et al., 1974; Purver, 2004), and disfluencies (e.g. the pause and restart in line 9 – Hough, 2015) – seen as 'performance errors' in traditional linguistics – crucial in the co-construction of meaning.

In (3), a chemistry tutor (J) prompts a student (A) to answer the question in line 1, illustrating the divergence and convergence complementarity that is key to driving dialogue forwards. The social roles of teacher and student constrain the way in which their several responses are interpreted and this interplay and meshing of factors belies distinctions such as form vs meaning, communication vs thought or speaker vs listener. From a standard individualistic perspective, one can characterise the exchange as indicating that from J's perspective, A's response in line 2 diverges from the expected answer. A finally produces the expected answer (thus demonstrating convergence with J's expectations) in line 10. This is a valid way of describing the process and could be how a single participant or observer might rationalise or abstract the dialogue process into a narrative that they construct post hoc. This meta-perspective is arguably the one that prevailed in the construction of dialogue systems (e.g. Kopp and Krämer, 2021) in the era before end-to-end statistical models.

However, this view neglects the fact that both participants operate in a context (a 'teaching context') that imposes normative constraints in what their actions should be as they perform the roles assigned to them by the sociocultural convention: there are no 'teacher' or 'student' roles outside this socially-afforded context. This is not necessarily a conceptualisation that is explicit in any individuals' real-time consciousness but it is an effect of the 'habitus' (a set of embodied dispositions, *solicitations*, e.g. (Dreyfus, 2013), or *effectivities* (Turvey, 1992)) that agents have acquired through enculturation. The characterisation of the interactive potential here is similar to Bickhard's 'situation convention' with the difference that it is not grounded exclusively through the participants' internal understanding or awareness. The practice is enabled outside the agents' brain processes to constitutively include extended temporal, material, and spatial processes converging in the interaction. In its turn, the process organisation that constitutes the practice constitutes the participants' (temporary) identities and the action possibilities afforded to them.

The exchange of information in the sense of 'semantic information' assumed in model-theoretic, denotational, or referential semantics is not the purpose of the interaction. Neither are Gricean or Neo-Gricean norms relevant in the sense of trying to figure out a speaker's communicative and informative intention. Instead, the task, or language game, here seems very similar to the elicitation tasks that current 'foundational' models are confronted with: sometimes they are required to complete a NL prompt given some additional context, or to produce an image by taking advantage of their experience with 'forms' of text and images that they have sifted over and compressed in their parameters and architecture (cf. Marcus, 2022). The functioning of these form-based results is then to be normatively determined within the overarching language game, which for foundational models is set by human users, thus minimising the agential properties of the models.

In the current case, the overarching goal is set by J and A's agency is minimised in the sense that A's responses are normatively judged as appropriate by J. From J's perspective, A's response in line 2 does not achieve the joint normative goal of the student-teacher context which A finally produces in line 10, namely, to enable A to respond appropriately when the situation requires retrieval of the type of elements that can be characterised as 'catalysts'. The naming word here (*catalyst*) has both linguistic and non-verbal affordances that are both targeted by the tuition. Inability to proceed

is explicitly conveyed by A's clarification requests which act as signals for J to produce prompts probing A's knowledge of word *forms* to induce the answer. After a cue in line 5 fails to elicit the required convergence, J exploits the predictability induced by the compound noun phrase *duck-billed platypus* to get A to produce the first syllables of the answer to the original question. Of course, J's purpose is not to just entrench word form associations with the word *catalyst* in A. Instead, it is taken for granted that the signs (forms) constituting the words have action implications for the constitution of A as a capable agent with respect to chemistry. Form and meaning then, or 'natural meaning' and 'non-natural meaning', are not separate categories but abstractions that in reality stand for qualitatively similar and interrelated processes within organisations of networks of affordances (Bickhard forthcoming cf. Skyrms, 2010).

Both participants' actions are subsumed under the context-specific normative perspective that their actions be relevant to the elicitation of some particular answer to the question posed by J, with both operating as a coherent system performing complementary actions towards that goal and compensating for each other's failings to contribute appropriately. This management of the divergent and convergent contexts is incrementally and locally managed, with a hierarchy of joint goals and subgoals emerging opportunistically. J and A can only have probabilistic expectations as to what they are required to do moment-by-moment and have to correct and adjust their performance based on the feedback received.

In this dialogue, there is an asymmetry between the speakers, as J is both the expert, and more powerful than A. In fact, this asymmetry is endemic, diagnostic of not just all child/adult (Duveen and Psaltis, 2013; Kunert et al., 2011) or expert/non-expert exchanges (Lu et al., 2007; Pilnick and Dingwall, 2011), but all interactions. Differences in experiences, cultural background, individual physiology and social communities all contribute to differences in our language use, meaning that we never share the "same" language as anybody we nevertheless successfully interact with (Clark, 1998). This raises an important practical question: How can we communicate successfully when individual differences in language use are not the exception but the norm?

We believe that the answer to this question relies on reconceptualising NL grammars as modelling a set of skills for interaction relative to social practices (Gregoromichelaki et al., 2019, 2020b), in common with distributed language models (Cowley, 2009) and the dialogical perspective (Linell, 2009) but within a formally articulated architecture that lends itself to implementation. We now sketch such a model.

4 DS-TTR

DS-TTR (Purver et al., 2010, 2011; Hough, 2015) is a system that combines the dynamic logic (PDL) architecture of Dynamic Syntax (DS, see e.g., Kempson et al., 2001; Cann et al., 2005) with probabilistic versions of Type Theory with Records (TTR, Cooper, 2005, forthcoming). TTR types are interpreted in DS-TTR in dynamic terms as affordances (Gregoromichelaki et al., 2019, 2020b; Eshghi et al., 2022), that is, type names are triggers for sets of PDL actions, just as syntactic/semantic categories in DS are labels for tree-building actions. Actions are expressed as probabilistically licensed transition events among the states of a dynamic system – see Fig. 1 where outgoing edges/actions from each node form a learnable (Eshghi et al., 2013) probability distribution conditioned on the current state. DS-TTR is thus articulated in terms of conditional and goal-driven actions whose accomplishment either gives rise to expectations of further actions, tests the environment for further contextual input, or leads to abandonment of the current strategy due to its unviablity in view of more competitive alternatives (see Fig. 1). Words, morphology, and syntax are, in this way, all modelled as indicators of opportunities for (inter-)action (Gregoromichelaki, 2018; Gregoromichelaki et al., 2019, 2020b,a). Participants' opportunities for action and their perspectives are modelled in a unified model of the whole system. Interactions are modelled as incrementally opening up a range of options so that selected alternatives can be pursued either successfully or unsuccessfully: even though a processing path might be initially highly favoured, it might nevertheless lead to an impasse so that processing is aborted and backtracking to an earlier state is required (Sato, 2011) due to the changing conditions downstream.

As Fig. 1 shows, edges correspond to DS actions; and nodes correspond to states defined by their predictive potential for further actions. However, one might also take a coarser-grained view of the DAG with edges corresponding to words (sequences of computational actions followed by

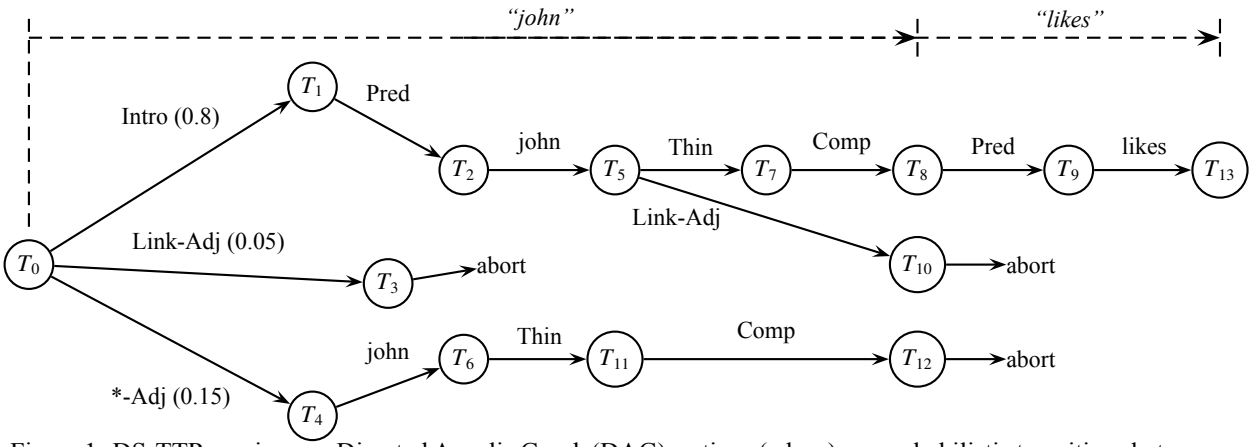


Figure 1: DS-TTR parsing as a Directed Acyclic Graph (DAG): actions (edges) are probabilistic transitions between partial trees (nodes).

a single lexical action) rather than single actions, and dropping abandoned parse paths (see Hough, 2015, for details).

On this view, DS-TTR parsing or generating a string of words or non-verbal tokens, induces some organisation of a state space of activity possibilities in combination with top-down actions ensuing from preexisting skills and dispositions of the participants involved (the 'grammar') (cf. Zadrozny, 2020). This either transforms the existing state space, adds new structural organisation to it, or removes existing paths through it. At each stage, a 'pointer' ($\Diamond$) determines the local point of modification; and locally, the immediate path trajectory moves through a tree-shaped state space with nodes as states traversed by means of constraints expressed by the modal operators (e.g. $\langle\downarrow\rangle$, $\langle\uparrow\rangle$, $\langle\uparrow_*\rangle$) of a modal tree logic (the Logic of Finite Trees; LOFT: Blackburn and Meyer-Viol, 1994) expressing topological relations among current or future anticipated (i.e. predicted) nodes. The tree-shaped organisation of local processing trajectories reflects the conceptualisation structure induced by the unfolding utterance in terms of function-argument articulations. More globally, the state space is presented as a directed acyclic graph (DAG) that records possible paths of actions in a landscape defined by what the grammar, acting as a controller of the normativity pertaining to linguistic actions, allows as predictions of future interaction possibilities. The context required for processing various forms of context-dependency *is* the path searches provided by the DAG, augmented by affordances pertaining to the 'form of life' (e.g. Bruineberg et al., 2018) within which the interaction takes place.

Given the basic property of *predictivity* that sustains the DS-TTR mode of explanation of linguistic phenomena, the task confronting a DS-TTR learner is similar to the self-supervised language modelling task and even closer to current Reinforcement Learning (RL) architectures. Eshghi et al. (2017a,b) show how this idea can be implemented in narrow dialogue domains, where DS-TTR action policies are learned through exploring environmental contingencies (affordances) and acquiring skills in predicting suitable trajectories within the evolving landscape of affordances via RL methods. Hence, an induced DS-TTR grammar can be seen as a generative model capturing the interaction potential of a situational context, the latter including agents and sociomaterial constructs as in distributed cognition research.

5 Modelling feedback in DS-TTR

Given these inherent properties, DS-TTR has lent itself particularly well to dialogue modelling and analysis of dialogue phenomena within a unified architecture. Dialogue is modelled as the incremental and interactive composition of action sequences triggered by words either from oneself (in production) or an interlocutor (in comprehension) in an incrementally evolving context, the DAG past or future defined trajectories constituting the context, enabling unitary explanations of ellipsis (Kempson et al., 2015), self-repair (Hough and Purver, 2012; Hough, 2015), split utterances (Howes et al., 2011; Howes, 2012; Kempson et al., 2016), clarification requests (Gargett et al., 2009; Eshghi et al., 2015) and other feedback (Howes and Eshghi, 2021). In particular, it provides a basis for modelling backchannels (indications of agreement) vs clarification requests (overt indications of needing further development to enable agree-

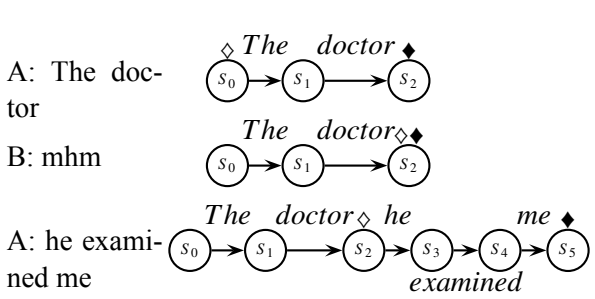


Figure 2: Backchannels as coordination pointers' movement in Interaction Control State-space (ICS)

ment), extensions vs corrections, hence 'repair', all as complementary procedural mechanisms for managing the types of transformations induced moment by moment in the ever evolving DAG space. As Eshghi et al. (2015) show, 'grounding' (the integration into the context of feedback) in a dyadic dialogue can be captured by including the perspective-relativisation of affordances: the DAG is augmented with two *coordination pointers*, the *self-pointer*, ♦, and the *other-pointer*, ◇, marking the points up to which the dialogue participants have each grounded the material. We dub this augmented context DAG, the *Interaction Control State-space* (ICS) - see Fig. 2.

Any utterance causes ICS pointer movement, and interlocutors each have their own ICS paths which can diverge, and re-converge as a result of clarification interaction and repair processes more generally. The self-pointer, ♦, on participant A's ICS tracks the point to which A has given evidence for reaching. The other-pointer, ◇, tracks where the other participant, B, has given evidence for reaching. For example, an utterance produced by A will move A's self-pointer on their own ICS to the rightmost node of their ICS; on B's ICS, it is the other-pointer that moves to the same location. On this model, the intersection of the path back to the ICS root from the self- and other-pointers is taken to be grounded, with the effect that parse or production search within this grounded pathway is precluded, thus removing the computational cost associated with finding alternative interpretation pathways, as well as formally explaining how conversations move forward.

This model has been shown to account for backchannels (Fig. 2), clarification interaction, and other-corrections (Eshghi et al., 2015; Howes and Eshghi, 2017, 2021). Clarification requests cause branching on the ICS, where the current path is abandoned and another branch constructed – a subsequent response plus the acknowledgement of this response eventually realigns the two coordination pointers, and the interlocutors' ICSs as a consequence (see Eshghi et al., 2015; Howes and Eshghi, 2021 for details). By contrast, backchannels and utterance continuations do not create new branches, but move the other-pointer forward on the current path.

6 DS-TTR in alignment with process and relational models of cognition and reality

DS-TTR and Interactivism (Bickhard, 2009, and elsewhere) share a lot in common. Both embrace the claim that the underlying foundation of linguistic theorising has to be reconsidered to a perspective that embraces the action-grounding and process metaphysics that standard representational frameworks have obscured. In this view, action dynamics are primary with processes being the most fundamental individuals (Seibt, 2018). Language processing is thus seen in both frameworks as transformation of a landscape of affordances (in DS-TTR terms) instead of decodings of denotational contents augmented by Gricean reasoning.

The two paradigms are thus strikingly congruent, yet they diverge in some respects. While agreeing that language is the fulcrum between what is mind-internal and -external, they diverge in the interpretation that they attribute to the process organisations that they invoke. The Bickhard view posits agent-internal representations (through *apperceptions*) as a necessary intermediate level of process in order to define error detectable by the agent. This is a crucial assumption for grounding Bickhard's notion of 'representation', albeit in non-standard dynamic terms. But from the DS-TTR point of view, this seems to presuppose that an agent has access only to its own dynamic mechanisms and processes, even when the agent is embedded in an overarching organisation like the one captured in a DS-TTR DAG. This means that the brain-internal perspective dominates the grounding of 'representation' even in such a ground-breaking dynamic model like the Interactivist model. In contrast, DS-TTR is more compatible with forms of radical realism, which construe the very existence of the objects of phenomenal experiences, including minds and languages, as products of interactions (e.g. Manzotti and Chella, 2018, cf. Laudisa and Rovelli, 2021;

Adlam and Rovelli, 2022), hence eliminating the need for a separate notion of mind-internal representations, without excluding them of course in certain circumstances. On this view, affordances are truly relational, generated and realised within distributed systems comprising multiple agents and within-agent levels. As in various forms of enactivism, social NL behaviours are understood as practices, with their normativity underpinned by a set of conditional actions (the 'grammar') inducing ongoing emergent flows that can be approximated, in more individualistic, abstract, and detached terms, as the often-studied notions of context, content, intentions, speech acts and the like. This radical extension of explanations of tools for use in communication as a core part of the grammar thus no longer corresponds to a capacity exclusively within the head of a single individual but is in some sense external to that, shared across participants. Moreover, the view of what an 'agent' is can be extended to non-biological artifacts, like artificial agents (Kockelman, 2011; Kiverstein et al., 2022). This is compatible with the view that process organisations are the fundamental explanatory factors of behaviours while metaphysical relationality implies that normativity can be attributed, albeit in a derivative sense, to the purposes of such agents (cf. Bickhard, 2021).

It is notable in this connection that the remit of data which DS-TTR is able and concerned to express corresponds remarkably closely to the insights of Conversational Analysis (CA), long widely ignored by theoretical linguists as doing no more than providing descriptions not amenable to formal characterisation, and in principle to be ignored due to merely constituting performance data (but cf. Ginzburg, 2012; Cooper, forthcoming).

Indeed the CA task was to provide a radically empiricist methodology to describe the interactions so characteristic of naturally occurring conversation. This can be given an internalist interpretation (cf. Ginzburg, 2012), but our aim here is to defuse the view that the skull or the human body provide a priori boundaries of where cognition, including grammars, is situated (cf. Albert and de Ruiter, 2018).

7 Future challenges

With grounding DS-TTR actions and types as affordances, there remains much work to be done, and at least one major problem. NLs universally display endemic context-dependence on the interpretations their words allow. Linguists are well aware of this fact, either addressing it by positing lexical ambiguities for every word of the language,[1] or attributing open-ended complexity of inference in the individual's capacity for language use. Against this challenge, the AI success in developing automated NL processing systems without any reference either to details of NL grammar formalisms or to such high-levels of inference stands in clear conflict with the abstract formalisms linguists have proposed – it is hard to envisage more damaging evidence against such approaches (Perconti and Plebe, 2019; Lappin, 2021). Much of this AI success has turned on large, neural language modelling techniques that instantiate the Firthian stance that the information-bearing load of words can be induced from the sets of words or affordances sharing the same local (multimodal) context window without any reference to intrinsic denotational content attributable to the words themselves (Gregoromichelaki et al., 2019).

In facing this challenge head on, work is currently exploring ways of combining the DS dynamic architecture with compositional Distributional Semantics tools (Purver et al., 2021). In this work, lexical items project *tensors* onto the interim emergent DS trees/states (instead of TTR record types), mapping onto vector spaces. This provides an explanatory basis from which the intrinsic non-determinism of lexical content can be modelled with content flexibility of NL expressions being essential to language variation and change (see, e.g., Gregoromichelaki et al., 2019). On this view, success in communication between participants is then predicted to rest in the emergent coordination due to the overlap shared by such spaces, for which feedback manifestly contributes as it conditions the shifting affordance landscape. This emergence, much in line with Bickhard's 'situation conventions' but externalised, plays a central role in refining emergent joint projects without requiring identity in understandings but, primarily, complementarity in action. Furthermore, work has been done in situating DS-TTR within embodied agents (Hough et al., 2020) giving non-verbal actions the same status as verbal utterances. Hence the claim that, far from defining a vehicle for communication leading to shared understanding of some defined denotational content, NL grammars are rather seen as comprising a set of skills for picking up interaction affordances within social practices.

[1] e.g., categorial grammar and its type polymorphism

References

Emily Adlam and Carlo Rovelli. 2022. Information is physical. *arXiv preprint arXiv:2203.13342*.

Saul Albert and J. P. de Ruiter. 2018. Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, 10(2):279–313.

Robert B. Arundale. 1999. An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics*, 9(1):119–153.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference*, pages 610–623, New York, NY, USA.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Mark H. Bickhard. 1980. *Cognition, Convention, and Communication*. Praeger.

Mark H Bickhard. 2009. The interactivist model. *Synthese*, 166(3):547–591.

Mark H. Bickhard. 2021. Emergent Mental Phenomena. In Robert W. Clowes, Klaus Gärtner, and Inês Hipólito, editors, *The Mind-Technology Problem*, pages 49–63. Springer, Cham.

Mark H. Bickhard. forthcoming. *The Whole Person*.

Patrick Blackburn and Wilfried Meyer-Viol. 1994. Linguistics, logic and finite trees. *Logic Journal of the Interest Group of Pure and Applied Logics*, 2(1):3–29.

Rishi Bommasani, Drew A. Hudson, and Ehsan Adeli et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. 2021. An enriched category theory of language. *arXiv preprint arXiv:2106.07890*.

Rodney A Brooks. 1990. Elephants don't play chess. *Robotics and autonomous systems*, 6(1-2):3–15.

Tom B. Brown et al. 2020. Language models are few-shot learners. *arXiv:2205.06175*.

Jelle Bruineberg, Anthony Chemero, and Erik Rietveld. 2018. General ecological information supports engagement with affordances for 'higher' cognition. *Synthese*, 196(12):5231–5251.

Jelle Bruineberg and Erik Rietveld. 2019. What's Inside Your Head Once You've Figured Out What Your Head's Inside Of. *Ecological Psychology*, 31(3):198–217.

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

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact. *International Journal of HCI*, 37(8):729–758.

Herbert H Clark. 1998. Communal lexicons. In Kirsten Malmkjær and John Williams, editors, *Context in Language Learning and Language Understanding*, chapter 4, pages 63–87. Cambridge University Press, Cambridge.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation. In *Procs of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. New York, NY, USA.

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

Robin Cooper. forthcoming. *From perception to communication*.

Stephen J Cowley. 2009. Distributed language and dynamics. *Pragmatics & Cognition*, 17(3):495–508.

Hanne De Jaegher and Ezequiel Di Paolo. 2007. Participatory sense-making. *Phenomenology and the Cognitive Sciences*, 6(4):485–507.

Terrence W Deacon. 2011. *Incomplete nature: How mind emerged from matter*. WW Norton & Company.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT. In *Procs of NAACL*, pages 4171–4186.

Mark Dingemanse. 2020. Resource-rationality beyond individual minds. *BBS*, 43:e9.

Hubert L. Dreyfus. 1990. *Being-in-the-World*. MIT Press.

Hubert L. Dreyfus. 2013. The myth of the pervasiveness of the mental. In J. K. Schear, editor, *Mind, Reason, and Being-in-the-World*. Routledge, London.

Gerard Duveen and Charis Psaltis. 2013. The constructive role of asymmetry in social interaction. In S. Moscovici, S. Jovchelovitch, and B. Wagoner, editors, *Development as a Social Processn*, pages 133–154. Routledge.

Arash Eshghi, Eleni Gregoromichelaki, and Christine Howes. 2022. Action Coordination and Learning in Dialogue. In *Probabilistic Approaches to Linguistic Theory*. CSLI.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matt Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the IWCS*, pages 261–271, London, UK. ACL.

Arash Eshghi, Matthew Purver, and Julian Hough. 2013. Probabilistic induction for an incremental semantic grammar. In *Proceedings of IWCS*, pages 107–118, Potsdam, Germany. ACL.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017a. Bootstrapping incremental dialogue systems from minimal data: Linguistic knowledge or machine learning? In *Proceedings of EMNLP*, pages 2220–2230.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017b. Interactional dynamics and the emergence of language games. In *CEUR Workshop Proceedings*, volume 1863, pages 17–21.

Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *CUI 19*.

Carol A. Fowler and Bert Hodges. 2016. Finding common ground. *New Ideas in Psychology*, 42:1–6.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Andrew Gargett, Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, and Yo Sato. 2009. Grammar resources for modelling dialogue dynamically. *Cognitive Neurodynamics*, 3(4):347–363.

James J. Gibson. 1966. *The Senses Considered as Perceptual Systems*. Houghton Mifflin.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Eleni Gregoromichelaki. 2018. Quotation in Dialogue. In Paul Saka and Michael Johnson, editors, *The Semantics and Pragmatics of Quotation*, pages 195–255. Springer.

Eleni Gregoromichelaki, Ronnie Cann, and Ruth Kempson. 2013. On coordination in dialogue. In Laurence Goldstein, editor, *On Brevity*. Oxford University Press, Oxford.

Eleni Gregoromichelaki, Stergios Chatzikyriakidis, Arash Eshghi, Julian Hough, Christine Howes, Ruth Kempson, Jieun Kiaer, Matthew Purver, Mehrnoosh Sadrzadeh, and Graham White. 2020a. Affordance competition in dialogue. In *Proceedings of the 24th SemDial*.

Eleni Gregoromichelaki, Christine Howes, Arash Eshghi, Ruth Kempson, Julian Hough, Mehrnoosh Sadrzadeh, Matthew Purver, and Gijs Wijnholds. 2019. Normativity, meaning plasticity, and the significance of Vector Space Semantics. In *Proceedings of the 23rd SemDial*.

Eleni Gregoromichelaki, Christine Howes, and Ruth Kempson. 2020b. Actionism in syntax and semantics. In *Dialogue and Perception DaP2018*, volume 2, pages 12–27.

Michael Haugh. 2008. Intention and diverging interpretings of implicature in the "uncovered meat" sermon. 5(2):201–228.

Michael Haugh and Yasuko Obana. 2015. Transformative continuations, (dis)affiliation, and accountability in Japanese interaction. *Text & Talk*, 35(5):597–619.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. 2018a. Running Repairs. *Topics in Cognitive Science*, 10(2):367–388.

Patrick G. T. Healey, Jan Peter de Ruiter, and Gregory J. Mills. 2018b. Editors' introduction: Miscommunication. *Topics in Cognitive Science*, 10(2).

Bert H. Hodges. 2022. Values Define Agency. *Adaptive Behavior*, page 10597123221076876.

Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Julian Hough, Lorenzo Jamone, David Schlangen, Guillaume Walck, and Robert Haschke. 2020. A types-as-classifiers approach to human-robot interaction for continuous structured state classification. *CLASP Papers in Computational Linguistics*, 2:28–40.

Julian Hough and Matthew Purver. 2012. Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the 16th SemDial (SeineDial)*, pages 136–144.

Christine Howes. 2012. *Coordination in Dialogue: Using Compound Contributions to Join a Party*. Ph.D. thesis, Queen Mary University of London.

Christine Howes and Arash Eshghi. 2017. Feedback relevance spaces. In *IWCS 2017*.

Christine Howes and Arash Eshghi. 2021. Feedback Relevance Spaces: Interactional Constraints on Processing Contexts in Dynamic Syntax. *Journal of Logic, Language and Information*, 30(2):331–362.

Christine Howes, Matthew Purver, Patrick G. T. Healey, Gregory J. Mills, and Eleni Gregoromichelaki. 2011. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311.

Edwin Hutchins. 1995. *Cognition in the Wild*. MIT Press.

Ruth Kempson, Ronnie Cann, Arash Eshghi, Eleni Gregoromichelaki, and Matthew Purver. 2015. Ellipsis. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*. Wiley-Blackwell.

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

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax*. Blackwell.

Julian Kiverstein, Michael D. Kirchhoff, and Tom Froese. 2022. The Problem of Meaning: The Free Energy Principle and Artificial Agency. *Frontiers in Neurorobotics*, 16:844773.

Paul Kockelman. 2011. Biosemiosis, technocognition, and sociogenesisy. *Current Anthropology*, 52(5):711–739.

Stefan Kopp and Nicole Krämer. 2021. Revisiting Human-Agent Communication. *Frontiers in Psychology*, 12.

Richard Kunert, Raquel Fernández, and Willem Zuidema. 2011. Adaptation in child directed speech. In *Proc of SemDial (LosAngelogue)*, pages 112–119.

Shalom Lappin. 2021. *Deep Learning and Linguistic Representation*. Chapman and Hall.

Federico Laudisa and Carlo Rovelli. 2021. Relational Quantum Mechanics. In *The Stanford Encyclopedia of Philosophy*, Winter 2021 edition. Stanford University.

Gene H. Lerner. 1991. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458.

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

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? In *Proceedings of EMNLP*, pages 2443–2453.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't Say That! In *Proceedings of the 58th ACL*, pages 4715–4728.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Procs of the 8th NAACL*, pages 733–743.

Per Linell. 2009. *Rethinking Language, Mind, and World Dialogically*. IAP.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xin Lu, Barbara Di Eugenio, Trina C. Kershaw, Stellan Ohlsson, and Andrew Corrigan-Halpern. 2007. Expert vs. Non-expert Tutoring. In *Computational Linguistics and Intelligent Text Processing*, pages 456–467. Springer.

Ewa Luger and Abigail Sellen. 2016. "like having a really bad pa". In *Procs of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297.

Riccardo Manzotti and Antonio Chella. 2018. Good old-fashioned artificial consciousness and the intermediate level fallacy. *Frontiers in Robotics and AI*, 5:39.

Gary Marcus. 2022. Horse rides astronaut. "`https://garymarcus.substack.com/p/horse-⊂rides-⊂astronaut?s=r`".

M. Mirolli and D. Parisi. 2011. Towards a Vygotskyan cognitive robotics. *New Ideas in Psychology*, 23(3):298–311.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2nd online edition.

Roger K. Moore. 2017. Is Spoken Language All-or-Nothing? In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots*, pages 281–291. Springer.

Alva Noë. 2004. *Action in Perception*. MIT press.

Pietro Perconti and Alessio Plebe. 2019. Deep learning and embodiment. In *AIC*, pages 10–21.

Martin Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *BBS*, 27:169–226.

Alison Pilnick and Robert Dingwall. 2011. On the remarkable persistence of asymmetry in doctor/patient interaction. *Social science & medicine*, 72(8):1374–1382.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of IWCS*, pages 365–369, Oxford, UK.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the 'I's and crossing the 'you's. In *Pros of the 14th SemDial*.

Matthew Purver, Mehrnoosh Sadrzadeh, Ruth Kempson, Gijs Wijnholds, and Julian Hough. 2021. Incremental composition in distributional semantics. *J. Log. Lang. Inf.*, 30(2):379–406.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied Human Computer Interaction. *KI - Künstliche Intelligenz*, 35(3):307–327.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Scott Reed et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Erik Rietveld, Damiaan Denys, and Maarten Van Westen. 2018. Ecological-enactive cognition as engaging with a field of relevant affordances. In *The Oxford handbook of 4E cognition*, volume 41. OUP.

Joanna Rączaszek-Leonardi, Agnieszka Dębska, and Adam Sochanowicz. 2014. Pooling the ground. *Frontiers in Psychology*, 5:1233.

Harvey Sacks, Emmanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Chitwan et al Saharia. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Yo Sato. 2011. Local ambiguity, search strategies and parsing in Dynamic Syntax. In E. Gregoromichelaki, R. Kempson, and C. Howes, editors, *The Dynamics of Lexical Interfaces*. CSLI Publications, Stanford, CA.

Johanna Seibt. 2018. Ontological tools for the process turn in biology. In Nicholson and Dupré, editors, *Everything flows*, page 113. OUP.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. *arXiv:1503.02364 [cs]*.

Brian Skyrms. 2010. *Signals*. OUP Oxford.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. *arXiv:1506.06714 [cs]*.

Dan Sperber and Deirdre Wilson. 1995. *Relevance*. Blackwell.

Sune Vork Steffensen, Frédéric Vallée-Tourangeau, and Gaëlle Vallée-Tourangeau. 2016. Cognitive events in a problem-solving task. *Journal of Cognitive Psychology*, 28(1):79–105.

Hao Tan and Mohit Bansal. 2019. LXMERT. In *Proceedings of EMNLP-IJCNLP*, pages 5100–5111.

Michael T Turvey. 1992. Affordances and prospective control. *Ecological psychology*, 4(3):173–187.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv*, (1506.05869v3).

Thomas Wolf et al. 2019. TransferTransfo. *arXiv 1901.08149*.

Wlodek W. Zadrozny. 2020. Towards Coinductive Models for Natural Language Understanding. *arXiv:2012.05715*.

Gesture and Part-of-Speech Alignment in Dialogues

Zohreh Khosrobeigi, Maria Koutsombogera, Carl Vogel
Computational Linguistics Group
Trinity College, the University of Dublin
Dublin 2, Ireland
{khosrobz, koutsomm, vogel}@tcd.ie

Abstract

This paper studies the relation between language and gesture in interaction by investigating the temporal alignment of gestures and the words they co-occur with in a corpus of task-based dialogues. Specifically, we examine the relationship of different semiotic gesture types (their quantity and duration) with the syntactic categories assigned to the words that coincide with the gesture. We observe that different types of gesture-word alignment present different distributions, as well as different associations depending on the gesture type.

1 Introduction

We seek to understand the interfaces between gesture and language. It has been suggested that gesture accompanying linguistic content has functions tied more closely to speakers' formulation of uttered thoughts than to facilitating hearers' grasp of a shared interpretation of expressed statements (Hadar, 1989; Tuite, 1993; McNeill, 1997; Kita, 2000). Evidence for the perspective is in the fact that people may frequently be witnessed gesturing while speaking without a visual modality of communication, via a telephone, for example, even if those gestures are reported to be smaller in extent and shorter in duration than those occurring during visual contact (Bavelas et al., 2008). Other aspects of language as used in communication are marked by economizing production effort: an introduction of an entity into discourse may be initially accompanied by adjective and prepositional phrases and relative clauses, but quickly after introduction, relegated to pronouns. In contrast, gesture consumes rather more physical energy than uttering overt nominals, and may be seen at all stages of a conversation.[1] However, gesture does not appear to be randomly distributed. Further, even if gesture is not performed primarily for the benefit of a listener, a listener may interpret gesture, partly on the basis of how they gesture themselves (to the extent that they are aware of how they gesture themselves).

Public gesture has systematic properties that enable consensus on the classification of a small number of semiotic gesture types – Beats, Symbolics, Iconics, and Deictics, for example, and further, those other movements are none of those.[2] Some research addresses the relationship between gesture and intonation (Steedman, 1991; Loehr, 2004; Jannedy and Mendoza-Denton, 2005; Loehr, 2012), and some work on gesture semiotic type and word meaning relation (Bernardis and Gentilucci, 2006; De Marco et al., 2015). While information structure and intonation outlines conform in a manner somewhat at odds with syntactic constituent structure, word-level categories have been modeled as carrying the relevant information (Steedman, 1991). Further, it has been found that there are strong asymmetric links between gesture,[3] using a more fine-grained set of gesture types than described above, and part of speech categories (Mehler et al., 2012), but without reporting on the effects of individual gesture types and part of speech types. These are among reasons it is interesting to explore word categories, starting with the part of speech categories that are addressed in this study.

Here we explore whether there are systematic facts in the distribution of gesture types in collab-

[1] We take it to be self-evident that moving arms, etc. requires more energy than the movements required by speech. However, for indirect support of this claim we not past work (Pouw et al., 2020) that shows greater variation in F0 and amplitude in vowel expression with gesture than without; thus, speech and gesture requires more energy than speech without gesture (the work cited explores the theory that synchrony of speech and gesture is more of a mechanical process than a cognitive one).

[2] Curiously, a "miscellaneous" movement may still contain private, and public meaning: scratching an eyebrow may be what one person does while puzzling on something, and what another does only when conversations are lively and the agent is confident it will not be noticed.

[3] They see stronger evidence of gestures being selected by words than vice versa.

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

orative conversations in relation to the syntactic categories of words that are nearby. One could imagine that Deictics mainly occur with nominals, Symbolics mainly with verbals, and Iconics split between those categories. However, it is necessary to be precise about what "with" means. MacNeill's hypothesis would suggest that gesture onset is typically prior to the accompanying word onset but with some extent of co-temporality.

The next section describes related work. Section 3 reports on the datasets used. Section 4 explains the methods followed for forced alignment, text-gesture alignment as well as POS tagging. The data profiling according to the alignment categories is reported in Section 5 and analyzed and discussed further in Section 6, while Section 7 presents the conclusions.

2 Related Work

Co-speech gestures are visible hand actions that are produced while speaking. Several views have been reported in the literature with respect to the role of gestures in communication, i.e., as an indispensable part of the language system (McNeill, 1992, 2005; Kendon, 2004) or the overall perspective of speaking as a multimodal construct (Cienki and Muller, 2008). There is a large amount of research surrounding the relationship between gesture and speech. This relationship can be viewed from its cognitive perspective, namely how gestures are linked to our thought (McNeill, 2005) as well as from its communicative perspective, i.e., how gestures regulate the organization of the interaction (Kendon, 2004). Theoretical research suggests that speech and gesture share a common conceptual origin and that they form a single integrated system (McNeill, 1992; McNeill and Duncan, 2000; Rieser, 2015); and that both speech and gesture have communicative functions that come from the same communicative intention (de Ruiter, 2000).

In this way, hand gestures help speakers talk, think, and disclose information that cannot be verbalized (Goldin-Meadow et al., 1993); and at the same time, performing gestures helps speakers organize visuo-spatial information into units that are compatible with the format of speech (Kita and Özyürek, 2003; Hahn and Rieser, 2010).

We study the temporal relation between gesture and speech seeking insight into the nature of their links. Words that match most closely the meaning of a gesture have been termed lexical affiliates (Schegloff, 1985). McNeill (1992) defined three rules of synchronization between gesture and speech, namely the phonological synchrony rule, predicting that a gesture stroke should occur before the most prominent syllable; the semantic synchrony rule predicting that co-occurring gestures and speech relate to the same idea unit; and the pragmatic synchrony rule predicting that co-occurring gestures and speech have the same pragmatic function (Wagner et al., 2014). While other works in this space address gesture morphology (Hahn and Rieser, 2010; Rieser and Lawler, 2020), we focus on the semiotic type of gestures without reference to gesture-internal phases.

The present study investigates the dependencies of gestures and the grammatical categories (part-of-speech - POS) of the words co-occurring with gestures. We study this in the totality of the gestures occurring in a multimodal corpus, and we consider gestures of all semiotic types. We use the manual transcripts of the corpus dialogues as well as existing manual annotations of gestures of dialog participants. To temporally identify lexical correlates, we use timestamps from manual word and gesture annotation. The dialog transcripts were further annotated automatically with POS tags.

3 Resources

3.1 Dataset

To study the distribution of the gesture types co-occurring with syntactic categories of words, we use the MULTISIMO corpus (Koutsombogera and Vogel, 2018), a multimodal dataset of three-party, task-based dialogues which were collected to investigate different aspects of collaborative interaction. The dataset consists of 18 dialogue sessions. In each session, two players collaborate with each other in English to answer three questions and rank the answers and are aided by a moderator who provides guidance. The dataset includes the video and audio of the dialogues, as well as a set of annotations including speech transcripts, gaze, laughter, and gesture annotations. There are 39 dialogue participants, 16 of which are native English speakers.

Hand gestures were manually annotated using the ELAN editor (Brugman and Russel, 2004). The entire duration of a gesture was annotated, i.e. the transition from a neutral position to the gesture as well as the return to the neutral position once the gesture concluded; that is, including preparation, stroke and retraction, as well as gesture holds, if

applicable. Also, the annotation scheme does not distinguish between the use of one or two hands performing the gesture. Once a gesture is visible, the start and end time of that gesture is marked and assigned with one of the following semiotic types: Beat, Iconic, and Deictic and Symbolic. The semiotic categories defined in the annotation scheme are based on McNeill (McNeill, 1992), who, in turn, built on Peirce's semiotic types (Peirce, 1931).

Beat gestures are utilized in rhythm with utterances in order to emphasize what is being said or to improve the coherence of the statement for the listener. *Iconic* gestures provide a pictorial representation of any concrete or abstract concept, e.g. moving the hand upwards or downwards while ranking the answers. *Deictic* gestures are usually depicted by pointing at a particular object or individual, whether they are real or imaginary. They are commonly used in the corpus by one participant to point at another participant, as if to encourage a contribution to the discourse. *Symbolic* gestures are culture-specific gestures where the relation between their shape and the accompanying speech is based on social conventions, such as the thumbs up gesture (to denote agreement) or the OK symbol.

In addition to those four types, the label N/A was used for visible hand movements, which, however, did not have a communicative function in the dialogue. Gestures were annotated by one rater and the annotation was validated by a subject matter expert, who had frequent interaction with the rater to monitor the task and to discuss, among others, difficult or ambiguous cases.

Apart from the data described above, we extracted some additional features and data from MULTISIMO: In order to analyze the gestures along with the rest of the data, at first, gesture timestamps of each player were exported from ELAN. Then, the files were pre-processed to keep the information related to the onset, offset and type of gestures. The duration of gestures, of 14 out of 18 dialogues, (mean = 1573 msec) is approximately five times greater than the duration of words (mean = 300 msec), and there are fewer than 50 gesture tokens in each session. The number of each semiotic gesture type (1004 instances of gestures), as well as the number of spoken word tokens, are counted in Table 1.

#Beat	#Iconic	#Deictic	#Symbolic	#N/A	#Word
374	251	64	15	300	12862

Table 1: Count of gestures per gesture type and count of spoken words of all dialogue players.

3.2 New Dataset for Text Alignment

An important aspect of dialogue analysis is understanding the factors that influence the alignment of the communication channels available – linguistic content, back-channels, social signals, laughter, gesture, gaze, and so on. All aspects convey meaning, although not always about the dialogue's linguistic content (sometimes, about the participants' level of engagement, sometimes about their personal relationships, etc.). Nonetheless, we take the linguistic content as the focal point of dialogue, and seek to understand alignment with respect to the linguistic form of that content. This entails requiring knowledge of the timing of the words spoken – temporal onset and offset for each item.

The onset and offset of linguistic content are important information when studying the relation of different channels in multimodal interaction. To identify the relation between audio and text in the dialogues, we use two streams of information: the audio channel of participants' speech and the transcript of the dialogues, performed at an utterance level. To be able to align the audio and the transcript at a word level, we labeled the onset and offset of each word in speech manually using the transcripts and monophonic audio files for each speaker with the Praat software (Boersma and van Heuven, 2001). Through Praat, the audio files and their corresponding transcripts are processed to define the start and end of each word. The output CSV files include the onset and offset of words and the words. The text alignment was done for the participants that had the player role, for 14 out of the 18 corpus dialogues.[4] Each dialogue needed about 8 hours to label.

4 Methods

Our research aim here is to identify the major syntactic category that is used most frequently during, before, or after a hand gesture occurrence. We approach this using word-level rather than phrasal

[4]The alignment of the remaining four dialogues is currently in progress. The aligned CSV files are available from the MULTISIMO website (`http://multisimo.eu/datasets.html`).

constituent-level labeling of POS. To answer this question, we first performed temporal text and gesture alignment. Then, each word was labeled with its POS tag. Each of these steps is explained below.

4.1 Temporal Gesture-Word Alignment

Players perform gestures while speaking. The gestures may be short, long, or located in any part of an utterance. They can be semantically related to words that precede, follow, or are uttered simultaneously with the gesture. Using text alignment and gestures alignment data, an alignment of gestures and words is possible at different times of occurrence in relation to each other. To align gestures and words, the occurrence of spoken words was computed in rather than gestures at various times of happening. All possibilities of occurrence are seven categories. Table 2 encapsulates the explanation of each temporal gesture-word alignment and shows a graphic view for each alignment.

4.2 Part of Speech Tagging

NLTK (Bird et al., 2009) and TreeTagger (Schmid, 1994) are used to categorize words with POS labels. Applied to this data, the systems differ in many words. NLTK tags our dataset using 26 different POS labels, of which the most used is a noun. For instance, NLTK tags the verb "think" as "Noun" and the adjective "dirty" as "Noun". On the other hand, TreeTagger tags the dataset using 51 different POS labels. TreeTagger tags the above examples correctly. The reason for having different tags is that taggers consider different types for each POS. For instance, TreeTagger has several types of a verb, such as VBZ, VBB, VBI, and VBG. For our purposes and also given the relatively small size of the dataset at hand, broader syntactic category labels seem more appropriate. As a result, similar categories are mapped to one main category. Table 3 illustrates the mapping from TreeTagger POS categories to more general category labels. In total, 3807 tags are mapped to eight categories using TreeTagger (including words and non-word vocalizations). We tag non-word vocalizations (e.g "hmm", laughter, etc.) as NW.

Taggers work with high accuracy on well-structured and standard texts. But in natural dialogue, people do not talk solely in complete grammatical sentences – sometimes, utterances are sentence fragments or ungrammatical. As a result, we tokenized MULTISIMO transcripts, and fed each word in succession to taggers as input. Moreover, we did not normalize tokens since normalization to lemmas can confuse automatic syntactic labeling.

Table 4 shows the number of POS instances at different alignments using TreeTagger.

To see the reliability of NLTK and TreeTagger, their error rates are estimated on the basis of 183 randomly selected items. For NLTK, 46 of 183 tags are incorrect, and the error rate is 25.12%. For TreeTagger one of 183 tags is incorrect, and the error rate is 0.5%. Considering these error rates, only TreeTagger labellings are analyzed.

When a player performs a gesture and then says a word, some POS types are used more than others. Table 5 illustrates the most used POS for each type of gesture at different alignments.

4.3 Durations

With respect to gestures and the words with which they align, it is interesting to examine durations, not least because these include aspects of production time and execution time. Table 6 shows the central tendencies of durations for each gestural types.

5 Data profile

Each word instance participates in at most two alignment categories, falling into more than one category if the token duration overlaps with the duration of successive gestures. Of 12862 words spoken by the players, 9055 words do not align with any gesture (see Table 7 for the distribution of POS categories for these words). Also, 3707 words align with exactly one gesture, and 100 words align with two gestures and enter into two alignment categories. Table 7 illustrates the distribution of POS categories for unaligned words, words aligned with one gesture and with two gestures.

Gesture instances can also enter into more than one alignment category, for example, gest-word-with-overlap and word-gest-with-overlap, when a gesture happens with two different words at different intervals. Of 1004 gestures, 290 gestures enter exactly one category; 714 gestures enter more than one.

We emphasize that the relative frequency of these alignment categories in natural dialogue are not given, *a priori*. The fact that 714 gesture tokens and 100 words are in more than one alignment category necessitates that the instances analyzed in terms of their counts in the contingency tables below be *pairs* of gesture tokens and word tokens. Each pair is independent. For the three alignment

Alignment	Description	Pictures
short-gest	The duration of a gesture is shorter than the duration of a word and occurs within the word duration; hence, it includes only one word. The onset and offset of the gesture are inside the word timestamp.	
long-gest	The duration of a gesture is longer than the duration of a word. As a result, the gesture occurs with a few words simultaneously. The longest gesture in the dataset co-occurs with four words. The onset and offset of words are inside the gesture timestamp.	
gest-word-no-overlap	A gesture occurs before a word, and when the gesture is completed, the word is uttered. The offset of gesture is before the onset of word. The distance between the offset of gesture and onset of word is less than one millisecond.	
word-gest-no-overlap	A gesture starts immediately as soon as a word is finished. The onset of a gesture is after the offset of a word. The distance between the onset of gesture and offset of word is less than one millisecond.	
gest-word-with-overlap	A gesture starts before the beginning of a word, and it ends before that word ends. The offset of gesture is inside the word onset and offset.	
word-gest-with-overlap	A gesture starts in the middle of a word and finishes after the word. The onset of gesture is inside the word onset and offset.	
silent-gest	A person gestures without speaking.	

Table 2: Types of temporal gesture-word alignment.

Main	Abbr.	#Tags	Mapped Tags
Noun	NN	658	NN2=106, NN1=510, NN0=28, NP0=14
Verb	VRB	768	VBB=434, VBI=36, VBZ=130, VM0=79, VBD=40, VBG=29, VBN=20
Adjective	AJ	137	AJ0=134, AJC=1, AJS=2
Adverb	ADV	478	AV0=389, XX0=61, AVQ=21, AVP=7,
Determiner / Pronoun	DP	903	DT0=112, AT0=245, DTQ=25, CRD=48, ORD=39, EX0=16, DPS=24, PNI=13, PNQ=2, PNX=3, PNP=376
Preposition	PRP	195	PRP=145, PRF=46, TO0=4
Conjunction	CJ	182	CJS=46, CJC=134, CJT=2
Interjection	IJ	168	ITJ=168
Non-word	NW	315	NW= 315
Sum		3807	

Table 3: Main POS categories and mapped subcategories aligned to eight categories using TreeTagger.

Alignment	ADV	AJ	CJ	DP	IJ	NN	NW	PRP	VRB
short-gest	1	0	0	0	1	2	6	0	0
long-gest	344	97	145	745	115	398	144	171	619
gest-word-no-overlap	0	1	0	0	0	1	0	0	1
word-gest-no-overlap	1	0	1	0	0	0	0	0	0
gest-word-with-overlap	63	22	14	93	30	165	76	12	85
word-gest-with-overlap	69	17	22	67	22	92	89	12	63
Sum	478	137	182	905	168	658	315	195	768

Table 4: Counts of POS instances for each alignment.

categories for which the total number of observations exceeds 45, we construct a contingency table analysis to test whether there is an interaction between the alignment category and the part of speech distribution; the interaction is significant

Alignment	Beat	Iconic	Deictic	Symbolic	N/A
short-gest	NN=2 NW=2 ADV=1	NW=2			IJ=1 NW=2
long-gest	DP=295 VRB=264 NN=153 ADV=151	DP=284 VRB=221 NN=171	DP=73 VRB=52	VRB=7 DP=5 IJ=6	DP=88 VRB=75 ADV=56
gest-word-no-overlap	VRB=1 NN=1				AJ=1
word-gest-no-overlap	CJ=1	ADV=1			
gest-word-with-overlap	NN=74 DP=46 VRB=44	NN=60 DP=29 VRB=23	NN=18 VRB=9 DP=9	VRB, ADV, IJ=2	NW=26 NN=14 ADV=13
word-gest-with-overlap	NN=40 ADV=35 VRB=37	NN=27 DP=23 NW=22	NN=10 DP=6	NW=3	NN=14 ADV=10 VRB=10 NW=24

Table 5: The most used POS with each type of gesture at different alignments using TreeTagger.

Duration	Beat	Iconic	Deictic	Symbolic	N/A
Word mean	328.7	301.1	266.3	327.9	387.5
Word median	255.2	244.8	220.5	300.1	302.6
Word s.d.	489.0	222.1	170.7	157.5	372.0
Gest. mean	1687.6	2168.4	1609.6	1603.1	2136.0
Gest. median	1513.5	2040.0	1450.0	1640.0	1910.0
Gest. s.d.	891.8	985.3	719.8	643.8	1177.2

Table 6: Word and gesture millisecond duration statistics for aligned gesture-word pairs.

Aligned and unaligned words			
	Aligned words in one category (distinct)	**Aligned words in two categories**	**Unaligned words**
NN	627	32	1506
VRB	755	13	1562
AJ	132	5	298
ADV	470	8	1205
DP	891	14	1703
PRP	194	1	300
CJ	180	2	334
IJ	167	1	1032
NW	291	24	1115
SUM	3707	100	9055

Table 7: The distribution of POS for words which are aligned or are not aligned.

($\chi^2 = 281.66$, df $= 16$, $p < 2.2^{-16}$).

Analysis of the Pearson residuals reveals: for long-gest alignments, there are significantly more DP ($p < 0.05$), PRP ($p < 0.05$), and VRB ($p < 0.05$) observations and significantly fewer NN ($p < 0.05$), and NW ($p < 0.001$) observations than would be expected with no interaction; for gest-word-with-overlap alignments, significantly more NN ($p < 0.001$), and NW ($p < 0.001$) observations and significantly fewer CJ ($p < 0.05$), DP ($p < 0.05$), PRP ($p < 0.05$), and VRB ($p < 0.05$) observations than would be epxected with no interaction; for word-gest-with-overlap alignments, significantly more NW ($p < 0.001$) and significantly fewer DP ($p < 0.05$), PRP ($p < 0.05$) and VRB ($p < 0.05$) observations than would be expected with no interaction. Thus, there appears to be an interaction between the starting point and span of a gesture and the accompanying parts of speech – DP, PRP and VRB categories are prominent in long-duration gestures; NN and NW are prominent in shorter duration gestures, with NN being most prominent for the short gestures commencing before and ending during the aligned word (gest-word-with-overlap). Considering the token durations, note from Table 8, that for long-duration gestures, the categories significant for the extent of positive observations (DP, PRP and VRB) are also the shortest in duration for that alignment category. For short gestures commencing before and ending during the aligned word the most numerous category (NN) is the second longest in duration for the alignment category. For short gestures commencing in the middle of a token and ending after it, the categories significant in the extent of their positive count (NN and NW) are the longest in duration for the alignment category. Thus, significant counts are not always explained by shorter durations.

short-gest, gest-word-no-overlap, word-gest-no-overlap Alignments: Of seven alignment categories, three categories, short-gest Alignment (gestures are short and only include one word), gest-word-no-overlap Alignment (a player gestures and then says a word after the gesture), and word-gest-no-overlap Alignment (a player gestures after finishing a word) have a few instances of POS cate-

	Alignment					
	long-gest		gest word		gest word	
POS	mean	median	mean	median	mean	median
ADV	259.0	235.1	408.1	379.1	343.1	309.2
AJ	353.7	322.6	479.2	478.3	464.9	401.9
CJ	219.0	185.2	**320.4**	**270.6**	305.3	230.9
DP	166.6	140.4	**241.6**	**193.3**	**262.8**	**217.1**
IJ	281.4	270.8	387.2	346.7	255.1	254.9
NN	**400.5**	**382.2**	515.4	494.3	673.2	491.7
NW	**471.1**	**422.1**	717.5	622.5	849.4	663.6
PRP	166.5	153.4	**264.4**	**271.9**	**314.1**	**289.2**
VRB	198.1	171.0	**432.1**	**309.2**	**311.8**	**277.1**

Table 8: The millisecond *durations* of POS instances for frequent alignments. To save space gest word abbreviates gest-word-with-overlap, and gest word abbreviates word-gest-with-overlap. The figures that are boxed or in bold indicate the cells of the contingency of *counts* for which Pearson residuals were significant ($p < 0.05$), as described in the text: durations in cells are boxed where the corresponding count was significantly *greater* than would be expected with no interaction between the part of speech and alignment type; durations in cells are **bold** where the corresponding count was significantly *less* than would be expected with no interaction between part of speech and the alignment type.

gories aligned with gestures (see Table 5).

long-gest Alignment: In long-gest alignment, a gesture is long and co-occurs with a few words. Beat gestures are accompanied by 1092 words, and Iconics by 1037 words. Verb and DP are the most used POS categories in this temporal alignment (see Table 5). Table 9 shows the distribution of POS categories across gesture types.

long-gest Alignment					
POS	**Beat**	**Iconic**	**Deictic**	**SYMBOLIC**	**N/A**
ADV	151	104	33	0	56
AJ	40	39	3	1	14
CJ	51	65	12	0	17
DP	295	284	73	5	88
IJ	31	32	9	6	37
NN	153	171	25	2	47
NW	39	58	4	2	41
PRP	68	63	21	3	16
VRB	264	221	52	7	75
Sum	1092	1037	232	26	391

Table 9: The count of POS categories in long-gest alignment by gesture type.

gest-word-with-overlap Alignment: A player gestures and starts a word in during the gesture and finishes the gesture before the word. The most accompanied gesture types are Beats (n=258) and Iconics (n=162). NN is the most used POS with Beat, Iconic, Deictic, and N/A (excluding NW as non-vocalized POS) (see Table 5). Table 10, shows the distribution of POS and gesture types.

gest-word-with-overlap Alignment					
POS	**Beat**	**Iconic**	**Deictic**	**Symbolic**	**N/A**
VRB	44	23	9	2	7
NN	74	60	18	0	14
PRP	8	4	0	0	0
DP	46	29	9	0	9
AJ	8	6	2	0	6
ADV	31	12	5	2	13
CJ	6	5	1	0	2
IJ	8	8	3	2	9
NW	33	15	2	0	26
Sum	258	162	49	6	86

Table 10: The distribution of POS categories in gest-word-with-overlap alignment by gesture type.

word-gest-with-overlap Alignment: Gestures that start in the middle of a word and finish after the word are categorized as word-gest-with-overlap. The most accompanied gestures in this alignment are Beat (n=208), and Iconic (n=125). NN is used the most in the alignment (excluding Symbolic and non-vocalize POS) (see Table 5). Table 11 shows the POS-gesture types distribution.

word-gest-with-overlap Alignment					
POS	**Beat**	**Iconic**	**Deictic**	**Symbolic**	**N/A**
VRB	37	13	3	0	10
NN	40	27	10	1	14
PRP	3	4	2	0	3
DP	29	23	6	0	9
AJ	12	2	0	0	3
ADV	35	19	4	1	10
CJ	7	9	2	0	4
IJ	10	6	2	0	4
NW	35	22	5	3	24
Sum	208	125	34	5	81

Table 11: The distribution of POS categories in word-gest-with-overlap alignment by gesture type.

silent-gest Alignment: The last category is silent-gesture alignment, in which a person commences and completes a gesture without accompanying vocalization. There are 128 such gestures, and miscellaneous movement (N/A) is the most frequent type (n=109). Table 12 shows all gesture type counts, and Table 13 indicates durations.

6 Results and Discussion

One might hypothesize that the extended gesture duration indicates cognitive processing. It is reasonable to theorize that Beats are used in a man-

Counts: gest-silent					
SUM	**Beat**	**Iconic**	**Deictic**	**Symbolic**	**N/A**
128	11	6	1	1	109

Table 12: The distribution of gestures accompany silence, by gesture type.

Durations: gest-silent					
	Beat	**Iconic**	**Deictic**	**Symbolic**	**N/A**
mean	803.6	1205.0	890.0	1200.0	1622.3
median	669.0	1020.0	890.0	1200.0	1450.0
sd	409.5	907.1	NA	NA	841.5

Table 13: Duration (milliseconds) statistics of gestures accompanying silence, by gesture type.

ner that punctuates completely planned speech while Symbolics are used in support of forming the thought that is being spoken. One might therefore expect words spoken during Beats to take less time than those spoken through Symbolics. This overall contrast is not significant (Wilcox's $W = 2516, p = 0.08387$), but it is significant when restricting attention to verbs[5] (Wilcox's $W = 1027, p = 0.04075$).

Of the six gesture-word alignments, three (long-gest, gest-word-no-overlap, gest-word-with-overlap) involve a gesture commencing before the onset of an aligned word and two of these (the exception is gest-word-no-overlap) are among the three most frequent alignment types. The third frequent alignment type, we note below, does not have significant interactions with the count of aligned POS categories, but the other two frequent alignments do. The combination of gestures commencing before the aligned word and the interaction with the distribution of POS categories of those words are suggestive of a role of the gesture in the formulation of the unfolding speech. Beat gestures are used more than other gestures in players' conversations. Iconic, N/A, Deictic, and Symbolic are the next most used gestures, see Table 1. As Table 5 shows, short-gest, gest-word-no-overlap, and word-gest-no-overlap alignments are least frequently witnessed. The most frequently witnessed alignment is long-gest, followed by gest-word-with-overlap, and word-gest-with-overlap alignments.

Consider those gestures that have a duration that exceeds that of its first aligning word, inclusive of more words as well (long-gest). Figure 1 shows the Pearson residuals that result from the χ^2 analysis of the contingency table inherent in Table 9

[5]That is, we measure the contrast between Beat durations and Symbolic durations when accompanying verbs.

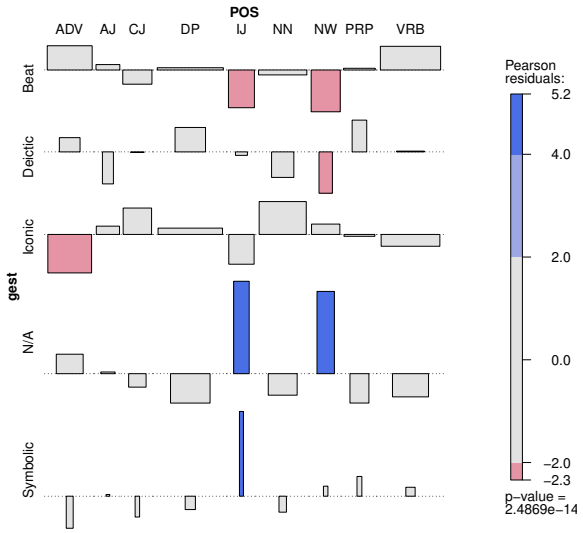


Figure 1: Residuals of interaction among gestures longer than nearby words (long-gest) and the syntactic categories of those words. The horizontal lines indicate the level of no interaction for a given row: upwards projections indicate counts in a cell that exceed what would be expected with no interaction; downwards projections indicate counts that are less than would be expected; shading indicates statistical significance ($p < 0.05$).

($\chi^2 = 134.34$, df $= 32$, $p = 2.487^{-14}$). In contrast to a null-hypothesis expectation of no interaction between gesture type and syntactic categories: Beats show a significant dearth with interjections and non-word vocalizations; Deictics show significant dearth with non-words; Iconics show significant dearth with adverbs; miscellaneous motions show significant co-occurrence with interjections and non-word vocalizations; Symbolics show significant co-occurrence with interjections (but we treat the effect of Symbolics with caution, given the low count of observations).

It is not surprising that Beats do not appear to be multi-modal exclamation marks for interjections or that Beats and Deictics are conspicuously missing from non-word vocalizations. It also makes sense for iconic gestures to neglect adverbs. It seems natural that miscellaneous motions accompany interjections and non-words.

Figure 2 shows the residuals of the χ^2 test of interaction between gesture types and parts of speech for the alignments in which a gesture starts before a word and ends in the middle of the word (gest-word-with-overlap; $\chi^2 = 72.526$, df $= 32$, $p = 5.572^{-5}$). In comparison with the distribution of counts that would be expected if there were no interaction be-

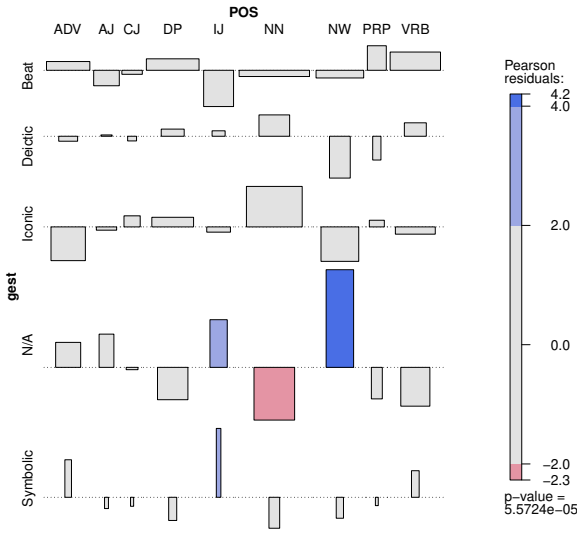


Figure 2: Residuals of interaction among gestures longer than nearby words (gest-word-with-overlap) and the syntactic categories of those words. The horizontal lines indicate the level of no interaction for a given row: upwards projections indicate counts in a cell that exceed what would be expected with no interaction; downwards projection idicate counts that are less than would be expected; shading indicates statistical significance ($p < 0.05$).

tween gesture type and part of speech, there are more non-words and interjections and fewer nominals aligned with miscellaneous movements; more Symbolics aligned with interjections.

Table 4 illustrates determiners followed by verbs and nouns are used the most around the gestures which the highest ones occur in longgest, gest-word-with-overlap and word-gest-with-overlap alignments. The interactions between gesture category and linguistic categorization of vocalizations are not significant for alignments in which the gesture starts in the middle of a word and ends afterwards (word-gest-with-overlap) – $\chi^2 = 31.955$, df $= 32$, $p = 0.47$.

Of the alignments for which there were sufficient interactions to meaningfully analyze the interaction between semiotic types and part of accompanying parts of speech, two demonstrated statistically significant interactions, and in both of those, the gesture onset preceded the linguistic content onset. The primary effects for the contentful semiotic types (i.e. not miscellaneous movements) was in a relative lack of gestures accompanying certain syntactic categories, but without systematic sensitivity to the whether the category is mainly populated by open-class or closed-class subcategories. Certain interesting trends are visible (e.g. beats occuring with relational categories; iconics with the nominal domain; deictics with nouns and verbs) but not statistically significant. While this work uses a more general typology of gesture types than (Mehler et al., 2012), we see more detail about where the relationships between gesture types and part of speech categories carry strong associations.

7 Conclusions

We have presented our observations of the counts and durations of gestures aligned with major syntactic categories assigned to vocalizations that accompany them, given a small number of possible alignment types. We think that the type of alignment (e.g., gesture onset prior to accompanying word onset) is revealing aspects of cognitive processing associated with the unfolding utterance. Of course, observations of different sorts than we have reported here would also be useful, but the alignments provided here will enable hypothesis testing regarding the interactions of gestures, syntactic categories, and their alignments. Of the six considered gesture-word alignment types, three are more frequently witnessed than the others, and within one of those, where gestures have a long duration from an onset before the first aligned word, there is noteworthy dearth of interjections and non-words with Beats, non-words with Deictics, adverbs with iconics; there is noteworthy coincidence of miscellaneous movement and interjections and non-words and interjections and Symbolics.

While the observations reported here are anchored in a single multi-modal dialogue corpus, the corpus involves distinct dialogues among a number of interlocutors, the dialogue settings do not impose particular constraints on gestures or part of speech categories. We intend to continue to explore gesture and word alignmeents in this and other multi-modal dialogue corpora.

Acknowledgments

This work was conducted with the financial support of Science Foundation Ireland under Grant No. 18/CRT/6223 and the GEHM research network (Independent Research Fund Denmark grant 9055-00004B). We are grateful to the anonymous reviewers who provided helpful constructive feedback on an earlier draft of this work.

References

Janet Bavelas, Jennifer Gerwing, Chantelle Sutton, and Danielle Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2):495–520.

Paolo Bernardis and Maurizio Gentilucci. 2006. Speech and gesture share the same communication system. *Neuropsychologia*, 44(2):178–190.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Paul Boersma and Vincent van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International*, 5(9/10):341–347.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Alan Cienki and Cornelia Muller. 2008. Metaphor, gesture, and thought. In Raymond W.Editor Gibbs, Jr., editor, *The Cambridge Handbook of Metaphor and Thought*, Cambridge Handbooks in Psychology, pages 483–501. Cambridge University Press.

Doriana De Marco, Elisa De Stefani, and Maurizio Gentilucci. 2015. Gesture and word analysis: the same or different processes? *NeuroImage*, 117:375–385.

Susan Goldin-Meadow, Martha Wagner Alibali, and R. Breckinridge Church. 1993. Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, 100(2):279–297.

Uri Hadar. 1989. Two types of gesture and their role in speech production. *Journal of Language and Social Psychology*, 8(3-4):221–228.

Florian Hahn and Hannes Rieser. 2010. Explaining speech gesture alignment in MM dialogue using gesture typology. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, pages 99–111, Poznań, Poland. SEMDIAL.

Stefanie Jannedy and Norma Mendoza-Denton. 2005. Structuring information through gesture and intonation. *Interdisciplinary studies on information structure: ISIS; working papers of the SFB 632*, (3):199–244.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Sotaro Kita. 2000. How representational gestures help speaking. In David McNeill, editor, *Language and Gesture*, pages 162–85. Cambridge University Press.

Sotaro Kita and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1):16–32.

Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2945–2951, Paris, France. European Language Resources Association (ELRA).

Daniel P Loehr. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory phonology*, 3(1):71–89.

D.P. Loehr. 2004. *Gesture and Intonation*. Georgetown University. PhD Thesis.

David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.

David McNeill. 1997. Growth points cross-linguistically. In Jan Nuyts and Eric Pederson, editors, *Language and Conceptualization*, 1 edition, pages 190–212. Cambridge University Press.

David McNeill. 2005. *Gesture and thought*. University of Chicago Press, Chicago.

David McNeill and Susan D. Duncan. 2000. Growth points in thinking-for-speaking. In DavidEditor McNeill, editor, *Language and Gesture*, Language Culture and Cognition, pages 141–161. Cambridge University Press.

Alexander Mehler, Andy Lücking, and Peter Menke. 2012. Assessing cognitive alignment in different types of dialog by means of a network model. *Neural networks*, 32:159–164.

Charles S. Peirce. 1931. *The Collected Papers of Charles Sanders Peirce, Vol. I: The Principles of Philosophy*. Harvard University Press, Cambridge.

Wim Pouw, Steven J Harrison, and James A Dixon. 2020. Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General*, 149(2):391.

Hannes Rieser. 2015. When hands talk to mouth. gesture and speech as autonomous communicating processes. In *SEMDIAL 2015 goDIAL: Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 122–130.

Hannes Rieser and Insa Lawler. 2020. Multi-modal meaning – an empirically-founded process algebra approach. *Semantics & Pragmatics*, 13:1–55.

Jan Peter de Ruiter. 2000. The production of gesture and speech. In David McNeill, editor, *Language and Gesture*, Language Culture and Cognition, pages 284–311. Cambridge University Press.

Emanuel A. Schegloff. 1985. On some gestures' relation to talk. In J. MaxwellEditor Atkinson, editor, *Structures of Social Action*, Studies in Emotion and Social Interaction, pages 266–296. Cambridge University Press.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Mark Steedman. 1991. Structure and intonation. *Language*, 67(2):260–296.

Kevin Tuite. 1993. The production of gesture. *Semiotica*, 93(1/2):83–106.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.

Coordinating taxonomical and observational meaning: The case of genus-differentia definitions

Bill Noble **Staffan Larsson** **Robin Cooper**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{bill.noble@, sl@ling, cooper@ling}.gu.se

Abstract

Genus-differentia definitions exhibit the dual nature of lexical semantic meaning—they incorporate both "hard" *X is a Y* relations between words, as well as "soft" aspects of meaning which can be supported or challenged by observation. Modeling such definitions as contributions in dialogue requires that we accommodate the fluidity of linguistic resources, while respecting the dual nature of the relations that hold between lexical items. In this paper, we use a Probabilistic Type Theory with Records (ProbTTR) to characterise genus-differentia definitions by describing the update they license to the common ground of a dialogue.

Metalinguistic dialogue is one way for speakers to align on the meaning of words. This is common, for example, between adults and child language learners (Clark, 2007):

(1) a. Naomi: *mittens*.
b. Father: gloves.
c. Naomi: *gloves*.
d. Father: when they have fingers in them they are called gloves and when they are all put together they are called mittens.

But such interactions also take place between adults engaged in a joint activity (Brennan and Clark, 1996):

(2) a. A: A docksider.
b. B: A what?
c. A: Um.
d. B: Is that a kind of dog?
e. A: No, it's a kind of um leather shoe, kinda preppy pennyloafer.
f. B: Okay, got it.

In both of these examples, the participants have a joint perceptual scene to help ground the meaning of the word, but that need not always be the case. Definition is also a common coordination strategy in *word meaning negotiations* that take place on text-based social media (Myrendal, 2019).

In this paper, we consider a particular definition paradigm known as a *genus-differentia* definitions. Consider the following (imagined) exchange between an expert ornithologist and aspiring birder:

(3) a. A: You know what a corvid is, right?
b. B: Yeah, sure. We have jays and crows in the garden sometimes.
c. A: A raven is a large black corvid.
d. B: Oh, okay.

Each of the above examples can be analysed as including a genus-differentia definition (Table 1). Furthermore, it seems reasonable to expect that each exchange results in some update to the *common ground* (Clark, 1996) of the participants.

Discussion of genus-differentia definitions can be traced back at least as far as Aristotle.[1] For Aristotle, each genus must be separated into species by some external *differentia*. Some species, acting as genera themselves, may be further differentiated into subspecies. We adopt some of the language of the Aristotelian tradition (genus, species, differentia), but rather than metaphysics, we are interested in genus-differentia definitions as a conventionalised resource for linguistic agents to coordinate on the meaning a word or phrase.

Genus-differentia definitions convey two kinds of information about the definiendum:

1. **taxonomical information** – A *X is a Y* relationship between the genus and the definiendum

2. **observational information** – One or more *features* that help to *differentiate* the definiendum from other species of the same genus

[1] See especially Books VI and VII of *Topics*.

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

Ex.	definiendum	genus	differentia
(1)	mittens	mittens ∨ gloves	fingers are all put together
(2)	docksider	shoe	leather
		pennyloafer	preppy
(3)	raven	corvid	large, black

Table 1: Three examples metalinguistic coordination analysed as genus-differentia definitions. While (3) fits neatly into the paradigm, the other two deviate somewhat. In (1), the genus is not explicitly stated, but can be taken to be a join type encompassing both *mittens* and *gloves* (see Cooper and Larsson, 2009). In (2), two alternative definitions are given, each with their own genus and differentia.

Marconi (1997) argues that there are two ways for speakers to be competent with the use of a word. *Referential competence* is the ability to map words to individuals or events in the world. If someone can identify a raven by sight (or by call, or by observing its behavior), they might be considered referentially competent with *raven*. This aspect of competence seems to be what is mainly at issue in the argument that at least some aspect of lexical semantic meaning may be associated with a perceptual classifier—a cognitive resource for identifying instances of a class, given some perceptual input (Larsson, 2013; Schlangen et al., 2016). On the other hand, *inferential competence* supports the ability to draw inferences based on the use of a word in context. In a community of bird watchers, one might be expected to infer from an utterance like *I saw a raven* that I saw a corvid. Someone who doesn't make that inference might be considered incompetent with the word *raven*, since part of the meaning of *raven* that they are corvids. Formal semantics in the Montagovian tradition, if it considers lexical semantics at all, focuses on inferential aspects of meaning, for example with meaning postulates (Carnap, 1952; Zimmermann, 1999).

Genus-differentia definitions are interesting to consider from the perspective of interaction because describing the result of grounding an utterance like (3-c) requires a framework that accounts for the dual nature of lexical meaning. We have essentially two desiderata for the shared meaning of *raven* that results from grounding (3-c):

D1 Raven is a species of the genus corvid.[2] This means two things: First, there is an intensional inferential relation from species to the genus. That is, there is no situation (actual or hypothetical) in which something might not be a corvid given that it is a raven, since the definition stipulates that being a corvid is part of *what it means* to be a raven. Second, being a raven is mutually exclusive with each of the sibling species of corvid.[3]

D2 Given that something is a corvid, being large and black (relative to corvids) is positive evidence for being a raven. However, this does not mean that ravens are a *type of* black thing. Any inference from *raven* to *large and black* is defeasible (for example, the speakers may entertain the possibility of albino raven, even if it happens to be extensionally true that all ravens are black). Furthermore, our account should accommodate the possibility that some differentia are interpreted in a way that is sensitive to the context given by the genus. For the sake of example, we will assume that this is the case for *large* but not for *black*.

Our analysis of (3) and therefore these desiderata is admittedly *ad hoc*. Indeed, the use of genus-differentia definitions as a metalinguistic resource is probably a source of variation across different communities of practice. The analysis that leads to these desiderata is partly motivated by the very fact that it requires us to distinguish between taxonomical and observational information about the meaning of *raven*.

We will come back to these desiderata in Section 4 after developing some formal machinery that we can use to express them more precisely. Section 1 introduces Probabilistic Type Theory with Records (ProbTTR). Section 2 describes a way of representing multiclass classifiers in ProbTTR, and Section 3 describes *classification systems*, a kind of ProbTTR type system that encodes a taxonomy

[2] Since we are interested in lexical meaning, the taxonomical information relevant to us is information about *folk taxonomies*, which are a resource for a particular *community of practice* (Gumperz, 1972). Among botanists, a banana is a species of berry while a strawberry is not. The opposite may hold among cooks or in ordinary discourse.

[3] Exactly what the sibling species are may be underspecified in the common ground. In this case, it includes at at least *jay* and *crow*, given the context of (3-b). In other cases, the relevant sibling species may be inferable from the differentia.

with types that refer to multiclass classifiers for their witness conditions. Finally, in Section 4, we will put these tools together to give an analysis of example (3).

1 Probabilistic Type Theory with Records

Probabilistic Type Theory with Records (ProbTTR) is a type system that allows for probabilistic type judgments of the form

$$p(a : T) = r, \tag{4}$$

where $r \in [0, 1]$ is a real number. In settings where the type system is a resource for (or models cognitive processes of) an agent, (4) is taken to mean that the agent judges entity a to be of type T with probability r.[4]

Possibilities and witness conditions In ProbTTR, *witness conditions* are used to compute the probability that a given entity is of a given type. For basic types, $T \in \mathbf{BType}$, witness conditions assign probability dependent on a *possibility* external to the type system. A possibility can be a set theoretic model (in which case the witness conditions for basic types is one of set membership) or it can, as in this paper, be based on a collection of classifiers (see Section 3.2). Thus, we write

$$p(a :_M T) = r \tag{5}$$

to mean that a is of type T with probability r in possibility M. Statements like (4) should only be used for judgments that hold regardless of possibility, or as a shorthand where it is clear that only one possibility is being considered.

We have not explicitly introduced a probability space underlying type judgments. In general, this may not be formally necessary (see Scott and Krauss, 1966). However, if we did, the sample space would be the set of all possible sets of pairs of basic types and entities:

$$\Omega = \mathcal{P}(\mathbf{BType} \times \mathit{Ind})$$

where, for $A \in \Omega$, $\langle T, a\rangle \in A$ would mean that a is of type T in outcome A.

As long as both $\mathbf{BType}$ and Ind are countable (for the purposes of this paper, we may assume they are finite), the distribution is discrete and there is no difficulty in talking directly about the probability of events.

[4] See Cooper et al. (2015) for a more complete introduction to ProbTTR.

A key point that is elucidated by considering the sample space of basic type judgments is that probabilistic dependencies between type judgments on basic types are entirely determined by M.

Conditional probability We may speak of the *conditional probability* that an entity a is of type T_1 given that it is of type T_2, written $p(a : T_1 \mid a : T_2)$. If we wish to express the probability (in general) that something is of type T_1 given that it is of type T_2, this is written $p(T_1 \| T_2)$. The use of the double stroke is to distinguish this expression from the probability that something *exists* of type T_1, given that something *exists* of type T_2, which is written $p(T_1 \mid T_2)$. These conditional probabilities are understood extensionally, specific to a particular *possibility*. If, for example, we know that penguins only live in Antarctica, we would, for the types $\mathit{Penguin}$ (the type of situation in which there is a penguin) and $\mathit{Antarctica}$ (the type of situation in Antarctica), judge $p(\mathit{Antartica} \| \mathit{Penguin})$ to be 1 (or close to 1) on the basis of this contingent fact.

Structured types The witness conditions of structured types are a function of the structure of the type and its components. For example, given types T_1 and T_2, the meet type $T_1 \wedge T_2$ has, witness conditions based on the Kolmogorov (1950) equation for conjunctive probability (Cooper et al., 2015):

$$\begin{aligned} p(a : T_1 \wedge T_2) &= p(a : T_1) \cdot p(a : T_2 \mid a : T_1) \\ &= p(a : T_2) \cdot p(a : T_1 \mid a : T_2) \\ &= p(a : T_2 \wedge T_1) \end{aligned} \tag{6}$$

In addition to types defined with $\wedge$, $\vee$ and $\neg$, ProbTTR defines *record types* as structured types—given a record s and record type R, $p(s : R)$ is a function of type judgments of the fields of s (see Cooper et al. (2015) for details).

1.1 Hard and soft relations between types

Subtype relation In TTR, T_1 is said to be a *subtype* of T_2, $T_1 \sqsubseteq T_2$ if and only if anything of type T_1 is also of type T_2 for any possibility M, (Cooper, forthc, p. 285). Extending this to ProbTTR, we can say,

$$T_1 \sqsubseteq T_2 \text{ iff } p(a :_M T_1) \leq p(a :_M T_2), \tag{7}$$

for any entity a and possibility M.

Naturally, it is not always necessary to check these conditions explicitly.[5] Subtype relations can be implicit in the structure of the types, as in the case of meet types. If $T_3 = T_1 \wedge T_2$, by the definition of the meet type we have $T_3 \sqsubseteq T_1$ and $T_3 \sqsubseteq T_2$.

In other cases, whether two types stand in a subtype relation may depend on what is meant by *all possibilities*. If we literally mean all possible assignments of probability to basic type-entity pairs, then two basic types will never stand in a subtype relation, since there will always be possibilities where $p(a :_M T_1) > p(a :_M T_2)$ and *vice versa*.

If, on the other hand, we restrict our attention to some class of possibilities $\mathcal{M}$, then subtype relations between basic types are possible. Witness conditions are one way to limit the possibilities under consideration and can therefore introduce probibalistic dependency between types.

Evidential relation We introduce a "soft" relation between types in ProbTTR, which captures the notion that T_2 is *evidence for* T_1 in the context of some type T^*. Two types stand in this relation with respect to T^* if learning that something is of type T_2 increases the probability that it is of type T_1:

$$T_1 \prec_{T^*} T_2 \text{ iff } p(T_1 \| T^*) < p(T_1 \| T_2, T^*) \quad (8)$$

This relation is also contingent, relative to a particular possibility.

1.2 Representing probability distributions

In the next section, we will define a type for probabilistic multiclass classifiers—that is, classifiers that compute the probability that a given entity belongs to each of several mutually exclusive classes. To that end, we must first encode discrete categorical probability distributions in TTR, since the output of the classifier takes that form.

Larsson and Cooper (2021) introduce a type theoretic counterpart of a random variable in Bayesian inference. To represent a single (categorical) random variable with a range of possible (mutually exclusive) values, ProbTTR uses a *variable type* $\mathbb{A}$ whose range is a set of *value types* $\mathfrak{R}(\mathbb{A}) = \{A_1, \ldots, A_n\}$. We might have, for example, $\mathfrak{R}(Animal) = \{Bird, Reptile, \ldots\}$.

[5] Indeed, it may not even be possible, depending on the notion of possibility since the "extension" of types with witness conditions based on classifiers is indeterminate (Larsson, 2020b).

We will use short-hands $Animal$, $Bird$ etc, for the situation where some individual is an animal, bird, etc.:

$$Animal = \begin{bmatrix} \text{x} & : & Ind \\ \text{c} & : & \text{animal(x)} \end{bmatrix}$$

$$Bird = \begin{bmatrix} \text{x} & : & Ind \\ \text{c} & : & \text{bird(x)} \end{bmatrix}$$

For a situation s, a probability distribution over the m value types $A_j \in \mathfrak{R}(A), 1 \leq j \leq m$ belonging to a variable type $\mathbb{A}$ can be written (as above) as a set of Austinian propositions, e.g.,

$$\{ \begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & A_j \\ \text{prob} & = & p(s : A_j) \end{bmatrix} \mid A_j \in \mathfrak{R}(\mathbb{A}) \} \quad (10)$$

However, we will also have use for an alternative representation of probability distributions, that indexes the probability assigned to each type with a unique label associated with the type:

$$\text{idx}(\{ \begin{bmatrix} \text{sit} & = s \\ \text{sit-type} & = A_j \\ \text{prob} & = p(s : A_j) \end{bmatrix} \mid A_j \in \mathfrak{R}(\mathbb{A}) \}) = \begin{bmatrix} \text{lbl(A}_1\text{)} & = & p_1 \\ \vdots & = & \vdots \\ \text{lbl(A}_\text{n}\text{)} & = & p_n \end{bmatrix}$$

where $p_j = p(s : A_j)$ and $\text{lbl}(A_j)$ is a unique label for $A_j \in \mathfrak{R}(\mathbb{A})$. This means that for a set of probabilistic Austinian propositions P_s, that concern a situation s, $\text{idx}(P_s).\text{lbl}(A_j) = p_j = p(s : A_j)$.

2 Multiclass Classifiers in ProbTTR

In this section we extend the TTR classifier defined by Larsson (2013) to give probabilistic type judgments in multiclass setting.

Larsson (2013) shows how perceptual classification can be modelled in TTR and Larsson (2020a) reformulates and extends this formalisation to probabilistic classification. Adapting the notation of a probabilistic TTR classifier to the current setting, a probabilistic perceptual (here, visual) classifier $\kappa_{\mathbb{A}}$ corresponding to a variable type $\mathbb{A}$ provides a mapping from perceptual input (of type $\mathfrak{V}$ e.g., a digital image) onto a probability distribution over value types in $\mathfrak{R}(\mathbb{A})$, encoded as a set of probabilistic Austinian propositions.

We also want to explicitly parametrise our classifier. A classifier $\kappa_{\mathbb{A}}$, would thus be a function of type:

$$\Pi \rightarrow Sit_{\mathfrak{Y}} \rightarrow \{ \begin{bmatrix} \text{sit} & : & Sit_{\mathfrak{Y}} \\ \text{sit-type} & : & RecType_{A_i} \\ \text{prob} & : & [0,1] \end{bmatrix} \mid A_i \in \mathfrak{R}(\mathbb{A}) \} \quad (11)$$

where Π is the type of the parameters needed by $\kappa_{\mathbb{A}}$, and $Sit_{\mathfrak{Y}}$ is the type of situations where perception of some object yields visual information, and where $RecType_R$ is the (singleton) type of records identical to R, so that e.g.,

$$T : RecType_{Bird} \text{ iff } T : RecType \text{ and } T = Bird$$

We take classifiers to be part of word meanings. We associate a word like "bird" with a type $Bird$ which is in turn associated with lexical entry in the form of a TTR record:

$$Lex(Bird) = \begin{bmatrix} \text{bg} & = & Sit_{\mathfrak{Y}} \\ \text{par} & = & \pi \\ \text{intrp} & = & \lambda r : \text{bg} \,.Bird \\ \text{clfr} & = & \lambda r : \text{bg} \,.\kappa_{Animal}(\text{par}, r) \end{bmatrix} \quad (12)$$

Assuming we have a function Lex that looks up the lexical entry related to a type (associated with a word), we also define a lookup function that gives us the classifier corresponding to a type:

$$Clfr(T) = Lex(T).\,\text{clfr}$$
$$Intrp(T) = Lex(T).\,\text{intrp}$$

Let us assume a s_{123} situation where a speaker points to a bird a and says “Bird!” (meaning “that is a bird”). We want to classify a perceived situation as being of the type $Bird$ or not, or in the probabilistic case, compute the probability of the judgment.

Now, to judge the probability with which a situation s is of a type $Bird$ (to continue with our example), the agent looks up the related classifier and applies it to s, which produces a probability distribution over different subtypes of $Animal$. The agent then looks up the probability associated with $Bird$. The general method for doing this can be written as:

$$p(s : T) = \text{idx}(Clfr(T)(s)).\,\text{lbl}(Intrp(T)(s))$$

In our case:

$$p(s_{123} : Bird) = \text{idx}(\kappa_{Animal}(\pi, s_{123})).\,\text{lbl}(Bird)$$

3 Classification systems in ProbTTR

To represent both taxonomical and observational relations between types, we will embed a *classification system* in ProbTTR. A classification system has two components, a *taxonomy* (Section 3.1), which is a set theoretic object representing an ontological hierarchy, and a collection of *classifiers* (Section 3.2) associated with the taxonomy. Ultimately the classifiers will provide witness conditions for certain basic types and the taxonomy will be fully encoded in the type system, but first we define the structure in set theoretic terms so that we can create a ProbTTR system with the correct subtype relations.

3.1 Taxonomy

A taxonomy is a rooted tree structure defined by a tuple,

$$\mathbf{T} = \langle T, D, t^* \rangle, \quad (13)$$

where T is a set of *taxons*, $D \subseteq T \times \mathcal{P}(T)$ is a set of *distinctions* on T, and $t^* \in T$ is the root taxon.

To elaborate, T is simply a finite set of labels and D provides the hierarchical structure of the taxonomy. *Distinctions* (elements of D) take the form $\langle g, S \rangle$, where $g \in T$ and $S \subset T$, and $|S| \geq 2$. We say that the taxons g and s stand in a genus-species relationship if there is some $\langle g, S \rangle \in D$ such that $s \in S$. Then s can be said to be a *species of* g. Alternatively, we can say that g is *the genus of* s.

This requires certain restrictions on $\mathbf{T}$. Namely, that it is:

- **Acyclic**: There are no cycles. I.e., no chain of distinctions $\{\langle g_1, S_1 \rangle, ..., \langle g_n, S_n \rangle\}$ such that $g_2 \in S_1, ..., g_n \in S_{n-1}$ and $g_1 = g_n$.
- **Rooted**: There is no distinction $\langle g, S \rangle \in D$ with $t^* \in S$.

- **Uniquely connected**: For every $t \neq t^*$ there is exactly one $\langle g, S\rangle \in D$ such that $t \in S$.[6]

Importantly, this still allows for multiple distinctions in which the same taxon acts as a genus. In other words, we can have $\langle g, S\rangle, \langle g, S'\rangle \in D$ where $S' \neq S$. For example, we might imagine a taxonomy in which both $\langle Animal, \{Bird, Reptile, ...\}\rangle$ and $\langle Animal, \{Carnivore, Herbivor, Omnivore\}\rangle$ are distinctions.

The *uniquely connected* constraint allows us to define a function

$$Dist : T \setminus \{t^*\} \to D \tag{14}$$

that gives, for each taxon, t (other than t^*), the distinction $Dist(t) = \langle g, S\rangle$ such that $t \in S$. For convenience we also define the functions $Genus$, and $Siblings$ such that

$$\langle Genus(t), Siblings(t)\rangle = Dist(t). \tag{15}$$

Note that under this definition, leaf taxons are those taxons for which there are no distinctions in D where the taxon appears as a genus.

3.2 Species Classifiers

In addition to the taxonomy, we have a collection of classifiers, $\mathbf{K}$ and parameters $\mathbf{P}$, each of which we index with elements of D, such that $\kappa_d \in \mathbf{K}$ is the classifier for distinction d provided with the appropriate parameters. This follows the intuition that a distinction in the taxonomy may be accompanied by an ability to *distinguish* among the relevant species. In general, we need only assume that we have classifiers for those distinctions that include at least one leaf taxon, since genus taxons can be defined as the join of their species in certain cases.[7] For now we will assume we have a classifier for each distinction in D.

3.3 The type system

Suppose we have a taxonomy $\mathbf{T} = \langle T, D, t^*\rangle$ and a collection of classifiers $\mathbf{K}$ on the distinctions of that taxonomy. Let Dom be a special type corresponding to the root of the taxonomy. We then define variable types $\mathbb{A}_d$ for each $d = \langle g, S\rangle \in D$ with $\mathfrak{R}(\mathbb{A}) = \{A_{s_1}, ..., A_{s_n}\}$ corresponding to $s_1, ..., s_n \in S$. Classifiers provide the witness conditions for the value types as described in Section 2. For a given entity a,

$$p(s : A_t) = \begin{cases} 1 & \text{if } t = t^* \\ \kappa_{Dist(t)}(a)(t) & \text{otherwise} \end{cases} \tag{16}$$

In other words, the probability assigned to A_t is 1 in the case of the root taxon, and otherwise determined by the classifier for the distinction corresponding to the variable in which A_t is a value type. These "auxiliary" value types we can give the witness conditions for the associated with the taxonomical categories as the product of the jugment of the genus and the axiliary type. For any object a,

$$p(a : T_t) = p(a : A_t) \cdot p(a : T'_t) \tag{17}$$

where

$$T'_t = \begin{cases} Dom & \text{if } t = t^* \\ T_{Genus(t)} & \text{otherwise} \end{cases}$$

This stipulates that the classifiers give us the probability that an individual is of each of the species types, *given* that it is of the genus type. Thus judgments about T_t correspond to an *absolute* judgment about belonging to the taxon.

Taken together, Equations 16 and 17 imply that for any a, $p(a : T_{t^*}) = p(a : Dom)$. In situations where the root taxon corresponds to all individuals (i.e., where $Dom = Ind$), we have $p(a : T_{t^*}) = 1$ for any a. It is also possible, however, to embed a classification system in an existing type system, as long it provides witness conditions for Dom. For example, if the classification system is specific to birds, we might embed it in a larger system that gives witness conditions for $Bird$.

3.4 Feature classifiers

In addition to the distinction classifiers, a classification system may include some number of types based on feature classifiers. A feature classifier takes any entity $a : Dom$ as input, and receives its witness conditions from a classifier that results in a probabilistic type judgement. In general, feature and distinction classifiers need not interact explicitly though, considered as random variables, there may be probabilistic dependence between them. Distinction classifiers may be defined in terms of

[6] A weakness of insisting on a tree structure is that we cannot have taxons that appear in multiple places in the taxonomy, whereas in folk taxonomies it would appear this is common. We would either need to say that the apparently duplicated taxon is actually part of a distinction at a higher level that encompasses both, or that it corresponds to two senses of the same word.

[7] See Marconi (1997, ch. 6) on "subordinate concepts".

feature classifiers, for example as Bayesian classifiers that take the result of feature classifiers as their input (see, Larsson and Bernardy (2021)).

In general, some of these feature types may be dependent types. Consider a type like $Tall$. Whether or not an individual is tall may depend on a comparison class (for example, a type in the taxonomy). Following Fernandez and Larsson (2014), we define dependent feature types with classifiers that take a threshold function as a parameter. For example,

$$\theta_{Large} : Type \to \mathbb{R}^+ \tag{18}$$

This gives the classifier the following type:

$$\kappa_{Large} : (Type \to \mathbb{R}^+) \to Type \tag{19}$$

4 Combining the observation and taxonomical aspects of genus-differentia definitions

With this formal machinery in place, we return to the project of characterising the result of grounding (3-c). First, let's lay out what is shared among speakers A and B before (3-c) is grounded.

We will assume that A and B share a classification system with $Bird$ at its root as part of their common ground. Utterance (3-d) establishes that a type for the lexical entry of *corvid*, for which we will use Cor, is a type in this system, and that there is a distinction on Cor such that $\mathfrak{R}(Cor) \supseteq \{Jay, Crw\}$, where Jay and Crw are the lexical entries for *jay* and *crow*—that is, for all species types of Cor given by the common ground , S (including at least Jay and Crw), $S \sqsubseteq Cor$. The witness conditions for each $S \in \mathfrak{R}(Cor)$ are given by a multiclass classifier κ_{Cor}. Since $Cor \sqsubseteq Bird$, we may also assume that $Dist(Cor)$ exists and that there is a classifier $\kappa_{Dist(Cor)}$, though it need not be common ground what the genus of Cor is.

Furthermore, we will assume we have types Lrg and Blk, whose witness conditions are given by feature classifiers. For the purposes of the example, we will assume that Blk is basic type that gets its witness conditions from a feature classifier, κ_{Blk}, whereas $Lrg : Type \to Type$ is a dependent type with a classifier that depends on threshold function θ_{Lrg}. Thus, the witness conditions for $Lrg(Cor)$ are given by $\kappa_{Lrg}(\theta_{Lrg}(Cor))$. This leaves open the question of exactly how θ_{Lrg} is defined, but we may assume that the value of $\theta_{Lrg}(Cor)$ depends in some way on the parameters of the classifier that defines the witness conditions for Cor, namely $\kappa_{Dist(Cor)}$.

Returning to our desiderata, we want to construct a type, Rav, such that:

$$\sum_{T \in Species(Cor) \cup \{Rav\}} p(T \| Cor) = 1 \tag{20a}$$

$$Rav \sqsubseteq Cor \tag{20b}$$

$$Rav \prec_{Cor} Lrg(Cor) \wedge Blk \tag{20c}$$

Here (20a) and (20b) formalise **D1** and 20c formalises **D2**.

4.1 Constructive approach

As discussed previously, one motivation for formalising this example and the interactive semantics of genus-differentia definitions in general is to expose some crucial distinctions in lexical semantics that are often overlooked. In this section, we give what is a rather straight-forward and intuitive solution to the challenge we have given ourselves, but one that fails to adequately make the distinction between taxonomical and observational lexical information.

In this solution, we attempt to directly construct a new type Rav out of the common ground types already available. The most straight-forward way to do this is with meet types:

$$Rav = Cor \wedge (Lrg(Cor) \wedge Blk) \tag{21}$$

This definition is intuitively appealing—(3-c) is saying that ravens are large *and* black *and* corvids. Furthermore, this definition does actually satisfy the desiderata stated so far.

To maintain (20a), we can redefine each existing species type S as:

$$S' = S \wedge \neg Rav \tag{22}$$

We have $Rav \sqsubseteq Cor$, satisfying (20b), since by the Kolmogorov (1950) definition of the meet type (6), for any possibility M and any entity a,

$$\begin{aligned} & p(a :_M Rav) \\ =& p(a :_M Cor) \cdot p(a :_M Lrg(Cor) \wedge Blk \mid Cor) \\ \leq& p(a :_M Cor) \end{aligned}$$

Finally, (20c) holds since it follows from the definition of Rav that, $p(Rav \| Lrg(Cor) \wedge Blk, Cor) = 1$ and, assuming there are

at least some non-large, non-black corvids, $p(Rav \| Cor) < 1$.[8]

However, the definition of the meet type (6) implies we also get $Rav \sqsubseteq Lrg(Cor)$ and $Rav \sqsubseteq Blk$. It does not make sense for Rav to be a *subtype* of large corvids or of black things (consider again the possibility of an albino raven). Put another way, it should be possible to construct a hypothetical possibility M and entity a such that:

$$p(a :_M Lrg(Cor) \wedge Blk) = 0 \text{ and } p(a :_M Rav) > 0 \quad (23)$$

In the next section, We will consider this a new desiderata along with the constraints in (20). Instead of constructing the type directly from existing types, we posit a basic type without explicit witness conditions, but with some constraints that are derived from by the genus-differentia definition.

4.2 Underspecifed approach

Cooper (forthc) treats types as having an existence independent of their witness conditions. Two types can share the same witness conditions, for example, and still play different roles in an agent's type system. Part of the motivation for doing this is that an agent can reason about a type and its relation to other types without specifying witness conditions for that type. This is in contrast to predicates in first-order logic, for example, which don't have any meaning independent of the model theoretic entities they are interpreted as.

We would like to interpret definitions like (3-c) as giving rise to an underspecified type; that is, a type without explicit witness conditions. Instead, we assert the following relationships between the new underspecified type Rav and other existing common ground types:

$$Rav \sqsubseteq Cor \quad (24a)$$

$$p(Lrg(Cor) \wedge Blk \| Rav) = 1 \quad (24b)$$

Notice that neither of these two conditions give us direct witness conditions for Rav. The first condition says that anything (in any possibility) that is a raven is also a corvid. The second condition says that anything that is a raven is, with probability 1, is large (for a corvid) and black. Note that (24b) is a constraint on the type's witness conditions given the current possibility, meaning that we can not infer $Rav \sqsubseteq Lrg(Cor) \wedge Blk$, since nothing prevents us from constructing a possibility in which (23) holds. In other words, albino ravens are still possible.

Clearly condition (20b) is satisfied by construction. This may be a bit unsatisfying, but it is worthwhile to consider that asserting $Rav \sqsubseteq Cor$ amounts to adding Rav as a witness condition to Cor. Put another way, for any entity a and possibility M, $P(a :_M Cor) \geq P(a :_M Rav)$.

In order to satisfy (20a), we need to redefine the witness conditions of the existing species types to "make room" in the probability distribution for Rav. How to do this depends somewhat on how completely the distinction is specified in the common ground. If there is an *other corvid* type, $Other$, we might just redefine the classifier for that type so that for any entity a, $\kappa'_{corvid}(a)(other) = \kappa_{corvid}(a)(other) - f(a)$, where f is such that $0 < f(a) < \kappa_{corvid}(a)(other)$. Alternatively, we might take some probability from each class. Either way, the solution should be a function of a that depends on the differentia, but exactly what that function is is not common ground since (24b) gives a unidirectional conditional—all ravens are large and black, but there may still be large, black, non-raven corvids.

It remains to be shown that (20c) holds. In the following, let $D = Lrg(Cor) \wedge Blk$ and S be the set of types representing each of the sibling species of Cor, including Rav.

$$\begin{aligned} & p(Rav \| D, Cor) \\ =& \frac{p(Rav \| Cor) \cdot p(D \| Rav, Cor)}{\sum_{T \in S} p(T \| Cor) \cdot p(D \| T, Cor)} && (25) \\ =& \frac{p(Rav \| Cor) \cdot p(D \| Rav)}{\sum_{T \in S} p(T \| Cor) \cdot p(D \| T)} && (26) \\ >& p(Rav \| Cor) \cdot p(D \| Rav) && (27) \\ =& p(Rav \| Cor) && (28) \end{aligned}$$

In the above, (25) follows from Bayes rule and the fact that $\sum_{T \in S} p(T \| Cor) = 1$, and (26) follows from $Rav \sqsubseteq Cor$. For (27), we must assume that

$$\sum_{T \in S} p(T \| Cor) \cdot p(D \| T) \leq 1.$$

This is the same assumption we made in the previous approach, which we argue follows from

[8] This assumption is justified by a pragmatic requirement of genus-differentia definitions that the differentia do at least some work to *differentiate* the definiendum from other species of the genus.

the pragmatics of genus-differentia definitions—namely that not all non-raven corvids are large and black. Finally, (28) follows directly from (24b).

In this approach, the type for *raven*, Rav is defined only in terms of its relationship to types corresponding to other terms in the utterance. A notable feature of this solution is that everything we learn from the definition can be stated in terms of witness conditions for types that already exist: In the case of *corvid*, we know that anything that witnesses the type Rav is a witness for the type Cor. This holds intensionally, meaning that it is true independent of possibility. In the case of *large* and *black*, we know *extensionally* that anything that is a raven will be large and black.

Speaker B learns the type Rav and the constraints associated with it (24) based on the definition offered by A in (3-c). After (3-d), this type and the associated constrains are added to the common ground.

5 Conclusion

The main goal of this paper was to develop a framework that can deal with the distinction between taxonomical and observational lexical information. We argue that this distinction is one that speakers make in metalinguistic interaction, as in genus-differentia definitions. In order to account for this distinction, we use a type system in which intensional relations between types can be reasoned about independently of their witness conditions, which depend on facts about the world.

Our account has been agnostic to the implementation of the classifiers involved. This is justified, in part, by the fact that we describe updates to the conversational common ground, rather than individual agents' abilities. However, it may also be interesting to consider what effect a dialogue like (3) may have on speaker B's ability to recognise ravens. This is related to the machine learning task of *zero-shot classification*, in which an existing classifier is adapted to recognise instances of previously unknown classes based on external information (such as a natural language descriptions). Future work should consider how zero-shot classification can be analysed from an interactive perspective.

References

Susan E Brennan and Herbert H Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

Rudolf Carnap. 1952. Meaning postulates. *Philosophical Studies*, 3(5):65–73.

Eve V. Clark. 2007. Young Children's Uptake of New Words in Conversation. *Language in Society*, 36(2):157–182.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

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

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic Type Theory and Natural Language Semantics. In *Linguistic Issues in Language Technology, Volume 10, 2015*. CSLI Publications.

Robin Cooper and Staffan Larsson. 2009. Compositional and ontological semantics in learning from corrective feedback and explicit definition. In *Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Raquel Fernandez and Staffan Larsson. 2014. Vagueness and Learning: A Type-Theoretic Approach. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 151–159, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

J Gumperz. 1972. The Speech Community. In Pier Paolo Giglioli, editor, *Language and Social Context: Selected Readings*. Harmondsworth : Penguin.

A. N. Kolmogorov. 1950. *Foundations of the Theory of Probability*. New York: Chelsea Pub. Co.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Staffan Larsson. 2020a. Discrete and Probabilistic Classifier-based Semantics. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 62–68, Gothenburg. Association for Computational Linguistics.

Staffan Larsson. 2020b. Extensions are Indeterminate if Intensions are Classifiers. In *SemDial 2020 (WatchDial) Workshop on the Semantics and Pragmatics of Dialogue*, page 10, Waltham, MA and online.

Staffan Larsson and Jean-Philippe Bernardy. 2021. Semantic Classification and Learning Using a Linear Tranformation Model in a Probabilistic Type Theory with Records. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 14–22, Gothenburg, Sweden. Association for Computational Linguistics.

Staffan Larsson and Robin Cooper. 2021. Bayesian Classification and Inference in a Probabilistic Type Theory with Records. In *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, pages 51–59, Groningen, the Netherlands (online). Association for Computational Linguistics.

Diego Marconi. 1997. *Lexical Competence*. Language, Speech, and Communication. MIT Press, Cambridge, Mass.

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

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1213–1223, Berlin, Germany. Association for Computational Linguistics.

Dana Scott and Peter Krauss. 1966. Assigning Probabilities to Logical Formulas. In Jaakko Hintikka and Patrick Suppes, editors, *Studies in Logic and the Foundations of Mathematics*, volume 43 of *Aspects of Inductive Logic*, pages 219–264. Elsevier.

Thomas Ede Zimmermann. 1999. Meaning Postulates and the Model-Theoretic Approach to Natural Language Semantics. *Linguistics and Philosophy*, 22(5):529–561.

Focus negation in formal grammar

Kata Balogh
Heinrich-Heine-Universität Düsseldorf
balogh@hhu.de

In this paper, we introduce a proposal towards a formal grammatical model that captures different types of negation uniformly, in terms of communicative functions and pragmatic structuring. The central objective of the work presented here is the analysis and formal modeling of the relation between focusing and negation, where next to the logico-semantic understanding of negation, the information structural interpretation plays a crucial role. The grammatical model proposed here is modular, with separate but interrelated representations for syntax, semantics and information structure, where the latter two together determine the (discourse) context-based interpretation of the sentence. We argue for an analysis of negation that targets the newly conveyed information (i.e., its communicative function) determined by the focus structure of the sentence, hereby accounting for the focus sensitivity of negation. The semantic representation of the sentence is given as a decompositional frame, which reflects a mental representation / description of the event expressed in the given sentence.

1 Types of negation

Dating back to the earliest discussions on negation (Aristotle; the Stoic School; Jespersen, 1917; Klima, 1964), there are different types distinguished, which lead to related, but still different notions. For example, Aristotle distinguished 'predicate denial' and 'term negation', philosophers of the Stoic School talk about 'external negation' and 'internal negation', Jespersen (1917) distinguishes 'nexal negation' and 'special negation', and Klima (1964) distinguishes 'sentence negation' and 'constituent negation'. Regardless of the differences between these notions (see, e. g., De Clercq, 2020), a crucial aspect of distinguishing these 'negation types' is the domain that the negation operates on. Despite the recognition of the different types of negation, formal syntactic/semantic accounts capture negation in the locigo-semantic terms and mostly investigate sentential (or propositional) negation, and related phenomena that are crucial at the syntax-semantics interface (e.g., the relation of sentential negation and quantification, the interpretation of negative indefinites, the analysis of negative polarity items and negative concord). The work presented here is inspired by the other type, which is generally underrepresented in current formal grammars and semantic/pragmatic approaches. This negation type is often referred to as 'focus negation', reflecting its tight relation with narrow focus structure. While in the locigo-semantic understanding of negation the two types can be captured uniformly in terms of a propositional operator, the two differ in their information structural interpretation. We argue for an analysis and introduce a proposal of a formal grammatical model, where the interpretational differences of the two negation types are captured within the information structure of the sentence, where negation scopes over the given focus domain. Hence, in information structural terms the different focus types reflect broad versus narrow scope negation. This basic assumption is in line with the analysis of (Vallduví, 1990).

2 Focus sensitivity of negation

Information structure, and hence focusing, manifests itself in different layers of natural language: in interpretation and in structure building. In interpretation, focusing can be treated semantically: as introducing alternatives (Rooth, 1992) or structuring semantic content (Krifka, 2001), and pragmatically: relating to the QUD (Roberts, 2012) or in terms of pragmatic structuring (Lambrecht, 1994). Structure building effects of focusing manifest itself in various languages (e.g., Hungarian, Basque) in terms of triggering dedicated syntactic operations and configurations (e.g., É. Kiss, 1995).

It is widely accepted that the interpretation of a range of linguistic expressions is dependent on

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

the information structure of the utterances in which they occur (König, 1991; Krifka, 2001; Beaver and Clark, 2008). This holds for focus sensitive particles (e.g., *only, also*), as well as for negation. This observation holds across languages and the phenomenon is referred to as *focus sensitivity*. See, e. g., (1), where the interpretation of the exclusive operator (*only*) depends on the placement of focus, hence the focus structure of the sentence.

(1) a. Pim only saw [MIA]^{F} at the party.
⇝ Pim saw Mia, and noone else, at the party

b. Pim only saw Mia at the [PARTY]^{F}.
⇝ Pim saw Mia at the party, and nowhere else

Current approaches to focus sensitivity are rather restricted to the field of formal semantics/pragmatics, however, despite their fairly uniform semantics, focus sensitive elements vary across languages with respect to their structural behaviour, which in turn strongly affects their modeling in formal grammar. Leading grammar theories and formalisms that capture information structural phenomena (CCG, LFG, HPSG)[1] do not systematically address focus sensitivity. These accounts generally acknowledge both aspects of information structure (i.e., interpretation and structure building), but they often concentrate on only one of them, or lack the formal means in their architecture to equally address both aspects.

In information structural terms, the two major types of negation differ in their focus domain they operate on. *Focus negation* takes a narrow scope, while *sentential/propositional* negation takes a broad scope. As we will discuss later in more detail, these domains correspond to narrow and broad focus respectively. Under narrow scope negation, also affixal negation (e.g., *unhappy*), inherent negation (e.g., *deny*) and negative quantification (e.g., *no girls*) are often understood. Although they share the property of having a narrow scope, we argue that these represent different types. Under the type of 'focus negation', we understand the type, where the negative particle, which also expresses sentential negation, operates on a single constituent instead of the whole proposition. In the examples below, square brackets indicate the domain the negation operates on, and capitals indicate where the main stress falls.

[1] See, for example, Steedman (2000, 2019), Dalrymple and Nikolaeva (2011), Engdahl and Vallduví (1996).

(2) [Pim did not introduce Sam to MIA].

(3) a. Pim did not introduce [SAM] to Mia.
b. Pim did not introduce Sam to [MIA].

In (2), the negation takes a broad scope, and operates on the whole proposition. The negation in (3-a) and (3-b), however, takes a narrow scope: it only operates on the constituent that is marked as the narrow focus of the sentence. Similarly to the examples in (1), the interpretational difference between (3-a) and (3-b) is due to the different focus structures, hence sensitive to focusing. The focus sensitivity of negation is explicitly addressed by Beaver and Clark (2008), who claim that negation is 'quasi focus sensitive', which is best analyzed as a pragmatic implicature. We argue that the relation between negation and focusing is more tight, and should be part of the grammatical system. This is supported by the fact that in certain languages, the two negation types are structurally different, with a direct relation to the default focus marking. For example, in Hungarian, the negative particle *nem* 'not' can appear right before the predicate (4-a) or right before the preverbal narrow focus (4-b), directly reflecting the above negation types.

(4) a. Alex nem csókolta meg Samu-t.
Alex not kissed VPRT Sam-ACC
'Alex did not kiss Sam.'

b. Alex nem Samu-t csókolta meg.
Alex not Sam-ACC kissed VPRT
'It is not SAM whom Alex kissed.'

In a compositional analysis, the scope of the operator is the semantic content of the expression that stands in a given structural relation with it. For sentential negation this leads to the insertion of the logical operator above the predicate, which provides the intended interpretation. The reading of 'sentential negation' in (2) is straightforwardly captured by the formula $\neg introduce'(pim', sam', mia')$, where the logico-semantic operator of negation is applied to the whole proposition. In focus negation, however, a structural relation where only the given constituent is in the scope of the negative particle is not sufficient to give the right interpretation. In (3-a), we cannot simply apply negation to the content of the focal object. That is not meaningful, it does not even provide a well-formed formula. In (3-a), semantically (or truth-conditionally) it also holds that 'Pim did not introduce Sam to Mia', but it has an additional contribution: the *identification* expressed by focusing is targeted as well. The

sentence in (3-a) expresses that 'the one Pim introduced to Mia is not Sam'. To capture the correct contribution of focus negation, we need a formal grammar that accesses the focus structure and its communicative function: e.g., identification in case of narrow argument focus. We introduce our proposal towards such a model, beginning with its application to focus negation and then extending it to sentential negation in a uniform way.

3 Proposal

The formal analysis of any linguistic phenomenon requires a two-sided approach: theoretical claims need to be verified by empirically valid and formally exact models, and formal models must be built on solid theoretical grounds. Therefore, in our proposal, we build upon the formalized version of Role and Reference Grammar (Kallmeyer et al., 2013; Osswald and Kallmeyer, 2018), which facilitates such an approach. This formal grammar is based on a solid theoretical framework, Role and Reference Grammar (RRG; Van Valin and LaPolla, 1997; Van Valin, 2005), with a strong typological and cross-linguistic perspective. The formal specification of this grammar is defined in terms of Tree-Wrapping Grammar (Kallmeyer et al., 2013; Osswald and Kallmeyer, 2018), strongly inspired by Tree-Adjoining Grammar (Joshi and Schabes, 1997). The current developments of this grammar lack a formal specification and modeling of information structure, which asks for an extension.

3.1 Theoretical base

We argue for the cross-linguistic validity of the claim that negation generally has a direct access to the focus structure of the utterance (Van Valin, 2005), and next to its logico-semantic contribution, it operates on the contribution by focusing, i.e., on the information conveyed. In this paper, we discuss this latter, information structural aspect of negation. To capture our proposal, we first need to specify what exactly the contribution of focus is to the interpretation of the sentence. We argue for a context-sensitive perspective on the matter, and follow the theory of information structure by Lambrecht (1994), who claims that beyond the semantic content of the sentence, focusing leads to its *pragmatic structuring*. This structuring reflects the communicative functions: what information is conveyed and how this information is transferred between the discourse participants. The core aspect is the transfer of information and its relation to the Common Ground, the set of propositions shared by the interlocutors.

The 'pragmatic presupposition' of the sentence is the information content that is part of the discourse context shared by the discourse participants, and the 'pragmatic assertion' is the newly provided information, in relation to the pragmatic presupposition. Both concepts are lexico-grammatically defined, hence they are determined by the grammatical organization of the sentence. In the following, we systematically use the notions 'presupposition' and 'assertion' in the above sense, thus regarding 'pragmatic presupposition' (e.g., Stalnaker, 1974; Lambrecht, 1994) as opposed to 'conventional/semantic presupposition'.

Lambrecht (1994) defines focus structure as "*the conventional association of a focus meaning with a sentence form*" (Lambrecht, 1994, p. 22). He distinguishes three different focus structures based on the domain (i.e., scope) of the focus in the given sentence, and presents the systematic ways natural languages encode these structures in their morphosyntax. The core distinction is given on basis of whether a single constituent or multiple constituents are included in the focus domain. In this respect, we distinguish narrow focus and broad focus respectively. Broad focus is further divided into 'predicate focus', where the focus domain includes all parts except the topic and 'sentence focus', where the focus domain is the entire utterance. The predicate focus construction correlates with the topic-comment distinction, and is referred to as the unmarked focus type.

(5) a. [Pim saw MIA]^{F} (sentence focus)
b. Pim [saw MIA]^{F} (predicate focus)
c. [PIM]^{F} saw Mia. (narrow focus)
Pim saw [MIA]^{F}

The communicative functions of these focus structures are different: introducing an event or a referent (sentence focus), providing information of a topic (predicate focus) and idenitification of an entity with respect to an open proposition (narrow focus). All these functions correspond to the relation between presupposition and assertion, which is determined as the newly conveyed information. In the sentence *Pim saw* $[MIA]^{F}$ in (5-c), the focus is the semantic content of the object noun phrase, while the new information (i.e., the pragmatic assertion) is not this content itself, but the identification

relation between the entity represented by the focal noun phrase and the open proposition 'pim saw x' given as the pragmatic presupposition (6). In the predicate focus construction, the pragmatic presupposition is the availability of a referent as the topic and the pragmatic assertion is the content predicated of this topic. Finally, in sentence focus constructions, the pragmatic assertion is the proposition, introducing an event.

(6) Pim saw [MIA]^{F}
$\rightsquigarrow$ presupposition: 'pim saw x'
(= open proposition)
$\rightsquigarrow$ assertion: 'x = mia'
(= identification)

In focus negation, the negation operator targets the identification, i.e., the pragmatic assertion, and not merely the content of the focal constituent. According to this view, at the level of the interpretation of the sentence, the semantic content and the information structural interpretation are represented at distinct, but yet related levels. To model the major types of negation, 'sentential negation' and 'focus negation', a grammatical model is required that provides access to these levels and that explains the relation between syntactic structure, semantic content and information structural interpretation. Role and Reference Grammar (RRG; Van Valin and LaPolla, 1997; Van Valin, 2005) is a linguistic theory that offers the sufficient means to satisfy these above requirements.

RRG is a surface oriented grammar theory, developed from a strong typological and theoretical perspective. One of the theory's main aim is to capture both the universal characteristics of natural languages and the given language specific features. The general architecture of RRG is modular, with different levels of representation called 'Projections' and well-defined linking relations between them to model the interfaces. The syntactic representation (the layered structure of the clause, Figure 2) captures universal notions in terms of predicate-argument relations, as well as language-specific aspects in terms of special syntactic positions. The syntactic representation is given in two closely related projections, the 'Constituent Projection' and the 'Operator Projection'. The semantic representation is based on the classification of predicates by (Vendler, 1967) and adapted from the decompositional system of (Dowty, 1979). The center of the grammatical model of RRG is the bi-directional *linking algorithm* between the syntactic and the semantic representations, capturing both language production and comprehension.

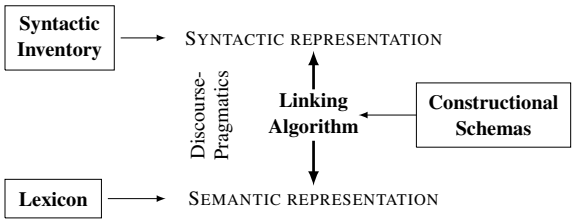


Figure 1: The general architecture of RRG

The universal properties of the clause structure are represented in the *layered structure of the clause* (see Figure 2), where the elements render semantically motivated universal characteristics of an utterance.

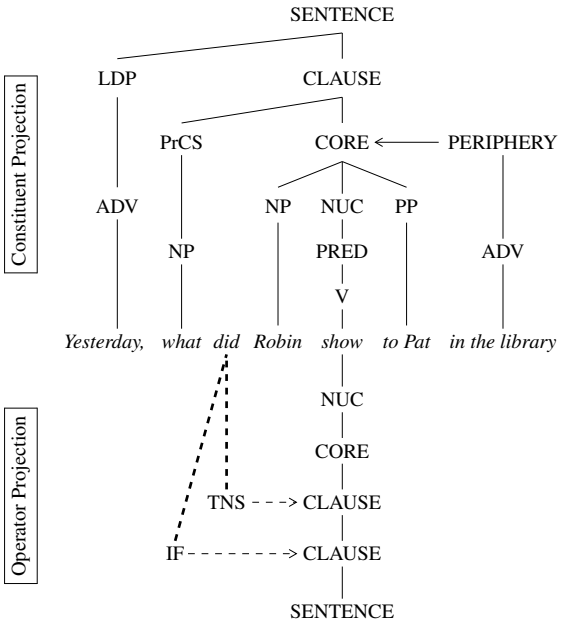


Figure 2: The layered structure of the clause in RRG

The basic elements of the layered clause structure are the NUCLEUS, containing the predicate, the CORE, containing the predicate and its core-arguments (both direct and oblique arguments), the PERIPHERIES, housing adjunct modifiers and adverbs and finally the CLAUSE, that contains the Core and the Peripheries. Next to these semantically motivated universal properties, there are also language-specific aspects represented in the syntactic structure. The presence of corresponding syntactic positions is language specific. Operators such as tense, aspect, modality and illocutionary force are not given in the constituent projection of the clause

but are represented in the separate 'Operator Projection'. The main layers can each be modified by one or more operators. The layered clause structure in RRG is motivated by theoretical and typological considerations, and as such it applies to different types of languages equally: to languages with fixed word order (e. g., English), to languages with free word order (e. g., Dyribal), to head-marking languages (e. g., Lakhota), to dependent-marking languages (e. g., Japanese), and so on.

In the general architecture of RRG, as part of the discourse pragmatics of the sentence, the focus structure is represented in a separate projection, called 'Focus Structure Projection'. Within this projection, RRG distinguishes the *actual focus domain* (AFD), the syntactic domain that corresponds to the *focus (domain)* in Lambrecht's terms, and the *potential focus domain* (PFD), where the focus can occur. Both syntactic domains include one or more *information units* (IU), which are the minimal phrasal units in the syntactic representation. The distinction between the PFD and the AFD is cross-linguistically relevant. Although in English, the PFD is always the entire clause, this is not generally the case in other languages. See, for example, Italian, where the PFD excludes any prenuclear elements (see Van Valin and LaPolla, 1997), or Hungarian, where the structural topic position is clause-internal, but external to the PFD. The information units are linked to syntactic domains in the constituents structure, and the focus domains include one or more information units. Hereby, it can represent the various focus structures. Figure 3 illustrates the RRG representation of narrow object focus and predicate focus respectively.

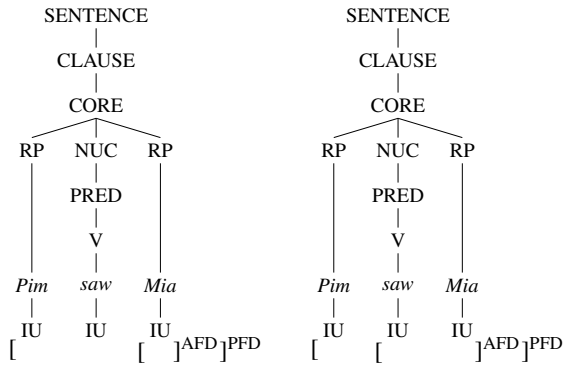


Figure 3: RRG's Focus Structure Projection

This representation shows the IUs, which are linked to syntactic domains in the constituent structure, and the focus domains, that each include one or more IUs. Hereby, it represents the various focus structures, as proposed by Lambrecht (1994). What is missing from this approach is the modeling of the interpretational effects of the different focus structures in terms of pragmatic structuring, which is crucial in the analysis of negation. We propose this extension in a formalized version of RRG (see Section 3.2). The extension requires a specification of the nature and role of information units, the ways of determining the presupposition-assertion distinction on basis of the focus structure, and its relation to the pieces of semantic information.

3.2 Modeling focusing and negation

In our proposal, we argue that negation operates on the pragmatic assertion, which is determined by the focus structure of the sentence. To capture this, pragmatic structuring needs to be derived, based on the given focus domains. This asks for an extension of the Focus Structure Projection. The information contained in the elements of the pragmatic structuring is derived on basis of the pieces of semantic information contributed by the constituents. This is essentially captured by the notion of 'information unit', which represents a given syntactic domain and its semantic content.

Our analysis is based on the theoretical developments of (classical) RRG and Lambrecht's theory of information structure, which both lack a precies formal definition. For the formal modeling and further extensions we use the formalized version of RRG (fRRG) as proposed by Kallmeyer et al. (2013) and Osswald and Kallmeyer (2018). fRRG has important advantages, of which a major one is that semantic composition is on a par with syntactic composition, i.e., semantic construction can be carried out compositionally. Syntactic templates come with (pieces of) semantic representations, given as *decompositional frames* (Petersen, 2015; Löbner, 2017), formally defined as *base-labelled typed feature structures* (Kallmeyer and Osswald, 2013). The nodes in the syntactic trees are provided with feature structures, containing interface features, which establish the link between syntax and semantics: they mediate between syntactic and semantic composition. The syntactic operations trigger the composition of the semantic representations, thereby deriving the meaning representation of the sentence. The semantic composition proceeds by *unification*. Figure 4 below illustrates the tree templates for deriving *Pim saw Mia* before

composition. By combining the trees templates (via substitution here), the feature structures are unified and the meta-variables are identified (e.g., $\boxed{1} = x$). The semantic representation of the final tree is calculated by unification of the semantic content of the participating trees.

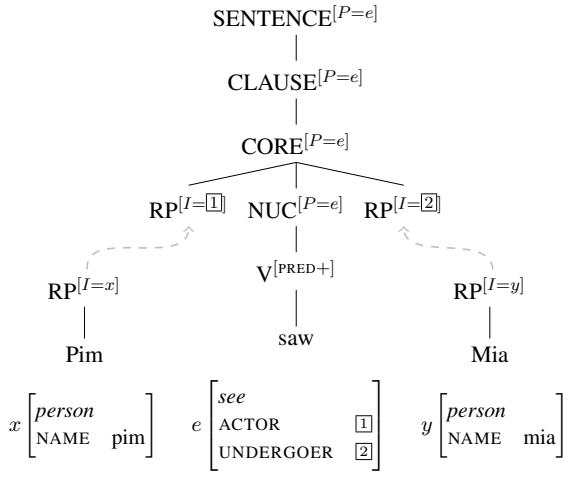


Figure 4: Syntax-semantics interface in fRRG

Recall that within the syntactic structure, operators (e.g., negation, tense) are represented in the separate 'Operator Projection'. In the linearization *Pim did not see Mia*, negation is analyzed as a core-operator. In the semantics, this leads to an operator that is applied to the content of the domain in the CORE, i.e., the whole proposition.[2]

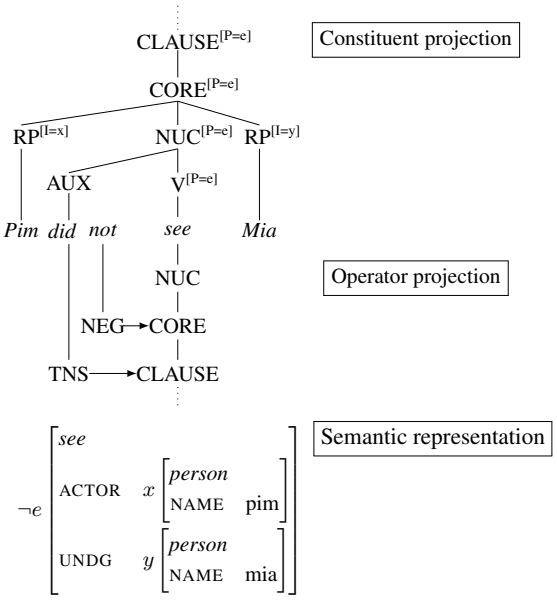


Figure 5: Syntactic and semantic projections

[2]The proper definition of negation in frame semantics is beyond the scope of this paper. The representation by ¬ here is simplified.

The difference between the 'focus negation' and the 'sentence negation' interpretation lies in the respective information structures of the sentences. Their semantic representation is the same. To capture this, we argue that the contribution of negation to the interpretation of the sentence also enters information structure. Hence, for the full analysis, we must extend Figure 5 with the representation of the information structure of the resepective utterance, where negation also plays a crucial role. We argue that negation operates on the pragmatic assertion, i.e., on the newly conveyed information, not on the semantic representation of the focused constituent. It is represented within the 'Information Structure Projection',[3] which contributes the context-dependent meaning component of the sentence. The pragmatic assertion is determined by the focus structure of the sentence, that contains the information units and the different focus domains. Following Van Valin (2005), we distinguish the actual focus domain (AFD) and the potential focus domain (PDF). Additionally, we also represent the non-focus domain (NFD), that can be straightforwardly derived based on the AFD/PFD structure. The information units have a central role establishing the link between the syntactic domains and the corresponding semantic content. The IUs are linked to the syntactic structure by features on the respective nodes, and to the pieces of semantic content of these syntactic domains. The focus structure is a triple of the focus domains: AFD, PFD and NFD. These focus domains are represented as sets of information units, and provide the focus-background division: the focus is the unification of the semantic content of the IUs in the AFD, while the background is the unification of the semantic content of the IUs in the NFD.[4] The communicative function, i.e., the newly conveyed information is dependent on the focus structure, and defined as a special relation between the focus and the back-

[3]The 'Information Structure Projection' is an extension of RRG's 'Focus Structure Projection' proposed by Balogh (2021), which not only represents the focus structure, but also the topic-comment division. This is necessary for a comprehensive representation of the information structure of the sentence, and also for capturing various linguistic phenomena where focus structure and topic structure interact, e.g., the linearization constraints of *also*. In order to simplify the representations here, we only give the focus structure, that is directly relevant to our discussion regarding negation. However, keeping in mind that the projection contains more, we keep referring to it as 'Information Structure Projection'.

[4]Preserving the specifications of the meta-variables as determined by the syntax-semantics interface; see Figure 4.

ground. In case of a narrow focus construction, this relation is the 'identification' between the focus and the missing information in the background, the open proposition. This equals the 'pragmatic assertion' in Lambrecht's (1994) terms, while the 'pragmatic presupposition' is the same as the background in the focus-background division. Figure 6 below illustrates the extended 'Information Structure Projection' for the sentence in (7) with narrow (object) focus structure above its syntactic and semantic representations given in Figure 5.

(7) Pim did not see [MIA]^{F}

Information Structure Projection (for (7))

information units: $\{IU^x, IU^y, IU^e\}$

focus structure:

$\langle AFD, PFD, NFD\rangle = \langle\{IU^y\}, \{IU^x, IU^y, IU^e\}, \{IU^x, IU^e\}\rangle$

focus-background division:

$$\left\langle\ y\begin{bmatrix} person & \\ \text{NAME} & \text{mia}\end{bmatrix},\ e\begin{bmatrix} see & \\ \text{ACTOR} & x\begin{bmatrix} person & \\ \text{NAME} & \text{pim}\end{bmatrix} \\ \text{UNDG} & \boxed{2}\end{bmatrix}\ \right\rangle$$

pragmatic assertion: $\text{NEG}(\boxed{2} = y)$

Figure 6: Information structure projection of (7)

Modeling of narrow subject focus (8) is straightforward. Note that the syntactic and semantic structures, as well as the information units are equivalent in example (7) and example (8). The difference is in the focus-background division, which derives the different content of the identification. Straightforwardly, the relation between focus and background is of the same nature for both (i.e., identification).

(8) [PIM]^{F} did not see Mia

Information Structure Projection (for (8))

information units: $\{IU^x, IU^y, IU^e\}$

focus structure:

$\langle AFD, PFD, NFD\rangle = \langle\{IU^x\}, \{IU^x, IU^y, IU^e\}, \{IU^y, IU^e\}\rangle$

focus-background division:

$$\left\langle\ x\begin{bmatrix} person & \\ \text{NAME} & \text{pim}\end{bmatrix},\ e\begin{bmatrix} see & \\ \text{ACTOR} & \boxed{1} \\ \text{UNDG} & y\begin{bmatrix} person & \\ \text{NAME} & \text{mia}\end{bmatrix}\end{bmatrix}\ \right\rangle$$

pragmatic assertion: $\text{NEG}(\boxed{1} = x)$

Figure 7: Information structure projection of (8)

The above approach correctly captures the meaning contribution of 'focus negation', where the negation operator takes narrow scope, and in the interpretation it applies to the identification evoked by narrow focus. As such, it is represented within the Information Structure Projection as well, rather than merely in the semantic representation of the sentence. Based on this analysis, an important question arises, how to capture 'sentential negation', which is standardly analyzed as the negation operator directly applies to the semantic content of the predication. We argue that in sentential negation, negation also applies within the information structure projection, targeting the communicative function. In *Pim did not see Mia*, without narrow focus, the negation operates on the predication. The underlying sentence has a broad (predicate or sentence) focus structure. In both, the AFD contains the predicate, the difference is in the topic-comment distinction. For our concerns here, the determinant aspect is whether the predicate is part of the AFD, so regarding space limitations, the precise characterization of the effects of the topic structure is left for further discussion. In broad focus structures, the pragmatic assertion of the sentence is the statement that the event described by the frame as the semantic representation exists, and must be added to the common ground. When negation applies to broad focus, it targets this pragmatic assertion, stating that the event represented by the frame does not exist. For *Pim did not see Mia* with a broad focus structure, the syntactic structure, the semantic representation and the IUs are the same as before. The interpretational difference is due to the different focus structure and the corresponding communicative function.

Compared to the narrow focus structures in Figure 6 and 7, the information units are the same, but the focus structure is different, as the information unit corresponding to the predicate is part of the focus. This in turn leads to a different pragmatic structuring, and a different communicative function, where the relation between focus and background is not of an 'identification', but stating the existence of the event. As before, this contribution is targeted by negation. In the focus-background division, the actual focus domain contains the whole proposition, and the background is either empty, or rather contains a contextual restriction ('Restr' above) to some time-/space-frame or alike, relative to which the (non-)existence of such an event is pragmatically asserted.

(9) [Pim did not see MIA]^{F}

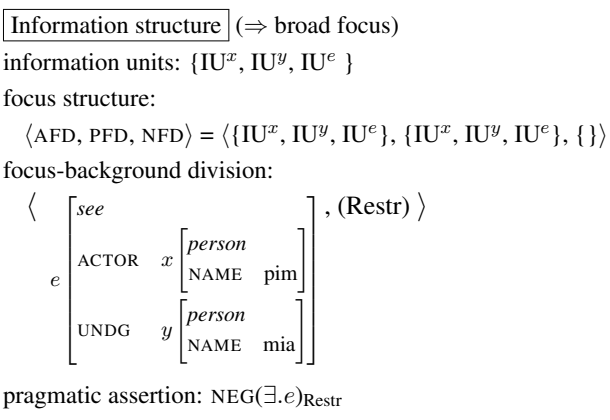


Figure 8: Information structure projection (broad F)

4 Conclusion and further issues

The paper addressed a surprisingly underrepresented linguistic phenomenon, 'focus negation', where the crucial issue is how to link the logical semantic understanding of negation as a unary propositional operator and the meaning contribution of negation operating on a single (non-propositional) constituent. Although this type of negation is generally acknowledged, an analysis and formal modeling of it is still missing. The issue is not straightforward, as it goes beyond the mere semantics of the sentence, and asks for an approach where information structure, in particular the focus structure, of the sentence interacts with negation at the syntax-semantics interface.

In this paper, we introduced a proposal towards a grammatical model that captures 'focus negation' and 'sentential negation' uniformly, in an information structure based perspective. The meaning component of the sentence is an interplay between the semantic representation, a mental representation/description of an event, and the information structural interpretation given in terms of pragmatic structuring. We proposed a two-level approach, where negation has access to and operates on the pragmatic assertion, rather than it merely enters the semantic representation. The proposal offers a way to capture 'sentential negation' and 'focus negation' in a uniform way, correctly dealing with the interpretation of the latter type as well. In the proposed grammatical model, semantic representations are given as decompositional frames, which are descriptions/minimal models of events.

The grammatical model we proposed is based on solid theoretical grounds as given by Role and Reference Grammar (Van Valin and LaPolla, 1997; Van Valin, 2005), formally defined using Tree-Wrapping Grammar (Kallmeyer et al., 2013; Osswald and Kallmeyer, 2018) and decompositional frames (Petersen, 2015; Löbner, 2017; Kallmeyer and Osswald, 2013). For the analysis we proposed the necessary extensions to the framework, regarding both the theoretical and the modeling side.

We proposed here the basic ideas of a uniform analysis and formal modeling of the two types of negation. Nevertheless, there are still several issues to resolve for a comprehensive analysis of natural language negation and the interface between syntax, semantics and information structure (i.e., discourse pragmatics). From the theoretic point of view the most urgent issue is how to analyze the relation between the contribution of negation in semantics and in information structure. Furthermore, we must extend the analysis for further constructions, in particular for constructions where the focus falls on the verb (i.e., narrow verb focus), where it falls on a constituent within a complex noun phrase (e.g., determiner, adjective, preposition and so on), constructions with multiple foci, and the relation between focus, negation and other scope taking elements. From a more structural perspective, we must extend the analysis to languages where the two types are distinguished in the morphosyntactic structure (e.g., Hungarian). These issues and further theoretical considerations are left for further investigation and development of the current proposal.

References

Kata Balogh. 2021. Additive particle uses in Hungarian: a Role and Reference Grammar account. *Studies in Language*, 45(2):428–469. Online-first: 2020.

David I. Beaver and Brady Z. Clark. 2008. *Sense and Sensitivity: How Focus Determines Meaning*. Wiley-Blackwell.

Mary Dalrymple and Irina Nikolaeva. 2011. *Objects and Information Structure*, volume 131 of *Cambridge studies in linguistics*. Cambridge University Press, Cambridge.

Karen De Clercq. 2020. Types of Negation. In Viviane Déprez and M. Teresa Espinal, editors, *The Oxford Handbook of Negation*, pages 58–74. Oxford University Press, Oxford.

David Dowty. 1979. *Word meaning and Montague Grammar*. Reidel, Dordrecht.

Katalin É. Kiss, editor. 1995. *Discourse Configurational Languages*. Oxford University Press, Oxford.

Elisabet Engdahl and Enric Vallduví. 1996. Information packaging in HPSG. *Edinburgh Working Papers in Cognitive Science*, 12: Studies in HPSG:1–32.

Otto Jespersen. 1917. *Negation in English and Other Languages*. A. F. Høst & Søn, Copenhagen.

Aravind K. Joshi and Yves Schabes. 1997. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, pages 69–123. Springer, Berlin.

Laura Kallmeyer and Rainer Osswald. 2013. Syntax-Driven Semantic Frame Composition in Lexicalized Tree Adjoining Grammars. *Journal of Language Modelling*, 1(2):267–330.

Laura Kallmeyer, Rainer Osswald, and Robert D. Van Valin, Jr. 2013. Tree Wrapping for Role and Reference Grammar. In *Proceedings of Formal Grammar 2012 and 2013*, pages 175–190, Berlin. Springer.

Edward Klima. 1964. Negation in English. In Jerry A. Fodor and Jerrold Katz, editors, *The Structure of Language*, pages 246–323. Prentice Hall, Englewood Cliffs, NJ.

Ekkehard König. 1991. *The Meaning of Focus Particles: A Comparative Perspective*. Routledge, London/New York.

Manfred Krifka. 2001. For a structured meaning account of questions and answers. In C. Féry and W. Sternefeld, editors, *Audiatur Vox Sapientia: A Festschrift for Arnim von Stechow*, pages 287–319. Akademie Verlag, Berlin.

Knud Lambrecht. 1994. *Information structure and sentence form*. University Press, Cambridge.

Sebastian Löbner. 2017. Frame theory with first-order comparators: Modeling the lexical meaning of punctual verbs of change with frames. In H. H. Hansen, S. E. Murray, M. Sadrzadeh, and H. Zeevat, editors, *Proceedings of the Eleventh International Tbilisi Symposium on Language, Logic, and Information*, (LNCS 10148), pages 98–117. Springer, Heidelberg.

Rainer Osswald and Laura Kallmeyer. 2018. Towards a Formalization of Role and Reference Grammar. In R. Kailuweit, L. Künkel, and E. Staudinger, editors, *Applying and Expanding Role and Reference Grammar*, (NIHIN Studies), pages 355–378. Albert-Ludwigs-Universität, Universitätsbibliothek, Freiburg.

Wiebke Petersen. 2015. Representation of concepts as frames. In Th. Gamerschlag, D. Gerland, R. Osswald, and W. Petersen, editors, *Meaning, Frames, and Conceptual Representation*, (Studies in Language and Cognition 2.), pages 43–67. Düsseldof University Press.

Craige Roberts. 2012. Information Structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1:75–116.

Robert Stalnaker. 1974. Pragmatic presuppositions. In M. Munitz and P. Unger, editors, *Semantics and Philosophy*, pages 197–214. New York University Press, New York.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA.

Mark Steedman. 2019. Combinatory Categorial Grammar: A Comparative Handbook. In András Kertész, Edith Moravcsik, and Csilla Rákosi, editors, *Current Approaches to Syntax*, volume 3 of *Comparative Handbooks of Linguistics (CHL)*, pages 389–420. De Gruyter Mouton, Berlin/Boston.

Enric Vallduví. 1990. Information Structure and Scope of Sentential Negation. In *Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society*, pages 325–337. Berkeley Linguistics Society.

Robert D. Van Valin, Jr. 2005. *Exploring the syntax-semantics interface*. University Press, Cambridge.

Robert D. Van Valin, Jr. and Randy J. LaPolla. 1997. *Syntax: Structure, Meaning and Function*. University Press, Cambridge.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca.

Conversation and mood in European Portuguese

Rui Marques
Faculdade de Letras, CLUL, Universidade de Lisboa
rmarques@letras.ulisboa.pt

Abstract

The literature on mood in Romance languages has identified the conditions that lead to the use of Subjunctive or of the Indicative moods. For syntactic contexts where only one of these moods is allowed, its obligatoriness follows from the semantics of the main clause, but, in cases of mood choice, the option is pragmatically driven. This paper focuses on cases of mood choice in European Portuguese, presenting data that suggests that mood choice is conditioned by the intended effect on the conversational context.

1 Introduction

A lot of debate has been devoted to the relation between (non) assertion and the Indicative and Subjunctive moods in Romance languages. The old idea that Indicative occurs in assertive sentences and Subjunctive in non-assertive contexts seems to account for most data concerning non-complement clauses, but faces important problems when complement clauses are taken into consideration. Accounts of mood in formal semantics allowed the understanding of the rationality lying at the opposition between Indicative and Subjunctive with no reference to assertion being needed. However, in some cases there seems to exist an undeniable relation between mood choice and assertion.

This paper proposes to investigate the relation between mood, context set, and dynamics of discourse, focusing on complement and adverbial clauses of European Portuguese (EP). Instead of focusing solely on the contrast between the Indicative and Subjunctive moods, the opposition between finite and infinitive clauses will also be considered, providing new insights concerning the relation between mood and dynamics of conversation.

The next section presents the traditional idea that Indicative occurs in assertive contexts, Subjunctive in non-assertive ones, and problems for it. Section 3 summarizes the conditions for the use of Indicative or Subjunctive in EP. These conditions account for the cases of lexically selected mood and for the interpretations obtained in cases of mood choice, which will be further deepened in sections 4 and 5. Notes towards a formalization of the given observations are presented at section 6 and Appendix contains authentic examples of the analyzed constructions.

2 Mood and (non) Assertive Speech Acts

Two traditional widespread ideas concerning the Indicative and Subjunctive moods in Romance languages are the *realis/irrealis* proposal and the assertion/non-asserion hypothesis. The first one claims that the Indicative/Subjunctive opposition mirrors the *realis/irrealis* distinction, Indicative occurring in sentences describing reality and Subjunctive in sentences linked to a virtuality level, such as those describing desires, possibilities, orders, and so on. This idea, though explaining the obligatoriness of the Indicative in different kinds of sentences that describe reality (e.g., complement clauses of predicates like the equivalents of *know* or *find out*, declarative unsubordinated clauses, causal clauses), faces two major problems: (i) it does not explain the selection of Indicative by fiction predicates, such as (the equivalents of) *dream*; (ii) it does not explain Subjunctive in complements of factive-emotive predicate, as, e.g., the equivalent of *regret*, as well as the obligatoriness of this mood in other factive contexts, as in (1) or (2):

(1) embora {esteja$_{\text{-SUBJ}}$ / *está$_{\text{-IND}}$} a chover, o dia está agradável
'Although it is raining, the day is pleasant'

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

(2) eles conseguiram que a reunião {fosse$_{\text{-SUBJ}}$ / *foi$_{\text{-IND}}$} adiada
'They managed the meeting to be postponed'

Concerning complement clauses of factive-emotive predicates, an ancient idea to explain the occurrence of Subjunctive in this context resorts to the distinction between assertion and presupposition, Indicative being the mood of assertion and Subjunctive the mood of non-assertion. This idea, which dates back at least to Hooper 1975, is grounded on the observation that the assertion of (3) will only be felicitous in a context where the speaker assumes that the complement clause belongs to the common ground:

(3) lamento que o teu gato {tenha$_{\text{_SUBJ}}$ morrido / *morreu$_{\text{_IND}}$}
'I regret that your cat has died.'

The occurrence of Subjunctive in this context is then explained as following from the fact that the complement clause is presupposed, not asserted.

However, the proposal that Indicative occurs in contexts of assertion, Subjunctive being the mood of non-assertion, faces several problems, among which the very concept of (non-)assertion[1]. Concerning complement clauses, the proposal that the Indicative and Subjunctive moods occur, respectively, in assertive and non-assertive contexts amounts to say that the main clause's predicate is assertive (hence an Indicative ruler) or non-assertive (hence a Subjunctive ruler). In EP, a group of verbs, as *acreditar* ('to believe'), accept both the Indicative as the Subjunctive in the complement clause:

(4) A: Achas que vamos ganhar o jogo?
'Do you think we will the match?'
B: a. Acredito que podemos$_{\text{_IND}}$ ganhar.
b. Acredito que possamos$_{\text{_SUBJ}}$ ganhar
'I believe we might win'

In this kind of sentences, the choice between Indicative and Subjunctive is dependent on the degree of belief being expressed. In (4), by choosing the Indicative, the speaker indicates that, in his opinion, there is a good possibility of winning the game, while the choice of Subjunctive indicates that such possibility is unlikely. In other words, Indicative signals a high degree of belief, Subjunctive a lower degree. Now, if the choice of mood is conditioned by the assertion/non-assertion opposition, (4) shows that the same predicate can be assertive or non-assertive. Given that the difference between (4a) and (4b) is the degree of belief being conveyed, it follows that assertive predicates (whose complement will be in the Indicative) will be the ones that express a full (or at least a high) commitment with the truth of the complement clause. Such is the case of factive-emotive predicates, which express the information that the attitude holder takes the complement proposition to be true, but they select the Subjunctive (see (3)). Hence, the assertion/non-assertion hypothesis faces the same problems as the *realis/irrealis* hypothesis.

A more reasonable interpretation of what is an assertive predicate would be based on the Stalnakerian concept of assertion (roughly, assertion of *p* is the addition of *p* to the common ground): assertive predicates will be the ones whose complement clause can be added to the common ground. However, one can easily think of examples of Indicative ruler's predicates whose complement clause is presupposed, not added to the common ground, in the same way as will be the case with factive-emotive predicates, as in (3). For instance:

(5) a. Todos sabemos que vamos$_{\text{_IND}}$ morrer um dia.
'We all know that we will die some day'

b. «Nós todos sabemos que o fumo é prejudicial, não é?»
(CETEMPÚBLICO *ext471815-nd-96b-2*)
'We all know that smoking is harmful, right?'

Conversely, it is also easy to find examples of Subjunctive rulers whose complement proposition is presented as new information, to be added to the common ground. For instance, consider the following example as part of a story which the public is hearing/reading for the first time:

(6) A situação era desesperante e muitas pessoas pensavam que nunca iriam sair dali. **Foi preciso que a tempestade passasse$_{\text{-SUBJ}}$** para que **o avião conseguisse$_{\text{-SUBJ}}$ finalmente levantar voo**! Mas **os esforços do piloto não impediram que um raio atingisse$_{\text{-SUBJ}}$ o avião**.
'The situation was hopeless and many people thought they would never leave. **It took the**

[1] Another major problem, as Palmer 1986 points, is that interrogatives are obviously non-assertive contexts (whatever concept of assertion is considered), thus the proposal not explaining the obligatoriness of Indicative in interrogatives, as *Que horas são?* ('what time is it?').

storm to pass for the plane to finally take off! But **the pilot's efforts did not prevent lightning from striking the plane**.'

In sum, the use of the Indicative or of the Subjunctive in complement clauses does not seem to be triggered by the issue of whether the complement proposition does or does not belong to the common ground. Both moods can occur in sentences that are taken to be part of the common ground prior to their utterance as in sentences that convey new information. Hence, an approach that bases the selection of mood on the kind of speech act doesn't seem tenable. Instead, the choice between one and another mood in complement clauses seems to be semantically driven, following primarily from the lexical meaning of the main predicate, not dependent on pragmatic issues.

Still, in some constructions the option for the Indicative or the Subjunctive mood is conditioned by whether the complement proposition is or is not presented as taken to be part of, or to be integrated in, the common ground. Such is the case of sentences as the following, which express a contrast between the speaker's belief at utterance time and his previous belief, in (7a), or someone else's belief, in (7b):

(7) a. Naquela altura, eu não acreditava que os Vikings chegaram$_{\text{-IND}}$ à América.
'At that time, I didn't believe that the Vikings reached America.'

b. Ele não acredita que os Vikings chegaram$_{\text{-IND}}$ à América.
'He does not believe that the Vikings reached America'

These sentences convey the information that, according to the speaker, the complement proposition is true. In the same kind of sentences, the Subjunctive might also occur, but, then, the truth of the complement proposition is not conveyed (i.e., such proposition might be true or false, no commitment with its truth value being conveyed):

(8) a. Naquela altura, eu não acreditava que os Vikings tenham$_{\text{-SUBJ}}$ chegado à América.
'At that time, I didn't believe that the Vikings reached America.'

b. Ele não acredita que os Vikings tenham$_{\text{-SUBJ}}$ chegado à América.
'He does not believe that the Vikings reached America'

Hence, in this kind of construction, by choosing the Indicative for the complement proposition, the speaker presents such proposition as one that belongs, or is to be added, to the common ground, while the choice of the Subjunctive states merely a negative epistemic state.

To summarize, the hypothesis that the Indicative/Subjunctive opposition mirrors the assertion/non-assertion distinction is too naïf to be an explanation for the distribution of these moods in EP (or in other Romance languages, presumably), but data as (7) and (8) show that some relation exists between mood and assertion. Thus, an account of mood in EP has to explain why is the Indicative obligatory in some clauses and the Subjunctive in others, despite the status of the proposition concerning its relation to the common ground, while in other cases the choice between one and another mood is grounded on whether the speaker intends to add the proposition to the common ground. In the following section, a semantic explanation for the first issue, detailed in Marques 2022, will be synthesized, after what, in the following section, the second issue will be resumed.

3 Indicative *vs* Subjunctive

The reason for some predicates to be Subjunctive rulers (i.e., the Subjunctive might occur in their complement clauses, the Indicative might not) and others to be Indicative rulers is nowadays understandable and can be expressed in a simple sentence (a slight amendment, justified and presented below, will be needed): Indicative is selected by those predicates whose meaning leads to consider only *p*-worlds (i.e., worlds where the proposition *p* is true), while the Subjunctive is selected by those predicates whose meaning leads to take into account (also) non-*p* worlds. Descriptively, the Indicative occurs in those sentences that are taken to be true and an epistemic or doxastic attitude is expressed towards them, otherwise (i.e., if the proposition is not presented as accepted to be true or if the attitude towards it is not an epistemic or doxastic attitude) the Subjunctive occurs. This explains why Subjunctive is selected by non-veridical predicates, as, e.g., predicates of desire (as the equivalents of *want*, *prefer*, etc.), deontic predicates (as the equivalents of *order*, *suggest*, etc.), modal predicates (as the equivalents of *be possible*, *be probable*, etc.), among others. Such predicates are non-veridical (in the sense of Giannakidou 1994, and several other texts of her), not expressing anyone's

compromise with the truth of the complement proposition. It also explains why the Indicative is selected by several veridical predicates, such as the equivalents of *know*, *verify*, *find out*, and others, which express an attitude of knowledge concerning the complement proposition, the equivalents of doxastic predicates as, e.g., *think*, or the equivalents of *verba dicendi*, as, e.g., *say*, *confess* or *assure*, and the equivalents of fiction predicates, as, e.g., the equivalents of *dream*. All these predicates indicate that the complement proposition is true in the model towards which it is evaluated. Such model is the one introduced by the main clause's predicate: a model that represents the epistemic state of the attitude holder, in the case of predicates like *think* and *verba dicendi*, the model that represents John's dream in a sentence like *last night, John dreamed that he was in Australia*, and so on. The most problematic cases are the Subjunctive clauses that describe facts. This is the case of complement clauses of factive-emotive predicates, as the equivalents of *regret*, *irritate*, *surprise*, and many others, as well as it is the case of complement clauses as those in bold in (6), above, and also of concessive clauses, where in EP the Indicative is also ruled out, as shown by (1), above. However, as synthesized in the following paragraph, the meaning of all of these constructions also involves the consideration of non-*p* worlds, which explains the obligatoriness of the Subjunctive.

Of all the cases where the Subjunctive is obligatory (in EP, allowed in other Romance languages) in sentences that describe facts, the most debated case is the one of complement clauses of factive-emotive predicates. The most common explanation for why these predicates take (in EP, accept in other languages) the Subjunctive is that they are gradable predicates, whose meaning leads to consider alternatives (see Villalta 2008; Godard 2012, Giannakidou & Mari 2016, 2021, a.o.). For instance, to say '*x* regrets that *p*' means that *x* would prefer if *p* were not true; one cannot say that 'it is fair that John resigned' without thinking of alternative worlds where John did not resign, and so on. However, as observed in Marques 2022, gradability does not explain all the cases where Subjunctive occurs in sentences describing facts. It explains, however, the selection of Subjunctive by some factive-emotive verbs, as *lamentar* ('*regret*'), *gostar* ('like') or *merecer* ('deserve'), as well as by adjectival predicates as the equivalents of *be* (*un*)*fair*, *be normal/strange*, and so on. Concerning factive verbs as the equivalents of *surprise*, *irritate*, and others, predicates whose argument structure is different from the preceding ones, the proposal was made that these are Subjunctive rulers because they express a causal relation. For instance, to say that 'Ana is surprised that it is raining' means that the fact that it is raining caused surprise on Ana. Given that, according to counterfactual theories of causality (see Lewis 1973, Salmon 1998, a.o.), causality involves the consideration of alternatives – A caused B means that if A had not occurred, all the rest being the same, B would not have occurred either –, the reason for these predicates to be Subjunctive rulers follows straightforwardly: their meaning involves counterfactual reasoning, leading to the consideration of non-*p* worlds (worlds where the complement proposition is false), hence they are Subjunctive rulers. The same explanation is extendable to the fact that Subjunctive is selected by predicates that express a necessary (as the equivalents of *be needed*) or a sufficient condition (as the equivalents of *be enough*). To say that, e.g., 'we had to climb the mountain to reach our destination' means that, if we had not climbed the mountain, all the rest being the same, we would not have reached the destination. Likewise, to say that, e.g., 'just a few drops of rain were enough for people to start leaving the stadium' means that, if no drop of rain had fallen, all the rest being the same, people might have not left the stadium.

As for concessive clauses, where the Subjunctive also occurs even if this is a veridical context, the proposal was made that this follows from the fact that concessive constructions express the information that an expectation following from *p* does not hold in every possible world that forms the context set. For instance, *the room is cold, although the heater is turned on* expresses the denial of expectation that the room is warm, an expectation that follows from the concessive clause.

Hence, the conditions for the use of Indicative or Subjunctive in EP can be stated as follows: if the (syntactic) context where a sentence S occurs leads to consider only worlds where S is true and the inferences (including conversational implicatures) following from S hold, the verb of S inflects in the Indicative; if the (syntactic) context where S occurs leads to consider worlds where S is false or where an inference following from S does not hold, the verb of S inflects in the Subjunctive.

This explanation accounts for the cases where only one of the Indicative and Subjunctive moods

is allowed as well as for cases where either of these moods may be used, as is the case of (7) and (8), above. In (7), the Indicative is used because the speaker describes his own opinion concerning the complement proposition, stating that his epistemic state at utterance time contains only worlds where such proposition is true. In (8) the Subjunctive is used because the speaker describes only the opinion of the attitude holder, stating that his epistemic state contains only worlds where the complement proposition is false. Thus, the Indicative is a mark that signals the consideration of only *p*-worlds, the Subjunctive one that signals that non-*p* worlds or worlds where an inference from *p* does not hold are to be considered.

In many cases, it is the meaning of the main clause's predicate (or, in the case of non-complement clauses, the meaning of the conjunction, or of another sentential operator) that leads to consider only *p*-worlds or (also) non-*p* worlds. But it (7) the use of the Indicative for the complement clause follows from pragmatics, not from the compositional meaning of the construction, which leads to the use of the Subjunctive, as in (8). The contrast between (7) and (8) provides sense to the traditional idea that Indicative is the mood of assertion, Subjunctive the mood of non-assertion: the speaker chooses between one or the other mood depending on whether he asserts the complement clause or not. Resorting to the Indicative is a device the speaker can use to signal that the complement clause belongs to (or is to be added to) the common ground.

In the two next sections the relation between mood choice and common ground will be deepened.

4 The case of (negative) epistemic commitment

In EP, the choice between Indicative and Subjunctive moods for complement clauses is available whenever the main clause is negative and the main predicate expresses a doxastic attitude:

(9) ele não {acredita / pensou / acha / disse / duvida / admite /…} que {tinha$_{_IND}$ / tivesse$_{_SUBJ}$} perdido as eleições!
'He {does / did} not {believe / thought / think / said / doubt / admit / …} that he has lost the elections!'

In all these cases, the use of Indicative indicates that the complement clause is true, according to the speaker, and is part of, or is to be added to, the common ground, while the use of the Subjunctive does not indicate what is the speaker's opinion concerning the truth value of the complement clause. Mari 2016 claims that in Italian the same kind of factor lies at the mood choice for the complement clause of *credere* ('believe') in affirmative sentences. According to her, (10a), with the Indicative, merely expresses the attitude holder's opinion concerning the truth of the complement clause, the question of whether such sentence is, in fact, true not being at issue, while the assertion of (10b), with the Subjunctive, presents the complement clause as a candidate to integrate the common ground:

(10) a. Gianni crede che Maria è$_{_IND}$ malata.
b. Gianni crede che Maria sia$_{_SUBJ}$ malata.
'Gianni believes that Mary is sick'

This proposal cannot be extended to EP, a language where, like in Italian, both the Indicative and the Subjunctive might occur in complement clauses of *acreditar* ('believe') in affirmative sentences. Regardless of whether the complement clause exhibits the Indicative or the Subjunctive, the sentence might merely describe the epistemic state of the attitude holder, as in (4), above, as it might be uttered in a context where the truth value of the proposition as a matter of fact is at stake, as shown by the following example:

(11) A: Did John really wrote that letter?
B: eu acredito que {escreveu$_{_IND}$ / tenha$_{_SUBJ}$ escrito}
'I believe he wrote / might have wrote'

Another piece of evidence that Mari's proposal is not extendable to EP comes from examples as (12):

(12) Ainda não acredito que ganhei$_{_IND}$!
'I still don't believe that I won!'

In these negative *believe*-clauses, where the main clause's subject identifies the speaker, the Indicative in the complement clause is only possible with a certain intonation showing surprise. This construction indicates that the complement proposition describes a fact. Clearly, the resort to the Indicative does not indicate that only the private epistemic state of the attitude holder is being described, as Mari claims to be the case in Italian, but that the complement proposition belongs (or is to be added) to the common ground.

To summarize, in EP, doxastic predicates accept both the Indicative as the Subjunctive in the com-

plement clause. In the case of affirmative sentences, the choice between one or the other mood depends on the degree of belief being expressed, the Indicative signaling a high, the Subjunctive a low, degree of belief (in other words, if the epistemic state of the attitude holder contains only *p*-worlds, the Indicative is used; if such epistemic state contains non-*p* worlds, the Subjunctive is used). In negative sentences, since a low degree of belief (the null degree) is expressed, the Subjunctive is the obvious mood, but the Indicative might also be used, to convey the information that, unlike what the attitude holder believes / believed at a previous time, the complement proposition is true. In other words, concerning doxastic predicates in EP, only in those cases where the main clause is negative and the complement clause is in the Indicative is the complement proposition presented as describing a fact; i.e., the complement clause is interpreted as if it were an independent clause. Hence, these cases – negative clauses with doxastic predicates and Indicative in the complement clause – are instances where two discourse units – the main clause and the embedded proposition – are at stake. In other words, two models are considered in the interpretation of the complement clause: the model representing the attitude holder's beliefs (at a previous time) and the one representing the speaker's belief (at utterance time) / the information shared by the participants in the conversation.

The conditions for the use of the Indicative or the Subjunctive moods provided on section 3 are coherent with these occurrences of the Indicative: by resorting to the Indicative, the speaker conveys the information that his epistemic state (and, presumably, the one of the other participants in the conversation) contains only *p*-worlds. In addition, the construction at stake shows that, at least in these cases, there is a relation between mood and discourse updating. Seeking to deepen the understanding of the relation between mood and context of assertion, in the next section infinitival clauses will also be brought into consideration.

5 Finite *vs* Infinitival clauses

As, e.g., Portner 1997 observes, in many cases where both an infinitival or a finite clause might occur there is no obvious semantic difference between the two constructions, as shown by the following examples:

(13) a. Penso chegar$_{\text{INF}}$ a tempo.
b. Penso que chego$_{\text{IND}}$ a tempo.
'I think I will arrive on time'

(14) a. Esperemos conseguir$_{\text{INF}}$ chegar lá!
b. Esperemos que consigamos$_{\text{SUBJ}}$ chegar lá!
'Let's hope we manage to get there'

By contrast, in other cases, the choice between an infinitival and a finite clause has semantic import, as shown by (15):

(15) a. É possível cultivar$_{\text{INF}}$ lá uvas.
'It's possible to grow grapes there'
b. É possível que se cultivem$_{\text{SUBJ}}$ lá uvas.
'It's possible that grapes grow there'

In (15b) the modal predicate has only an epistemic reading, which is unavailable in (15a). This shows that the option between a finite or an infinitival clause is not always a matter of free choice. Also in different kinds of adverbial clauses differences of interpretation are observable between infinitival and finite clauses.

5.1 *Before* and *until*-clauses

In EP, the verb of temporal clauses introduced by the equivalent of *before* or *until* may inflect in the Infinite or in the Subjunctive mood:

(16) a. Emigrou antes de a guerra começar$_{\text{INF}}$.
'(S)he emigrated before the war begun'
b. Emigrou antes que a guerra começasse$_{\text{SUBJ}}$.
'(S)he emigrated before the war would begin'

(17) a. Fica aqui até alguém te chamar$_{\text{INF}}$.
b. Fica aqui até que alguém te chame$_{\text{SUBJ}}$.
'Stay here until someone calls you'

In (16a) and (17a), the embedded proposition is presupposed, its truth surviving if the main clause is negated, contrary to what is verified in (16b) and (17b). At first sight, in the latter cases, the embedded sentence is either taken to be false or else as describing a possibility. However, other examples, as (18), show that, even with the Subjunctive, the embedded clause may be true:

(18) a. Sai antes que morras$_{\text{SUBJ}}$!
'Get out before you die!'
b. Vou ficar aqui até que morra$_{\text{SUBJ}}$.
'I will stay here until I die'

Thus, the difference between infinitival and finite clauses in *before* and *until*-clauses is not primarily related to the truth value of the embedded clause. Moreover, both the infinitival and the finite clauses

express temporal precedence between the situation described by the main clause and the one described by the embedded clause. However, the infinitival clause can only be felicitously asserted in a context where it is part of the common ground, whereas the Subjunctive clause may not belong to the common ground – as in (16b) and (17b) – or else it introduces in discourse a new topic – as in (18).

5.2 *Without*-clauses

Clauses introduced by *sem* ('without') are another case where the choice exists between an infinitival and a Subjunctive clause:

(19) a. Ganhou o jogo sem se esforçar$_{_INF}$ muito.
b. Ganhou o jogo sem que se tenha$_{_SUBJ}$ esforçado muito.
'(S)he won the game without a great effort'

Both sentences indicate that the embedded proposition is false, but (19b) conveys the information that such falsity was unexpected, contrary to (19a), which does not convey unexpectedness (see also examples A18 of the Appendix). If the unexpectedness of $\neg q$ follows from p plus world knowledge, the use of the infinitive in *p without q* is much more natural with than without an intonation indicating surprise. By contrast, the use of a subjunctive clause dismisses the use of a particular intonation:

(20) a. Caminhou em cima de brasas sem SE QUEIMAR$_{_INF}$!
b. Caminhou em cima de brasas sem que se tenha$_{_SUBJ}$ queimado.
'(S)he walked over embers without getting burned'

This shows that the infinitival proposition is adequate to retrieve a proposition that belongs to the common ground (or that is expected given the information belonging to the common ground), while the subjunctive clause forces the consideration of possibilities outside the common ground. More precisely, in sentences of the form *p without q*, the subjunctive may occur in q if $\neg q$ is unexpected, while the infinitive may occur if the normalcy of $\neg q$ is assumed. If $\neg q$ is unexpected and infinitive is used, resort to a suppletive device, as intonation, will be needed.

5.3 *Because*-clauses

If, as stated above, causality involves counterfactual reasoning and, therefore, leads to the use of Subjunctive, one could expect Subjunctive to be the mood occurring in causal clauses. However, the Subjunctive might only exceptionally occur in some (affirmative[2]) causal clauses, as (21), and even in these cases it is not obligatory, the indicative being also acceptable, if not preferred:

(21) "No dia 4 de Outubro, como **estivesse**$_{\text{-SUBJ}}$ bastante pior, voltei à Urgência do Hospital de São José, onde uma médica me diagnosticou «conjuntivite bilateral purulenta»".
(CETEMPÚBLICO, par=ext471198-nd-94b-1)

'On the 4th of October, as I was much worse, I returned to the Emergency Department of the Hospital de São José, where a doctor diagnosed me with «bilateral purulent conjunctivitis»'

The explanation I propose for Indicative to be used in causal sentences, while Subjunctive is obligatory in complement clauses of causal predicates, as in *a chuva fez com que a prova {fosse$_{_SUBJ}$ / *foi$_{_IND}$} adiada* ('the rain caused the race to be postponed') is that sentences of the form *p because q* do not express a causal relation between p and q in the same way as causal predicates. Sentences as *p caused q* mean that if p had not occurred, all the rest being the same, q would not have occurred either. As for sentences of the form *q because p*, they indicate that, among the necessary conditions for q, the speaker highlights p as being the most relevant one. A nice example that sustains this claim is the answer that Edmund Hillary, the first man to climb Mount Everest, will have given when he was asked why he climbed the mountain: "because it was there". Obviously, the mountain being where it is does not cause anyone to climb it. It is, however, a necessary condition for the climbing event, in addition to other necessary conditions, such as the

2 Under the scope of negation, as in other sentences where the causal clause is not presented as true, Subjunctive might occur: *não saiu porque estivesse$_{_SUBJ}$ incomodado, mas por outra razão* ('he did not leave because he was upset, but for another reason') / *ou porque estivesse$_{_SUBJ}$ doente ou porque houvesse greve de transportes, o certo é que faltou à aula* ('either because he was sick or because there was a transport strike, the truth is that he missed the class'). Infinitive is also possible in these (negative) constructions. The Indicative might also occur, but only if the negative operator is an instance of metalinguistic negation.

willing to climb the mountain, the ability to do it, and so on.

Given this, let us consider infinitival and finite causal constructions:

(22) a. Ela chegou atrasada porque se perdeu$_{_IND}$.
b. Ela chegou atrasada por se ter$_{_INF}$ perdido.
'She arrived late because she got lost'

In the same way as observed in clauses introduced by *before*, *until* or *without*, the utterance of the infinitival sentence is adequate in a context where such proposition is part of the common ground, while the finite clause may introduce new information in discourse. In other words, the infinitival clause is useful to retrieve a proposition that is already known by the addressees, while, with the Indicative, the assertion of the causal sentence consists in the same process as the assertion of an independent declarative clause: by uttering it, the speaker expresses his belief that the proposition is true and presents it as a piece of information to be added to the common ground, if it is not yet part of the common ground. Evidence that a finite causal sentence may update the context of assertion, while an infinitival clause can only point to a proposition whose acceptance is shared by the participants in the conversation, can be found in the following dialogues:

(23) A: Ficou em casa porque estava$_{_IND}$ a chover.
'(S)he stayed home because it was raining'
B: Não! {Não estava a chover! / Ficou em casa porque estava de quarentena!}
'No! {It was not raining! / She stayed home because she was in quarantine!}'

(24) A: Ficou em casa por estar$_{_INF}$ a chover.
'(S)he stayed home because it was raining'
B: Não! {#Não estava a chover! / Ficou em casa porque estava de quarentena!}

5.4 Complement clauses

Complement clauses of some verbs, as the equivalents of *say*, *think* or *believe*, are another case that suggests that the choice between infinitival and finite clauses is pragmatically triggered. Basing on an example of Mandy Simons (see, e.g., Simons 2007, 2019, a.o.), the observation arises that the choice of an infinitival or a finite complement has different effects on the discourse:

(25) A: How will the weather be there?
B: A Ana {disse / pensa} que está$_{_IND}$ a chover. / Duvido que esteja$_{_IND}$ a chover.
'Ana {says / thinks / believes} that it is raining / I doubt that it is raining'

(26) A: How will the weather be there?
B: A Ana {disse / pensa} estar$_{_INF}$ a chover.
'Ana {says / thinks / +-believes} it to be raining'

While B's answer in (26) describes only Ana's opinion, in (25) it also allows the complement proposition to be interpreted as an answer to A's question. Thus, also in this kind of sentences, data suggests that an Indicative proposition may add new information to the context of conversation, contributing to update of the common ground, contrary to infinitival clause, whose assertion has no effect on the information shared by the participants.

6 Conclusion and notes towards formalization

The observed data allows the following conclusions:

• Subjunctive instructs the hearer to consider non-*p* worlds or worlds where an expectation following from *p* does not hold.

• Indicative instructs the hearer to consider only *p*-worlds and where the expectations following from *p* hold.

• Infinitive instructs the hearer to retrieve a proposition that is part of the common ground or is expected, its assertion not providing any change in the context of assertion.

• Finite moods in complement clauses of some verbs allow the complement proposition to be added to the context of conversation, contrary to the Infinitive.

Seeking to capture formally the above observations, let us consider some basic notions used in modal semantics and in dynamic semantics (see, e.g., Portner 2009 or Fintel & Gilles 2007):

M – Model of evaluation (the model representing the state of information against which the proposition is evaluated)

Cg – Common Ground (the set of propositions that participants in the conversation agree to accept as true)

C – Context Set (the set of propositions compatible with the Common Ground)

Since a proposition denotes a set of possible worlds (the worlds where the proposition is true),

the Context Set is a set of possible worlds. I assume that this set is ordered; i.e., some worlds of C are closer to what is expected than others. For instance, the possibility that a huge meteorite will hit the Earth in a near future, even if compatible with Cg, is less likely than, e.g., that elections for the Italian Parliament will be anticipated. Thus, possible worlds where a huge meteorite will hit the Earth in a near future are more distant than worlds where there will be anticipated elections for the Italian Parliament, even if all these worlds are part of the Context Set (C). Being ordered, C will contain a sub-set of Best worlds, those which are closer to what is expected given what is assumed:

Bc – The subset of C that is closer to Cg (i.e., Bc contains worlds where the expectations following from what is assumed are met)

Each proposition is evaluated against a Model. In the case of non-subordinated propositions, the model against which they are interpreted is C, the set of possible worlds that models the context of assertion. The assertion of a simple proposition *p*, as *it is raining*, is made against a context of assertion (or an information state) C, and, if *p* is accepted by the participants in the conversation, the assertion of *p* results in a new context, which is the subset of C that contains all but the non-*p*-worlds:

$$c + p = \text{c*}\ (\text{c*} = [\text{c}/\neg\text{p}] = c \cap p)$$

Hence, the meaning of a sentence corresponds to its Context Change Potential (CCP). I assume that also adverbial clauses are evaluated against C, as well as complement clauses of non-attitudinal predicates, as, e.g., *prevent* (as in *the hurricane prevented the plane from landing*) or *lead to* (as in *bad weather led Maria to give up the trip*). Complement propositions of attitudinal predicates are evaluated against the model introduced by the attitude predicate.

Given this, I propose that adverbial clauses have the following CCP (Figure 1 schematizes the information):

$c + p\_{INF} = c^* \mid ((\text{Bc}^* \cap p) \neq \varnothing) \wedge (\text{Bc} \cap p) \neq \varnothing))$

(*p* is already part of the Common Ground or an expectation that follows from what is assumed in the context of assertion)

$c + p\_{SUBJ} = c^* \mid ((\text{Bc}^* \cap p) \neq \varnothing) \wedge ((\text{Bc} \cap p) = \varnothing)$

(the assertion of a Subjunctive proposition *p* in a context *c* leads to consider worlds outside Bc; i.e., *p* refers an unexpected possibility)

$c + p\_{IND} = (\text{c} \cap p)$

(the assertion of a proposition *p* in the Indicative removes non-*p* worlds from the context set; no restriction is given concerning whether *p* is part of Cg, Bc, or whether it is outside Bc; i.e., *p* may be known, expected or new information in discourse).

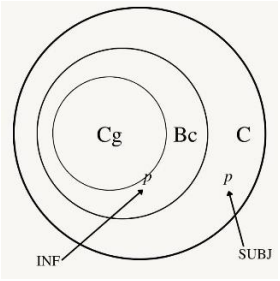


Figure 1 – Infinitival and Subjunctive adverbial propositions

The idea is that adverbial infinitival propositions are part of what is known or expected in discourse, while adverbial Subjunctive propositions, on the contrary, have the presupposition that the state of affairs described by the adverbial proposition is unexpected in discourse.

Concerning complement propositions of attitudinal predicates, they are evaluated against a model M_i that represents the epistemic state of the attitude holder *i*. Identically to the model representing the context of assertion, there is a set of propositions that the attitude holder takes for granted (what (s)he knows) and a superset, which is a (ordered) set of propositions compatible with what (s)he knows or takes for granted. Hence, the model against which complement clauses are evaluated is identical to the one represented in Figure 1. As seen above, the Indicative signals that M_i contains only *p*-worlds, the Subjunctive leading to the considerations of non-*p* worlds. Subjunctive rulers as, e.g., *duvidar* ('doubt') indicate that the epistemic state of the attitude holder contains non-*p* worlds, and the interpretation of factive-emotive predicates, as *lamentar* ('regret'), involves counterfactual reasoning, leading to the consideration of non-*p* worlds. Hence, in both cases the Subjunctive forces to search for worlds outside the center of the model of evaluation (i.e., Subjunctive indicates to search for possible worlds outside what is known or assumed).

As for infinitival complement clauses, it was observed that only finite clauses may be integrated in the context of assertion of the main clause (see (9) and (25)-(26)). In other words, an infinitival clause

is simply evaluated against a model M*i*, while a finite complement clause is evaluated against a model M*i* and may also be evaluated against C, the context of assertion of the main clause. Hypothetically, the complementizer (which is obligatory in the case of finite complementation and absent in infinitival complementation) introduces the instruction to check ($C \cap p$), in the case of Indicative complement clauses, or ($C \cap \neg p$), in the case of Subjunctive complement clauses. That is, Indicative signals the consideration of only *p*-worlds, Subjunctive instructing to consider non-*p* worlds, and the complementizer would give the instruction to check the sustainability of the complement proposition in C.

In sum, the picture that emerges is that the assertion of an infinitival proposition does not make any change in the context of assertion, being merely an instruction to check the existence of *p*-worlds in the context of evaluation (which is M*i* in the case of complement clauses, C in the case of adverbial clauses); the assertion of a subjunctive clause gives the instruction to look outside the center of the model of evaluation, and the assertion of an indicative clause gives the instruction to consider only *p*-worlds. In addition, concerning non-infinitival propositions, if the model of evaluation is M_i, they may also be evaluated against C, not necessarily making any change in C.

Schematically, each of the considered moods would give the following instructions:

p-Infinitive: check that *p*-worlds are part of the model of evaluation

p-Subjunctive: search for non-*p* worlds (necessarily outside the center of M, which contains only *p*-worlds); If $(M \neq C) \rightarrow (\neg p \cap C) =?$

p-Indicative: remove non-*p* worlds from the model of evaluation; If $(M \neq C) \rightarrow (p \cap C) =?$

Acknowledgments

This work was supported by CLUL and FCT (ref. UIDB/00214/2020). This paper has been improved by three valuable anonymous reviews.

References:

von Fintel, K. & A. Gillies: 2007, An opinionated guide to epistemic modality, in T. Szaebo (ed.), *Oxford Studies in Epistemology*, Vol. 2. Oxford University Press, Oxford: 32-62.

Giannakidou, A. 1994. The semantic licencing of NPIs and the Modern Greek subjunctive. *Language and Cognition 4, yearbook of the Research Group for Theoretical and Experimental Linguistics.* Univ. Groningen.

Giannakidou, A. & A. Mari. 2016. Emotive predicates and the subjunctive: A flexible mood OT account based on (non)veridicality. In *Proceedings of Sinn und Bedeutung 20*: 288-305.

Giannakidou, A. & A. Mari. 2021. *Truth and Veridicality in Grammar and Thought: Mood, Modality, and Propositional Attitudes*. Chicago University Press, Chicago.

Godard, D. 2012. Indicative and subjunctive mood in complement clauses: from formal semantics to grammar writing. In C. Piñón (ed.), *Empirical Issues in Syntax and Semantics*, 9: 129–148.

Hooper, J. B. 1975. On assertive predicates. In P. Kimball (ed.); *Syntax and Semantics*, 4. Academic Press, New York: 91 - 124.

Lewis, D. 1973. Causation. *The Journal of Philosophy*, 70(17): 556–567.

Mari, A. 2016. Assertability conditions of epistemic (and fictional) attitudes and mood variation. *Proceedings of SALT* 26: 61–81.

Marques, R. 2022. Not Just Gradability – Explaining the Subjunctive in Factive Contexts. Communication presented at *Linguistic Symposium on Romance Languages 52*, Univ. Wisconsin-Madison.

Palmer, F. R. 1986. *Mood and Modality*. Cambridge University Press, Cambridge.

Portner, P. 1997. The Semantics of Mood, Complementation, and Conversational Force. *Natural Language Semantics*, 5: 167-212.

Portner, P. 2009. *Modality*. Oxford University Press, Oxford.

Salmon, W. C. 1998. *Causality and Explanation*. Oxford University Press, Oxford.

Simons, M. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6): 1034-1056.

Simons, M. 2019. The Status of Main Point Complement Clauses. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*.

Villalta, E. 2008. Mood and Gradability: An Investigation of the Subjunctive Mood in Spanish. *Linguistics and Philosophy*, 31(4): 467–522.

Appendix. Examples of the analyzed constructions retrieved from the electronic *corpus* CetemPúblico (https://www.linguateca.pt)

Subjunctive clauses that describe facts:

A-1 “Cheguei ao Rio numa noite de Verão, embora fosse[SUBJ] Inverno.” (*ext1144787-nd-93b-1*)

‘I arrived in Rio on a summer night, even though it was winter’

A-2 “(…) Jill Jolliffe não desistiu e, juntamente com a RTP, conseguiu que Dom Rotheroe concluísse[SUBJ] o projecto já iniciado.” (*ext12281-clt-96b-2*)

‘Jill Jolliffe did not give up and, together with RTP, managed to get Dom Rotheroe to complete the project already started’

A-3 “Em Lisboa, foi preciso que os estudantes fechassem[SUBJ] as portas a cadeado sob as luzes da comunicação social para que fossem[SUBJ] escutados.” (*ext858783-soc-95a-1*)

‘In Lisbon, students had to lock the doors under the lights of the media so that they could be heard.’

A-4 “A situação política na Argélia não impediu que cerca de cem mil pessoas tivessem[SUBJ] assistido, na capital, ao maior concerto realizado naquele país nos últimos cinco anos.” (*ext83903-clt-96b-1*)

‘The political situation in Algeria did not prevent around 100,000 people from attending, in the capital, the biggest concert held in that country in the last five years.’

A-5 “Só lamento que Souness tenha[SUBJ]\ dito que eu não tinha qualidade para jogar no Benfica.” (*ext41444-des-98a-1*)

‘My only regret is that Souness said that I didn't have the quality to play for Benfica.’

Believe-clauses:

A-6 “Eu acredito que ele não teve[IND] nada a ver com isso.” (*ext1222201-nd-96b-1*)

‘I believe he had nothing to do with it.’

A-7 “A regionalização está na Constituição e acredito que vá[SUBJ] para diante.” (*ext70224-opi-97a-1*)

‘Regionalization is in the Constitution and I believe it will be done.’

A-8 “Muitos americanos não acreditam que os europeus têm[IND] quatro ou cinco semanas de férias.” (*ext769223-eco-95a-2*)

‘Many Americans don't believe that Europeans have four or five weeks of vacation.’

A-9 “Por regra, as pessoas não acreditam que alguém se esgote[SUBJ] no cumprimento das suas obrigações.” (*ext1151109-nd-97b-2*)

‘As a rule, people do not believe that someone is exhausted in fulfilling their obligations.’

Before-clauses:

A-10 “O assaltante, que estava encapuzado, teve ainda tempo para a fechar no quarto de banho da loja antes de fugir[INF].” (*ext769965-soc-95b-1*)

‘The assailant, who was hooded, still had time to lock it in the store's bathroom before escaping.’

A-11 “Crêm alguns que Fujimori decidiu encabeçar o golpe antes que os jovens turcos do Exército o depusessem[SUBJ].” (*ext17092-pol-92a-2*)

‘Some believe that Fujimori decided to lead the coup before the young Turks in the army deposed him.’

A-12 “O jogo é ocupar posições antes que os norte-americanos, um dia, regressem[SUBJ].” (*ext77837-pol-95a-1*)

‘The trick is to take positions before the Americans one day return.’

Until-clauses:

A-13 “A vizinhança diz ter sido alertada para o que estava a acontecer por um automobilista que ia a passar e que resolveu apitar até alguém surgir[INF] à janela.” (*ext268037-soc-97b-1*)

‘Neighborhood says they were alerted to what was happening by a passing motorist who decided to whistle until someone came to the window.’

A-14 “São como máquinas de ferro que prosseguem o seu caminho até que alguém rebente[SUBJ] com elas.” (*ext1564991-clt-94b-1*)

‘They are like iron machines that keep on going until someone blows them up.’

Without-clauses:

A-15 “Não era possível entrar ou sair do quartel sem levar[INF] tiros.” (*ext24850-pol-95b-2*)

‘It was not possible to enter or leave the barracks without being shot.’

A-16 “Imagine chegar à Polónia, no princípio dos anos 80, e perder-se na cidade de Szczebreszynie -- sem falar[INF] uma palavra de polaco, nem ter nascido com o dom natural para pronunciar quatro consoantes de uma só vez.” (*ext961749-eco-92a-1*)

‘Imagine arriving in Poland in the early 1980s and getting lost in the city of Szczebreszynie -- not speaking a word of Polish, nor being born with the natural gift for pronouncing four consonants at once.’

A-17 “Uma vasta operação da GNR realizada na quarta-feira, envolvendo seis centenas de militares dos distritos de Lisboa, Setúbal, Leiria e Santarém fiscalizou 3874 condutores sem que qualquer deles acusasse[SUBJ] excesso de alcoolémia.” (*ext411048-soc-95b-1*)

‘A vast GNR operation carried out on Wednesday, involving six hundred military personnel from the districts of Lisbon, Setúbal, Leiria and Santarém, inspected 3,874 drivers without any of them accusing excessive alcohol consumption.’

A-18 “Subitamente, sem que nada o fizesse$_{[SUBJ]}$ prever, recorda Emmanuel Desplechin, de 16 anos, «o autocarro flectiu à esquerda, inclinou-se, acabou por desabar e prosseguiu, de rojo, por 150 metros».”
(*ext19180-soc-95b-2*)

‘Suddenly, without anything to predict it, recalls Emmanuel Desplechin, 16 years old, «the bus turned left, leaned, ended up collapsing and continued, dashing, for 150 meters».’

A-18’ ??Subitamente, sem nada o fazer$_{[INF]}$ prever, (…)

‘Suddenly, without anything to predict it, (…)’

Because-clauses:

A-19 “Será que pensou que por ter$_{[INF]}$ contratado um campeão tinha garantido vitórias atrás de vitórias?”
(*ext1327948-des-98a-2*)

‘Did he think that because he hired a champion he had secured victory after victory?’

A-20 “Não participou porque foi$_{[IND]}$ precisamente no dia da festa que nasceu Maria Antónia.”
(*ext19275-clt-95a-1*)

‘(S)he did not participate because it was precisely on the day of the party that Maria Antónia was born.’

Finite / Infinitival complemente clauses:

A-21 “Um número mais restrito disse que tinha$_{[IND]}$ lido o livro.” (*ext97206-soc-97b-1*)

‘A more restricted number said that they had read the book.’

A-22 “Carlucci disse ter$_{[INF]}$ sido sempre partidário do apoio às «forças democráticas».”
(*ext26706-soc-91b-1*)

‘Carlucci said that he had always been in favor of supporting "democratic forces".’

A-23 “Considera-se a si próprio como um homem modesto e duvida estar$_{[INF]}$ à altura de tão altos cargos, mas, teoricamente, Jiang Zemin, 67 anos, é a figura mais poderosa da China, que desde há década e meia não concentrava tantos títulos num único dirigente.”
(*ext223530-pol-93b-1*)

‘He considers himself a modest man and doubts he is up to such high positions, but theoretically, Jiang Zemin, 67, is the most powerful figure in China, which has not held so many titles in a decade and a half in a single leader.’

A-24 “Quanto ao prazo avançado pela Câmara de Lisboa, duvida que se cumpra$_{[SUBJ]}$.”
(*ext227550-soc-96a-1*)

‘As for the deadline set by the Lisbon City Council, he doubts that it will be met.’

A-25 “O ídolo acha que tem$_{[IND]}$ poderes milagrosos e pensa ser$_{[INF]}$ responsável pela cura de várias crianças que sofriam de cancro.” (*ext591706-soc-93b-1*)

‘The idol thinks he has miraculous powers and thinks he is responsible for curing several children who suffered from cancer.’

A-26 “Hasse Ferreira pensa que tudo ficará$_{[IND]}$ resolvido este mês, sendo assim possível cumprir o plano de actividades para 1991, que estabelece o arranque da reconstrução.” (*ext46502-nd-91a-2*)

‘Hasse Ferreira thinks that everything will be resolved this month, making it possible to fulfill the activity plan for 1991, which establishes the start of reconstruction.’

Poster Abstracts

Understanding Fillers May Facilitate Automatic Sarcasm Comprehension: A Structural Analysis of Twitter Data and a Participant Study

Fatemeh S. Tarighat
UKRI CDT in NLP
University of Edinburgh
f.samadzadeh-tarighat@sms.ed.ac.uk

Walid Magdy
School of Informatics
University of Edinburgh
wmagdy@inf.ed.ac.uk

Martin Corley
School of PPLS
University of Edinburgh
martin.corley@ed.ac.uk

Abstract

Sarcasm detection models are often built based on self-annotated tagged data. However, fillers (e.g., *um* and *hmm*), deliberate use of which may indicate sarcasm, do not get enough attention in these models. We analyze five fillers in different categories of untagged tweets. We also present participant ratings of sarcasm, offensive language, language formality, and basic emotions in tweets with and without *um* and *hmm*. Our evidence, albeit weak, points to the importance of linguistic features such as these fillers in determining sarcastic meaning.

1 Introduction

Transcribed spoken language and user-generated online text are two of the main sources of training data for language models. Traditionally, fillers have been dismissed as noise in transcription of spoken language. However, the importance of understanding fillers and disfluencies of natural language has been emphasized in human-computer interaction research (e.g, Bates et al., 1993; Oviatt, 1995; Wigdor et al., 2016) with focused studies on real-time dialogue systems (Passali et al., 2022), question answering systems (Gupta et al., 2021), and autonomous vehicles (Large et al., 2017) in recent years. Moreover, sarcasm detection models do not account for linguistic details of their training resources and mainly rely on user-generated tags and indicators of sarcasm to flag remarks as sarcastic or non-sarcastic (Oprea and Magdy, 2020). This is all the more important as written language online is adopting elements of spoken language, e.g., the deliberate inclusion of fillers such as *uh* and *um* (also known as filled pauses and discourse markers). Whether spoken or written, fillers can convey sarcasm, among other things (D'Arcey et al., 2019).

With this in mind, we investigated 5 fillers and their potential sarcastic meanings on Twitter. Taking into account the type of tweets, we hypothesized that [1] users have position preferences when using fillers online and fillers appear more in the middle if the tweet is a stand-alone one and not in response to another tweet. [2] In contextually self-sufficient tweets, fillers are often perceived to deliver sarcasm. [3] Contextually independent tweets with filler somewhere in the middle get rated as sarcastic more than structurally similar tweets with filler appearing at the beginning or at the end.

2 Data Collection and Processing

We studied over 1.4 million English tweets containing *um*, *uh*, *hmm*, *erm*, *er*, and *#sarcasm* collected through the Twitter Application Programming Interface[1] using `twitter_collector`[2] over the span of 23 days. We excluded *#sarcasm* data from the study because most tweets including this tag did not include the fillers under investigation.

Our investigation focused on tweets that were classified as stand-alone, which could contain mentions (*@username*) or media but were not quotes, replies, or retweets. We reviewed random samples of these tweets to look for context-independent content to be used in our participant study. To ensure context-independence of the language in the tweets, we divided our sample into two groups; SELF-CONTAINED: tweets that only include text and emojis and MEDIA-URL: tweets that contain a form of media (e.g., image, GIF, video) and/or URLs (Table 1). 10% of the tweets analyzed included mentions.

	um	*uh*	*hmm*	*erm*	*er*
MEDIA-URL	33439	40934	42373	3302	19274
SELF-CONTAINED	106104	136066	167242	8006	56098
	139543	177000	209615	11308	75372

Table 1: Stand-alone tweets including each type of filler, with media content or links, or fully self-contained.

[1] https://developer.twitter.com/en/docs/twitter-api

[2] https://github.com/yalhariri/twitter_collector

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

	um	*uh*	*hmm*	*erm*	*er*
Beginning	0.34	0.27	0.38	0.38	0.06
Middle	0.62	0.69	0.45	0.54	0.89
End	0.04	0.04	0.17	0.08	0.06

Table 2: Proportions of tweets in each position, by matched filler in automatically selected database (612,838 tweets).

Table 2 shows the proportions of tweets which matched the search criteria containing each of the fillers under investigation at the beginning, in the middle, and at the end. As can be seen, there is a tendency for fillers to occur in the middle of tweets.

3 Participant Study

We created a pool of 2300 SELF-CONTAINED tweets by randomly selecting 10 tweets per day. We applied several rounds of filtering, e.g., to remove false positives such as 'ER' for 'emergency room'. We manually selected 48 tweets, 24 containing *um* and 24 containing *hmm*. Two independent NLP researchers conducted context sufficiency checks for us. Each set of 24 tweets included 8 with the filler at the beginning, 8 with the filler in the middle (defined as any word except the first or last), and 8 with the filler at the end. To investigate the specific role played by the fillers, we controlled for content, by creating versions of each of the 48 tweets which were identical in every respect, but had the fillers removed. The resulting 96 tweets were counterbalanced into two lists of 48, each including equal numbers of examples of each filler at tweet beginning, middle, and end, as well as a matched number of tweets with fillers excised from the same positions.

An experiment was administered via Prolific[3]. Participants were asked to rate tweets for sarcasm (SARCASM), offensive language (OFFENSE), language formality (FORMALITY), and emotions associated with the tweets (Ekman's six basic emotions, not discussed further here) in 5-point Likert and slider question formats. We assumed that self-contained tweets containing fillers should get rated as more sarcastic, more offensive, and less formal compared to their without-filler counterparts. We also wanted to know whether any effect of presence/absence of fillers was moderated by their positions in the tweets.

[3] https://www.prolific.co/

um			*hmm*		
Question	Position	Mean	Question	Position	Mean
Sarcasm	Beginning	3.15	Sarcasm	Beginning	3.59
	Middle	3.50		Middle	3.51
	End	2.67		End	2.85
Offense	Beginning	2.63	Offense	Beginning	2.89
	Middle	2.63		Middle	2.92
	End	1.99		End	2.35
Formality	Beginning	2.12	Formality	Beginning	2.27
	Middle	2.05		Middle	2.37
	End	1.95		End	2.41

Table 3: Mean ratings (0–5) for tweets with *um* or *hmm* present/absent in three positions, for SARCASM, OFFENSE, and FORMALITY.

96 participants took part in the study. We found weak evidence supporting our claims (Table 3). [1] For both *um* and *hmm*, SARCASM scores are slightly higher when fillers are present. [2] For OFFENSE, fillers seem to contribute to offensive tone with the highest contrast in *um* beginning and *hmm* middle. Also, *um* middle and end are the only instances where offensive language scores are slightly lower when the filler is present. Thus, they are the only instances that seem to slightly take the sting away from remarks. [3] Tweets with both fillers in all positions get rated as less formal when fillers are present. [4] Surprise is the only emotion that gets rated more when the filler is present.

4 Discussion

The present study is limited in scope and shows only weak evidence in support of its hypotheses. However, the numerical indication that inclusion of fillers increases the perception of sarcasm suggests that a larger-scale study is warranted. As our next step, we will study MEDIA-URL tweets along with quotes and replies in our data set in a similar fashion. We can then investigate fillers in self-annotated sarcastic tweets to check whether tweets are perceived sarcastic regardless of the filler in them. A better understanding of linguistic features such as fillers would allow us to train language understanding, prediction, and detection models with more accuracy.

Acknowledgments

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Infor-

matics, and School of Philosophy, Psychology & Language Sciences. Ethical approval for data collection and participant study was granted by the University of Edinburgh, School of Informatics Ethics Committee according to the Informatics Research Ethics Process, RT number 2021/29989.

References

Madeleine Bates, Robert Bobrow, Pascale Fung, Robert Ingria, Francis Kubala, John Makhoul, Long Nguyen, Richard Schwartz, and David Stallard. 1993. The bbn/harc spoken language understanding system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 111–114. IEEE.

J Trevor D'Arcey, Shereen Oraby, and Jean E Fox Tree. 2019. Wait signals predict sarcasm in online debates. *Dialogue & Discourse*, 10(2):56–78.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-qa: A benchmark dataset for understanding disfluencies in question answering. *arXiv preprint arXiv:2106.04016*.

David R Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics*, 63:53–61.

Silviu Vlad Oprea and Walid Magdy. 2020. The effect of sociocultural variables on sarcasm communication online. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–22.

Sharon Oviatt. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1):19–36.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. Lard: Large-scale artificial disfluency generation. *arXiv preprint arXiv:2201.05041*.

Noel Wigdor, Joachim de Greeff, Rosemarijn Looije, and Mark A Neerincx. 2016. How to improve human-robot interaction with conversational fillers. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 219–224. IEEE.

ConceptNet infused DialoGPT for Underlying Commonsense Understanding and Reasoning in Dialogue Response Generation

Ye Liu[1], **Wolfgang Maier**[1], **Wolfgang Minker**[2] and **Stefan Ultes**[1]
[1]Mercedes-Benz AG, Sindelfingen, Germany
{ye.y.liu,wolfgang.mw.maier,stefan.ultes}@mercedes-benz.com
[2]Ulm University, Ulm, Germany
wolfgang.minker@uni-ulm.de

1 Introduction

Many pre-trained transformer-based (Vaswani et al., 2017) language models (LMs) have been widely applied in spoken dialogue systems (SDS) and shown promising performance. However, the probing experiments in Zhou et al. (2021b) demonstrated that pre-trained LMs (Zhang et al., 2020; Roller et al., 2021; Lewis et al., 2020) fail to capture commonsense (CS) knowledge hidden in dialogue utterances, even though they were already pre-trained with numerous datasets.

To improve the CS understanding and reasoning ability of a pre-trained model and to build a dialogue agent like shown in Figure 1, we have two main contributions in this work. We firstly inject external knowledge into a pre-trained conversational model to establish basic commonsense. Secondly, we leverage this integrated commonsense capability to improve open-domain dialogue response generation so that the dialogue agent is capable of understanding the CS knowledge hidden in dialogue history on top of inferring related other knowledge to further guide response generation.

2 Enabling commonsense capability

To enable the commonsense capability of the pre-trained conversational model DialoGPT, commonsense triplets of ConceptNet (Liu and Singh, 2004)–a large-scale knowledge graph–are infused through efficient Adapter tuning (Pfeiffer et al., 2021). By utilizing the AdapterHub (Pfeiffer et al., 2020), the Houlsby Adapter (Houlsby et al., 2019) is used, which includes two bottleneck adapters in each transformer layer: one after the multi-head attention sub-layer and another after the feed-forward sub-layer. To efficiently integrate this external knowledge into DialoGPT, only parameters of Adapter layers are updated and the parameters of DialoGPT are frozen during training.

To achieve this, we adapt the work from Per-

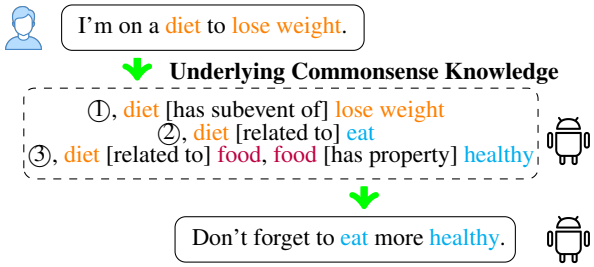


Figure 1: The ideal dialogue agent can understand the CS knowledge hidden in the dialogue history (①), meanwhile, infer the reasonable CS knowledge (② and ③) for further guiding an informative response generation. The key words/concepts are highlighted in orange for user utterance and blue for system response, respectively. The words highlighted in purple are middle concepts extracted in two-hop searching.

ozzi et al. (2014); Lauscher et al. (2020) and induce a synthetic corpus from ConceptNet through a bias random walk (Grover and Leskovec, 2016). We convert relations in ConceptNet into natural language phrases ("IsA" to "is a") and every relation is along with [] in collected data (Table 1) to distinguish relations from normal words/concepts. Finally, we created 359,421 data points and split them into 80%/10%/10% train/valid/test set. During CS_Adapter tuning, we add one special token "<|commonsense|>" to the DialoGPT tokenizer and insert it at the beginning of every prompt input (Table 1). Given the auto-regressive property of DialoGPT, four prompt templates (Table 1) are proposed and randomly chosen as input. Given the knowledge prompt, the ConceptNet integrated DialoGPT+CS_Adapter can generate series of CS triplets (like the data in Table 1).

3 CS-based Response Model

To enable commonsense-based open-domain response generation, we utilize the CommonsenseDialogues (Zhou et al., 2021a) dataset to continually train the DialoGPT+CS_Adapter presented in Section 2. This time, all parameters are updated.

data	autobraking [related to] automatic, automatic [derived from] auto, auto [related to] automobile, ...
prompt templates	<|commonsense|> autobraking [related to] <|commonsense|> autobraking [related to] automatic, <|commonsense|> autobraking [related to] automatic, automatic [derived from] <|commonsense|> autobraking [related to] automatic, automatic [derived from] auto,

Table 1: One data example in synthetic corpus from ConceptNet and four prompt templates randomly as input.

model	perplexity↓	concepts Acc (%)	assertion Acc (%)
DialoGPT baseline	1.405	-	-
DialoGPT+CS_Adapter (ConceptNet Integration in 2)	-	56.88	47.29
DialoGPT+CS_Adapter (Response Model in 3)	1.365	62.43	45.27

Table 2: The automatic metrics of DialoGPT baseline, DialoGPT+CS_Adapter knowledge integration (Section 2) and DialoGPT+CS_Adapter response model (Section 3).

	annotator 1	annotator 2
yes *vs* no	87 *vs* 13	88 *vs* 12
positive agreement	93.71%	

Table 3: The human assessment results on generated CS triplets that do not officially exist in ConceptNet.

As shown in Figure 1, the goal is a dialogue agent that is capable of *understanding* CS knowledge hidden in the dialogue history (like ① in Figure 1). Furthermore, the agent is also capable of *reasoning* other CS triplets for guiding the response generation. For this, we extract the knowledge triplets of keywords hidden in the dialogue history and the response (like ② and ③ in Figure 1). For CS knowledge extraction, we firstly extract the key words in dialogue history and response reference. We adapt the work from Tang et al. (2019) and Zhong et al. (2021) that use TF-IDF and Part-Of-Speech (POS) features to select the keywords. Secondly, we extract the CS knowledge of these keywords from ConceptNet, i.e., one-hop and two-hop triplets with the keywords as root. Considering two-hop results includes triplets where the source and target concepts have no direct connection but share a common middle concept (③ in Figure 1).

During DialoGPT+CS_Adapter training, maximal 3 turns' dialogue context is as input for memory-efficiency, the extracted CS triplets, where the "<|commonsense|>" is inserted, and response are as label. Meanwhile, we add two new tokens: "[USER]" and "[SYSTEM]" to distinguish the user utterance from system response. Afterwards, the DialoGPT+CS_Adapter can generate both underlying CS knowledge and reasonable response.

4 Results and Discussion

To evaluate the CS knowledge integration in DialoGPT we use two automatic metrics. One is *concepts accuracy*, which represents the proportion of generated (head concept, tail concept) pairs that exist in ConceptNet without considering if the generated relation is officially correct. Another is *assertion accuracy*, which represents the proportion of generated (head concept, relation, tail concept) triplets that officially exist in ConceptNet. In order to further test our assumption–even if the generated commonsense triplets do not officially exist in ConceptNet, they still make sense for humans–we manually evaluate the generated CS knowledge. For this, two Master students with computational linguistic background were hired. And the human evaluation results shown in Table 3 support our assumption. The result comparison in Table 2 demonstrate that our final DialoGPT+CS_Adapter response model has comparative performance on knowledge generation compared to the plain DialoGPT+CS_Adapter after ConceptNet integration and lower perplexity (Serban et al., 2015) compared to the DialoGPT baseline.

In this work, we found several shortcomings that need to be discussed in our future work. One is that the relation distribution in ConceptNet is severely imbalanced which results in an over-generation of the "[related to]" relation. Another shortcoming is that our method of extracting CS triplets hidden in the dialogue is rule-based. It includes keywords extraction and knowledge extraction without considering the discourse information. A next step will be the application of neural network methods for knowledge extraction.

References

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Anne Lauscher, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8):434–441.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14568–14576.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021a. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore and Online. Association for Computational Linguistics.

Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2021b. Probing commonsense explanation in dialogue response generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4132–4146.

Real-life Listening in the Lab: Does Wearing Hearing Aids Affect the Dynamics of a Group Conversation?

Eline Borch Petersen, Els Walravens, Anja Kofoed Pedersen
WS Audiology, Scientific Audiology, Lynge, Denmark
eline.petersen@wsa.com; els.walravens@wsa.com; anja.pedersen@wsa.com

Abstract

Hearing science traditionally focuses on testing listening in isolation. Here we explore the effect of providing hearing aids to listeners with hearing impairment by analyzing the dynamics of a group conversation. First, a pilot study was conducted to identify suitable conversational starters. Using these starters, preliminary results from an experiment involving two normal-hearing and one hearing-impaired interlocutor are presented. The results show that when providing hearing-aid amplification to the hearing-impaired talker in close-to-quiet situations (noise at 50 dB SPL) and applying directional signal processing when conversing in noise (75 dB SPL), all talkers reduce their speech level. This effect could stem from the normal-hearing interlocutors no longer having to compensate for the communication difficulty experienced by the hearing-impaired interlocutor.

1 Introduction

When evaluating the detrimental effects of being hearing impaired (HI), hearing science has traditionally one-sidedly focused on the ability to listen. Some of these detriments can be, partly, compensated for by presenting amplified and processed sounds through hearing aids (HAs).

Recently, the hearing science community has requested more emphasis on '*encompassing the interactive nature of everyday communication*' (Keidser et al., 2020) into experimental designs. So far, a few studies have focused on exploring the communication between a HI and a normal-hearing (NH) interlocutor, showing the NH alters the spectral content of his/her speech (Beechey et al., 2020b; Hazan et al., 2019) and increases speech levels (Sørensen et al., 2021) when having a conversation with a HI interlocutor. Providing HI interlocutors with HA amplification caused the HI to initiate turn-taking faster (FTO floor-transfer offset), increase the articulation rate, and reduce the speech level. In response, the NH interlocutors also reduce the speech level when their HI conversational partner was wearing a HA as opposed to unaided listening (Beechey et al., 2020a; Petersen et al., 2022).

We are currently exploring whether similar effects of HA signal processing can be seen in a group conversation between a HI person and two NH persons. This includes identifying a suitable conversational task when increasing the group size from two persons to three.

2 Identifying Suitable Conversational Tasks

Studies within hearing science often evaluate within-subject changes e.g., of providing HA amplification or speech-enhancing HA processing strategies. As such, the study design must meet these demands: **1)** conversations must be replicable and natural. **2)** No learning/ training effect of the conversation task to avoid alterations in the conversational dynamics over time. **3)** The above should be realized for previously unacquainted interlocutors. Additionally, **4)** the task should not require visual acuity or physical activity.

As none of the existing methods for starting a conversation met the above criteria, we conducted a pilot study exploring three conversational starters prior to the actual experiment: **A) Consensus** questions where participants were to agree upon a common answer. **B) Picture cards** with keywords encouraging a conversation based on a theme. **C)** Four historical events to be put in chronological order in a **timeline**.

The goal of the study was to investigate whether the starter affected the conversational dynamics. A total of 10 examples of each starter were generated and tested in four groups of three NH interlocutors.

The timeline task showed significantly altered dynamics compared to the consensus and picture-card tasks: **i)** The turn-taking timing (median FTO) was significantly higher (33 ms longer, $p < 0.001$), **ii)** there were fewer floor-transfers (2.6 turns/min less, $p < 0.001$), **iii)** the speaking times between talkers were less balanced (difference between talkers was 9.75% higher, $p = 0.03$), and **iv)** more silence was present within the 5-minute conversations (24.9 s of additional silence, $p < 0.001$). None of these measures showed any training/learning effects over time.

These results indicate that the timeline task was less interactive. One participant noted to another "*you are thinking inside your head, you have to say it out loud*". And difference in speaking times was likely due to a difference in background knowledge between participants.

The timeline task was discarded, and for the experiment investigating the effects of HA processing, the picture cards and consensus questions were used to start the conversations.

3 Effect of Hearing Aid Processing on Group Conversational Dynamics

Using the two tasks described above to initiate conversations between a HI and two NH persons (one <35 years and one >50 years), we investigated how providing HA processing to the HI talker affected the group conversation. The effect of HA amplification was investigated in low 50 dB SPL background noise (unaided vs aided), while the effect of providing directional microphone sound-processing[1] was examined in high 75 dB SPL background noise (omni-directional vs directional).

Preliminary analysis of the first 10 triads shows that providing HA amplification to the HI interlocutor reduced the speech levels of all three talkers in low background noise (unaided vs aided: -1.2 dB, $p = 0.002$, **Figure 1A**). Similarly, in high levels of noise, improving the listening situation of the the HI interlocutor by providing directional signal-processing caused all three talkers to reduce their speech levels (omni vs dir: -0.7 dB, $p = 0.04$, **Figure 1A**). As might be expected, this change was most evident for the HI talker (omni vs dir: -2.0 dB, $p = 0.001$, **Figure 1B**).

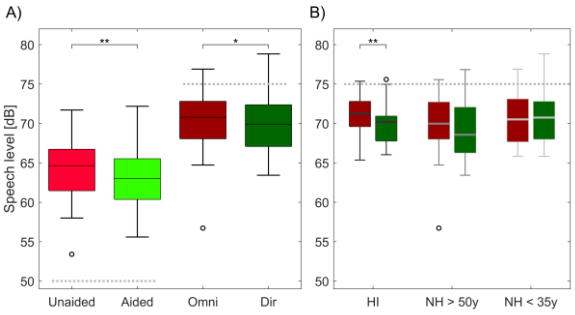


Figure 1: A) Effect of providing hearing aid amplification (unaided vs aided) at a low background noise level and of directional signal-processing (omni vs dir) in a high level of noise. The noise levels are indicated by dotted grey lines (50 dB and 75 dB respectively). **B)** The interaction between talker and effect of directional processing (omni vs dir) showing how the HI talkers are most affected by the alteration in hearing aid signal processing.

While conversing at a very positive SNR in the low level of noise (13.6 dB), the conversation in noise is conducted at low SNR (-4.9 dB). In a standardized speech-in-noise test using every-day sentences, an SNR of -2.5 dB corresponds to an intelligibility of around 50% for younger NH listeners (Nielsen & Dau, 2009). As all interlocutors in the current experiment were able to conduct a conversation at -4.9 dB SNR, the speech intelligibility seems to be much higher for real-life communication, than in the standardized laboratory tests of speech understanding in noise. This illustrates how the traditional single-sided focus on listening result in test scenarios which far from resemble every-day listening.

This is the first known attempt to investigate the effect of HI and HA signal processing on the dynamics of a group conversation. Preliminary results show that, despite only affecting the listening condition of the HI talker, HA processing causes all talkers to adjust their speech levels. Although the effect in noise is largest for the HI talker, it also affects the two NH talkers, potentially due to the NH talkers no longer having to make up for the communication difficulty experienced by the HI listener when providing adequate HA processing, improving audibility.

[1] Omni-directional processing preserve the auditory input, while directional processing combines the HA microphone inputs to emphasize sounds from the front, while attenuating noise from the back.

References

Beechey, T., Buchholz, J. M., & Keidser, G. (2020a).

Hearing aid amplification reduces communication effort of people with hearing impairment and their conversation partners. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1299–1311. https://doi.org/10.1044/2020_JSLHR-19-00350

Beechey, T., Buchholz, J. M., & Keidser, G. (2020b). Hearing impairment increases communication effort during conversations in noise. *Journal of Speech, Language, and Hearing Research*, *63*(1), 305–320. https://doi.org/10.1044/2019_JSLHR-19-00201

Hazan, V., Tuomainen, O., Kim, J., & Davis, C. (2019). The effect of visual cues on speech characteristics of older and younger adults in an interactive task. *Proceedings of ICPhS 19, Melbourne, Australia*, 815–819.

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., & Smeds, K. (2020). The Quest for Ecological Validity in Hearing Science : What It Is , Why It Matters , and How to Advance It. *Ear and Hearing*, *41*(Suppl 1), 5S-19S. https://doi.org/10.1097/AUD.0000000000000944

Nielsen, J. B., & Dau, T. (2009). Development of a Danish speech intelligibility test. *International Journal of Audiology*, *48*(10), 729–741. https://doi.org/10.1080/14992020903019312

Petersen, E. B., Macdonald, E. N., & Sørensen, A. (2022). The Effects of Hearing Aid Amplification and Noise on Conversational Dynamics between Normal-Hearing and Hearing-Impaired Talkers. *Trends in Hearing*, *26*, 1–18. https://doi.org/10.1177/23312165221103340

Sørensen, A., Fereczkowski, M., & MacDonald, E. (2021). The effects of noise and second language on conversational dynamics in task dialogue. *Trends in Hearing*, *25*, 1–17. https://doi.org/https://doi.org/10.1177/23312165211024482

"He hasn't done much to keep it up": Annotating topoi in the balloon task

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

Abstract

We present a preliminary study of the topoi employed by people with a diagnosis of schizophrenia in triadic dialogues discussing a moral dilemma with people who are unaware of their diagnosis. Results support the hypothesis that people with a diagnosis of schizophrenia are more consistent in their reasoning than healthy controls.

1 Introduction

Interacting with others frequently involves making common-sense inferences linking context, background knowledge and beliefs to utterances in the dialogue. However, sometimes it is not obvious how a particular contribution should be interpreted in terms of the underpinning assumptions warranting an inference. In dialogue involving participants who demonstrate atypical linguistic behaviour, such as people with a diagnosis of schizophrenia, the effects may be even more marked. In this exploratory study we consider the *topoi* – underpinning warrants – evoked in triadic dialogue involving people with a diagnosis of schizophrenia, focusing on the variety of topoi drawn on by patients and controls respectively.

2 Background

In addition to the traditional inter- and intrasentential structures normally assumed in linguistic theory such as questions, dialogue requires us to deal with phenomena such as clarifications, repair, overlap and split utterances. These can all be linked to reasoning in dialogue (Jackson and Jacobs, 1980; Breitholtz and Cooper, 2011; Breitholtz, 2014; Breitholtz and Howes, 2015). Reasoning in dialogue is *enthymematic*, that is, the arguments presented lack some premises which would be required in a fully logical chain of reasoning. Instead, enthymematic arguments (*enthymemes*) rely on notions or warrants in the minds of the listeners. These are often referred to as *topoi* (Aristotle, ca. 340 B.C.E./2007; Ducrot, 1988; Anscombre, 1995). When we interact we expect certain topoi to be common ground, or to be accommodated (adopted by dialogue participants) during the course of the interaction. If conversational participants access different topoi to serve as underpinnings for a particular argument, this may lead to misunderstandings and other disruptions in the dialogue. In this exploratory study we look at the topoi used in dialogues where participants are asked to make a decision regarding a moral dilemma.

3 Experiment

Building on work presented in (Howes et al., 2021), we take a more detailed look at the specific topoi provided in dialogues with a person with a diagnosis of schizophrenia, compared to control dialogues.

3.1 Data and analysis

The data (described elsewhere, e.g. Lavelle et al., 2013; Howes et al., 2021) consists of 38 triadic dialogues where participants discuss a moral dilemma and reach agreement about which of four people in a hot air balloon should jump to save the other three. Half of the dialogues include a person diagnosed with schizophrenia, with their two interlocutors unaware of their diagnosis, while the other half are between three healthy controls.

3.1.1 Annotation

As a point of departure we used the data from (Howes et al., 2021) and extended the annotations of turns which provided a reason to specific topoi. The authors developed a topos coding schema based on four sample dialogues, two involving people with a diagnosis of schizophrenia and two control dialogues, which was then given to two annotators to apply to the whole dataset. For each reason coded in the data, the annotators were asked

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

to choose one of several topoi from a drop down menu (see (1) and (2), below). The topoi differed depending on which of the balloon passengers the reason was related to, however, for this preliminary study we looked only at the reasons given for or against saving the balloon pilot, who is described in the instructions as the only one with any balloon flying experience.

The list of topoi given to the annotators included the following possible topoi:

(1) For saving the pilot
  (a) If the pilot is the only one who can fly the balloon you need to keep the pilot;
  (b) If the husband dies the wife might get upset;
  (c) If the pilot is thrown out, the child will be fatherless;

(2) For not saving the pilot
  (a) If one of the passengers is sacrificing themselves they might as well be thrown out;
  (b) If someone other than the pilot could fly the balloon, the pilot is expendable;
  (c) If you are going to leave a legacy then you can be sacrificed;
  (d) If you are married to someone you know some of what they know;
  (e) If piloting is not hard, then anyone can do it;
  (f) If the balloon will crash anyway, the pilot might as well be thrown;

If the annotators found that none of the given topoi was suitable they were instructed to add their own. Additional topoi supplied by the annotators during the annotation task were:

(3) For saving the pilot (all Annotator 1)
  (a) If someone is going to be difficult to throw over they should not be sacrificed
  (b) If a pair of people may reproduce they should be saved / If someone has a family they should be saved
  (c) The pilot of an aircraft is not necessarily responsible for accidents
  (d) If someone is a little person they don't weigh a lot

(4) For not saving the pilot
  (a) If someone is responsible for the situation they should be sacrificed (Ann1) / If someone is responsible for the crash they should jump (Ann2)
  (b) A reason to save someone has to be unique to the person (Ann1) / If there is nothing special about you, you can be sacrificed (Ann1) / If the others are more valuable you can be sacrificed (Ann2)
  (c) If someone can be replaced in their romantic relationship then they should be sacrificed (Ann1)
  (d) If you die someone else can take care of your child (Ann2)
  (e) If someone is an adult man they are heavy so throwing them is more effective (Ann1) / If someone is very fat they should be sacrificed (Ann1) / If someone is an adult man they weigh a lot (Ann1) / If someone is heavy then they can be thrown out (Ann2)
  (f) If someone is an adult man they can be sacrificed (Ann1) / If you are an adult man you are more likely to survive a fall from a hot air balloon (Ann1)
  (g) The person whose idea it was to throw someone off should be the one who gets thrown (Ann1)
  (h) If someone has lived a long time they should not be saved (Ann1) / If someone has lived a longlife they can be sacrified (Ann2)

As can be clearly seen in (4a) and (4h), for example, several of the additional topoi were recognised by both annotators, despite not appearing on the list. Those that received 3 or more annotations or clearly matched were therefore included as categories in their own right in the inter-annotator agreement calculations, with the rest being allocated to an 'other' category. This resulted in 4 categories for saving the pilot with Cohen's kappa $\kappa = 0.792$, and 12 categories for not saving the pilot $\kappa = 0.659$. For the following results we use the annotations from Annotator 1.

4 Results and Discussion

199 of the 206 (97%) reasons given for saving the pilot were taken from the topoi shown in (1), with 146 (71%) of these being annotated as (1a). The reasons for not saving the pilot were more diverse, with 151 of 215 (70%) coming from the list provided in (2), (48 (2b); 22%) and 39 (2e); 18%) and a further 40 from the added topos in (4a) (19%).

All dialogues contained at least two reasons for or against saving the pilot, with a range from 2 to 10. However, only 6 (33%) of the patients provided more than one reason for saving or not saving the pilot compared to 25 of their partners (66%; $\chi^2_1 = 5.21, p = 0.02$) and 40 of the healthy controls (70%). Additionally, in the control dialogues, arguments are more likely to be taken up by more than one participant – 57 out of 112 topoi (51%) are associated with turns by more than one participant in the same dialogue, compared to 24 out of 64 in the patient dialogues (38%) though this does not reach significance ($\chi^2_1 = 2.94, p = 0.086$).

This suggests, in line with the qualitative results of (Howes et al., 2021) that people with a diagnosis of schizophrenia are more consistent in their reasoning and use less varied arguments than non-patients. One such example can be seen in (5) where the patient argues that the pilot messed up and therefore should be thrown based on the topos that if someone is responsible for a situation they are the one that should be sacrificed – a topos that the patient returns to much later in the dialogue.

(5) **lines 58-62** If he messed up that to that point.
lines 132-135 I just feel if he messed up to this point, I don't know what he's doing there.

Acknowledgements

This work is part of the *Dialogical Reasoning in Schizophrenia* project supported by Riksbankens Jubileumsfond (RJ: P16-0805:1) and would not have been possible without our colleagues Robin Cooper and Mary Lavelle. We would also like to thank our annotators, Bill Noble and Anna Lindahl. Howes is also supported by the Swedish Research Council (VR 2014-39 *Centre for Linguistic Theory and Studies in Probability*).

References

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

Aristotle. 2007. *On Rhetoric, a theory of civic discourse* (translated by George A. Kennedy). Oxford University Press, Oxford. (original work published ca. 340 B.C.E.).

Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, University of Gothenburg.

Ellen Breitholtz and Robin Cooper. 2011. Enthymemes as rhetorical resources. In *Proceedings of the 15th workshop on the semantics and pragmatics of dialogue (LosAngelogue)*, pages 149–157.

Ellen Breitholtz and Christine Howes. 2015. Within reason: Categorising enthymematic reasoning in the balloon task. In *Proceedings of the 19th workshop on the semantics and pragmatics of dialogue (goDIAL)*, pages 160–161.

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

Christine Howes, Ellen Breitholtz, Mary Lavelle, and Robin Cooper. 2021. Justifiable reasons for everyone: Dialogical reasoning in patients with schizophrenia. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*. SEMDIAL.

Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*, 66(3):251–265.

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

Chat-o-matic: an online chat tool for collecting conversations of situated dialogue

Simon Dobnik and **Aram Karimi**
Department of Philosophy, Linguistics and Theory of Science (FLoV)
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
name.surname@gu.se

1 Introduction

One of the major challenges for modelling situated dialogue is a collection of quality natural language data in situational contexts. Such data involves long and sometimes persistent dialogue conversations over situated scenes between several conversational partners. Scenes may also change over time either because new events occur or conversational participants change locations so dialogue history has to be matched with different states in the perceptual environment. Collection of such data requires significant time as conversational participants must interact with each and with the surrounding environment. The amount of data that can be practically collected from a single participant is thus relatively limited which together with a high variability of both linguistic and perceptual information makes situated dialogue an *under-resourced task* in natural language processing.

One possibility taken in constructing corpora is to automatically generate the linguistic dataset from a set of rules (Das et al., 2018) but this has distinct limitations as the scope of language and vision that these rules cover is highly limited and the referring accuracy of these descriptions is unreliable (Aruqi, 2021). Several task-tailored solutions have been developed to collect multi-modal data with targeted participants either in real, e.g. (Dobnik, 2009), or virtual environments , e.g.(Stoia et al., 2008). There are also general solutions, most notably Dialogue Experimental Toolkit or DiET (Healey et al., 2003). This can be run on two or more participant computers where the clients are connected to a server where dialogue tasks are configured and recorded. The user interface resembles online messaging applications, with a chat window in which participants view the unfolding dialogue and a typing window in which participants can type and correct their contributions. In order to extend the pool of participants, over the last 10 years researchers have turned to online crowd-sourcing such as Amazon Mechanical Turk which has shown to produce good quality data based on completion rate and passing manipulation checks (Ipeirotis, 2011).

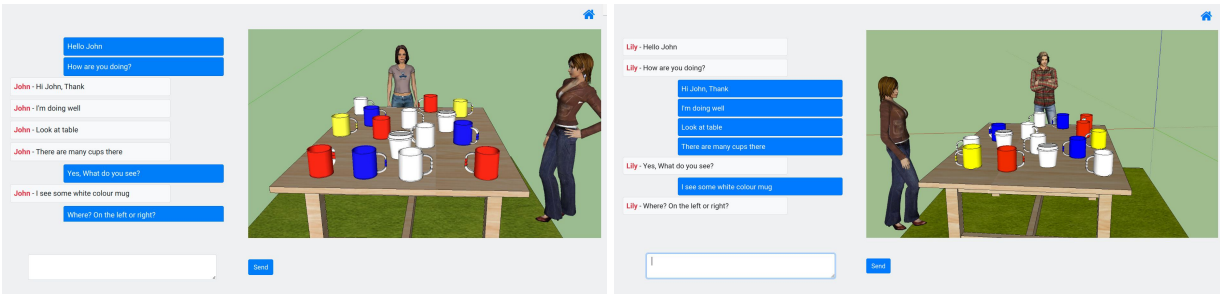
The crowd-sourcing platforms have been developed with the aim of collecting information from a single person at a time which is presented as a list of choices and input text. Since crowd-workers are normally paid by each input, time is an important issue for them, and they will try to accomplish the task as quickly as possible. Crowd-sourcing has been successfully used in research in computational linguistics, robotics and computer vision for tasks such as describing, annotation and providing speaker judgements. In order to set-up a task and the resulting data it is important to understand other tasks on the platform as well as the demographics and experiences of workers (Hitlin, 2016). Situated dialogue represents several additional changes to the standard crowd-sourcing setup. The first one is matching participants in real time and engage them in longer persistent conversations. The second is connecting visual environments with the textual chat interface. Several tools have been developed already that offer this kind of functionality, for example (Chernova et al., 2011; Manuvinakurike and DeVault, 2015; Das et al., 2017; Schlangen et al., 2018).

2 Semant-o-matic chat

We present a demo of an online chat tool that we have developed to collect conversational data from situated dialogue called Semant-o-matic chat or *Chat-o-matic*[1]. This allows us to setup data collection experiments both with targeted participants and with online crowd-sourcing platforms. The main motivation for developing our own tool rather than using existing solutions was to extend our

[1] https://www.dobnik.net/experiments/chat-o-matic

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.



(a) The View of Person 1

(b) The View of Person 2

Figure 1: View of the chat interface for different participants. Each participant is presented with a different view of the situated scene.

existing *Semant-o-matic* tool which we used for crowd-sourced collection of both linguistic (Rajestari et al., 2021) and situated data (Dobnik and Åstbom, 2017). The additional functionality includes the ability for participants to find each other online, engage in dialogue and presenting them with different scene views as shown in Figure 1.

2.1 Database

To make the tool more efficient and easier for researchers to analyse the data, we use a SQL database for storing all data. The table format is easy to understand and provides an organised and structural way to represent information (Davies, 2005). Additional features can be easily added in the future and the collected data can be searched and modified using SQL queries and filters which allow reference to data across columns and tables. In addition to conversations non-linguistic information such as answers from participants about their background, time during which the conversation took place and timings between turns can also be recorded.

2.2 Connecting with Mechanical Turk

The chat tool can be used openly by participants who can sign up for an account and then initiate a conversation with one of other users who are currently online. Participants can be attracted through target advertising of the task. Once initiated, the conversation is persistent and the participants can continue at any time later while keeping their participant roles (see Figure 1. We are currently testing the tool with AMT where we are using the same method as with the existing *Semant-o-matic* tool described in (Rajestari et al., 2021). A limited number of hits is published frequently to ensure that workers find our task. When signing up for the task, participants are issued with a unique id which becomes their username on *Chat-o-matic*. Participants are paid as the tasks are published per each turn after these are checked for quality.

2.3 The task

While *Semant-o-matic chat* can be setup for any situated dialogue task, it is currently setup for the data collection of the Cups corpus described in (Dobnik and Silfversparre, 2021; Dobnik et al., 2020) which was previously collected with targeted participants using the DiET chat tool. The goal of this task is to study referring over longer stretches of situated dialogue where each participants has a different view of a slightly divergent scene: some cups are missing for each but these can be seen by the other. We use typing indicators to keep user attention and provide real-time feedback that replicates the feelings and cadence of an in-person conversation. This also helps users to better manage their turn-taking. Participants start with a blank view and the view of Person 1 is assigned to the one making the first turn.

3 Conclusion and future development

Our ongoing efforts are focused on testing the tool on AMT. We are also developing functionality that will help us to overview dialogues in a web interface (without the need to directly issue SQL queries) and automatically calculate the amount of payment for the reviewed dialogues. The code of the tool will be released as open source. We intend to extend the tool to new tasks of situated dialogue and adding functionality such as recording of key-strokes and corrections through JavaScript and usage of dynamic multi-modal data such as sound and video.

References

Ali Aruqi. 2021. Embodied question answering in robotic environment: Automatic generation of a synthetic question-answer data-set. Masters in language technology (mlt), 30 hec, Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik and Nikolai Ilinykh, examiner: Staffan Larsson, opponent: Catherine Viloria.

Sonia Chernova, Nick DePalma, and Cynthia Breazeal. 2011. Crowd-sourcing real-world human-robot dialogue and teamwork through online multiplayer games. *AI Magazine*, 32(4):100–111.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335.

Mark Davies. 2005. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, 10(3):307–334.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

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

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of PotsDial - Semdial 2021: The 25th Workshop on the Semantics and Pragmatics of Dialogue*, Proceedings (SemDial), pages 50–60, Potsdam, Germany.

Patrick G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg J. Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, volume 25, pages 539–544, Boston, MA.

Paul Hitlin. 2016. Research in the crowdsourcing age: A case study. Technical report, Pew Research Center, Pew Research Center.

Panos Ipeirotis. 2011. Crowdsourcing using Mechanical Turk: quality management and scalability. Talk, Stern School of Business, New York University.

Ramesh Manuvinakurike and David DeVault. 2015. Pair me up: A web framework for crowd-sourced spoken dialogue collection. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 189–201. Springer.

Maryam Rajestari, Simon Dobnik, Robin Cooper, and Aram Karimi. 2021. Very necessary: the meaning of non-gradable modal adjectives in discourse contexts. In *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020*, volume 184 of *NEALT Proceedings Series, No. XX*, pages 50–58, Linköping, Sweden. Northern European Association for Language Technology (NEALT), Linköping University Electronic Press: Linköping Electronic Conference Proceedings.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. slurk – a lightweight interaction server for dialogue experiments and data collection. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Aix-en-Provence, France. SEMDIAL.

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

Speaker transitions in 2- and 3-party conversation

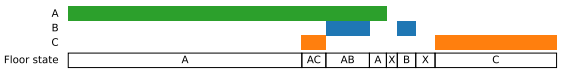
Emer Gilmartin
ADAPT Centre, Trinity College Dublin
Ireland
gilmare@tcd.ie

Marcin Włodarczak
Stockholm University
Sweden
wlodarczak@ling.su.se

1 Introduction

This paper reports ongoing work on temporal aspects of how participants manage conversation. We analyse dyadic phone conversations in the Switchboard (SWB) corpus (Godfrey et al., 1992) using a method previously employed on multiparty dialogue (Włodarczak and Gilmartin, 2021). The analysis is based on *floor state* - who is speaking or silent at any moment during interaction. By annotating *floor state intervals*, stretches of time during which a particular floor state holds, we can analyse *floor state transitions* or sequences of contiguous floor states. We are particularly interested in transitions between 'substantial' stretches of single party speech, to elucidate turntaking. We focus on transitions between stretches of single party speech in the clear of at least one second in duration (to avoid treating e.g. backchannels as turns). We distinguish *between speaker transitions* (BST) and *within speaker transitions* (WST). In WST, the speaker on either side of the transition is the same, as in turn retention, while in BSTs, the single party speech bounding the transition is by different speakers, as in turn change. To illustrate, Figure 1 shows a short exchange from a 3-party conversation. It involves 8 floor states – solo speech (**A**, **B**, **C**), overlaps (**AC**, **AB**) and general silence (**X**). Without the one second threshold we would treat this stretch as a series of three transitions: **A_AC_AB_A** from A to A, **A_X_B** from A to B, and **B_X_C** from B to C. However, looking at the transcript and the speech patterns, it seems more likely that the *longer* stretches of solo speech (**A**, **C**) delimit a single more complex transfer of floor possession from speaker A to C.

In previous work we found similarities in speaker transition distribution in different multiparty corpora. One-interval transitions were the largest class for all corpora studied, with a higher proportion of one-interval transitions in WST. However, less than half of between and within speaker transitions were accomplished with a single intervening interval of silence or overlap, indicating that turn change and retention are often a more complex sequence of events than a simple silence or short overlap. We found high levels of uniformity in the most common WSTs and BSTs found in different languages and settings (Gilmartin et al., 2020, 2019; Gilmartin, 2021; Włodarczak and Gilmartin, 2021; Gilmartin et al., 2018). We found considerable complexity and growing incidence of participation by more speakers with transition length, and that silent intervals account for a significant part of transition duration. Below, we analyse SWB to investigate whether our findings on multiparty talk hold for dyadic phone conversations.



A …cameras. You're just not normally as aware of it.
C Yeah
B It's true.
B Yeah.
C There's two on every bus.

Figure 1: An excerpt from a 3-party casual conversation corresponding to a between-speaker transition, **A_AC_AB_A_X_B_X_C**, from speaker A to C with six intervening intervals (**AC**, **AB**, **A**, **X**, **B**, **X**). *Top*: Temporal organization of individual speakers' contributions (represented as color bars) and the resulting floor states. *Bottom*: Simplified transcript. Speakers' contributions are color-coded for consistency.

2 Data and Annotation

We used the 2438 dyadic phone conversations (259 hours) in the Switchboard-1 Telephone Speech Corpus: Release 2, with the Mississippi University ISIP word level transcription. Transcripts were processed using Praat and Python to create speech and silence labels with all non-speech sounds suppressed to silence, resulting in 520135 talkspurts.

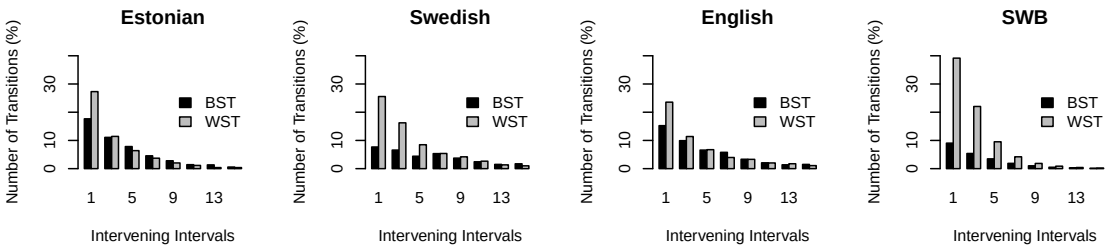


Figure 2: Distribution of Between and Within Speaker Transitions in Switchboard and 3 other corpora

From these we generated BST and WST labels, as described above. We used annotations from 3-party dialogue data from our previous studies to compare with the SWB results –three-party spontaneous conversations in Estonian (Lippus et al.) and Swedish (Włodarczak and Heldner, 2017), and collaborative conversational games in English (Litman et al., 2016). This data set contained 22106 talkspurts in 9 hours and 51 minutes hours of conversation.

3 Results

Results are first presented for SWB, and then contrasted with results on multiparty corpora.

Distribution of Speech, Silence and Overlap SWB has lower incidence of silence and overlap than the multiparty datasets, and higher incidence of single-party speech in the clear.

Distribution of speaker transitions SWB yielded 256,655 speaker transitions in 259 hours of talk, an average of one every 3.7 seconds. In the 3-party data, there was an average of one transition every 4.7 seconds. The vast bulk (over 99%) of transitions in SWB comprised fewer than 16 intervening intervals (approximately 99%). There were vanishingly few transitions involving even numbers of intervening intervals (47 out of 256,655). One-interval transitions are the largest class, and the frequency of transitions decreases with increasing numbers of intervening intervals. All of these results reflected our earlier findings for 3-party data.

Distribution of BSTs and WSTs In SWB, 78.28% of transitions are WST, greatly outnumbering BSTs. WSTs account for 81% of 1-interval transitions, 80% of 3-interval, with proportion falling with increasing numbers of intervals to 60% of 15 interval transitions. Figure 2 shows the split between BSTs and WSTs for odd number interval transitions in SWB and in the 3-party conversations. In SWB, 47.72% of all transitions (41.65% of BSTs and 50.03% of WSTs) were accomplished with one intervening interval, 27.14% (24.77% of BSTs and 28.15% of WSTs) with two intervening intervals, and 12.86% (15.98% of BSTs and 12.16% of WSTs) with 3 intervening intervals

4 Discussion

SWB has less silence and overlap and more speech in the clear than the 3-party data - this may be due to modality; on the phone, speakers may wait for their interlocutor to finish before commencing to speak, and may give less verbal feedback in overlap. It could also reflect differences between dyadic and multi-party talk. The distribution of speaker transitions largely reflects results from the 3-party data (and also from 4- and 5- party data analysed in (Gilmartin, 2021). The largest category are 1-interval transitions, even-number interval transitions are extremely rare, and the number of transitions drops off with increasing numbers of intervals. The proportion of 1-interval transitions in SWB is greater than in 3-party, but still only accounts for 47.7% of all transitions, highlighting how most transitions involve more than a single silence or overlap, even in dyadic phone conversations. The higher incidence of WSTs than BSTs in SWB reflects results in the 3-party data. WSTs more dramatically outnumber BSTs in SWB than in the 3-party data. This could reflect long turns being taken in SWB, perhaps because participants were strangers, or indeed, may be a feature of telephone conversation.

Our analysis has shown that more than half of all BSTs and WSTs involve more than one intervening interval of speech, silence or overlap between longer stretches of single party speech. This reflects previous results on multiparty spoken interaction, implying that turn change and retention even in dyadic phone conversations exhibit a level of complexity that is not covered by modelling them as a simple gap or overlap.

5 Acknowledgements

This work was conducted with the support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant Number 13/RC/2106. The work was also funded by Swedish Research Council project 2019-02932 *Prosodic functions of voice quality dynamics* to Marcin Włodarczak.

References

E. Gilmartin, Christian Saam, Carl Vogel, Nick Campbell, and Vincent Wade. 2018. Just talking - modelling casual conversation. In *Proceedings SIGdial 2018*, pages 51–59, Melbourne, Australia.

Emer Gilmartin. 2021. *Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis*. Ph.D. thesis, Trinity College Dublin.

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

Emer Gilmartin, Mingzhi Yu, and Diane Litman. 2019. Comparing speech, silence and overlap dynamics in a task-based game and casual conversation. In *Proceedings of ICPhS 2019*, pages 3408–3412.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, pages 517–520.

Pärtel Lippus, Tuuli Tuisk, Nele Salvestre, and Pire Tiras. Phonetic corpus of Estonian spontaneous speech.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

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

Marcin Włodarczak and Mattias Heldner. 2017. Respiratory constraints in verbal and non-verbal communication. *Frontiers in Psychology*, 8:708.

"Apparently acousticness is positively correlated with neuroticism" Conversational explanations of model predictions

Alexander Berman and **Christine Howes**
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
alexander.berman@gu.se, christine.howes@gu.se

Abstract

This paper describes an experiment that collects human dialogues about predictions of participants' personality traits on the basis of their music preferences, and presents preliminary results. This type of data can inform the design of explanatory dialogue systems, and the method can straightforwardly be adapted to other domains and statistical models.

1 Introduction

When machine-learning models inform high-stakes decisions, such as in healthcare, it is important to understand what the models' estimates are based on. Under the umbrella term "explainable AI" (XAI), various techniques have been developed for explaining estimates from models that are otherwise considered opaque, such as deep neural networks. One of the most popular techniques involves constructing a simpler, linear approximation of the prediction to be explained (Ribeiro et al., 2016). However, most work in XAI has primarily targeted machine-learning experts, has not assessed explainability in naturalistic settings, and has not accounted for the interactive nature of human explanations (Miller, 2019; Arya et al., 2019; Weld and Bansal, 2018; Simkute et al., 2021). Specifically, Lakkaraju et al. (2022) report that users of current explanation techniques lack interactivity and conversational possibilities.

This paper presents a method for collecting human dialogues revolving around judgements by statistical models, as a basis for informing the design of explanatory dialogue systems and yielding requirements for XAI techniques. In a similar vein, previous work has collected dialogues where the explainer is a dialogue system (Kuźba and Biecek, 2020) or a researcher acting as the system (Hernandez-Bocanegra and Ziegler, 2021), as well as dialogues that do not specifically involve statistical estimates (Moore and Paris, 1993; Madumal et al., 2019). As far as we are aware, no previous work has collected explanatory dialogues revolving around model predictions to inform the design of XAI, with human participants/informants in both roles.

2 Experiment

Our experiment collects human explanatory dialogues about a model's predictions of personality traits from music preferences. Firstly, participants listen to 30-second excerpts of 10 tracks and rate them on a 4-point hedonic scale (like/dislike slightly/very much). In the second part, participants are paired up with each other and are randomly assigned the role of either explainee or explainer. They then chat with each other using an online text chat interface (see figure 1). Explainers, but not explainees, are given access to prediction results (estimated personality traits), information about the statistical model and what the personality traits mean, global and local feature contribution plots, and feature values (plots of the explainee's music preferences), as well as an interactive exploration enabling the explainer to make predictions for hypothetical feature values.

Since participants are paired up with each other, we avoid known issues of bias when using confederates (Kuhlen and Brennan, 2013), enabling an open-ended investigation. A high level of data protection is achieved by not asking participants about their names or contact information, not logging information that could link data to persons, and by screening collected utterances before storing them.

Tracks are featurised on the basis of 10 audio properties (energy, loudness etc.), and an explainee's ratings are aggregated into a fixed-size vector using weighted averaging. For each big-five personality trait (John et al., 1999), we train a logistic regression model to predict polarity (e.g. introverted or extraverted). As training data, we use listening histories from Last.fm and Spotify, audio features extracted from Spotify API, and psycho-

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

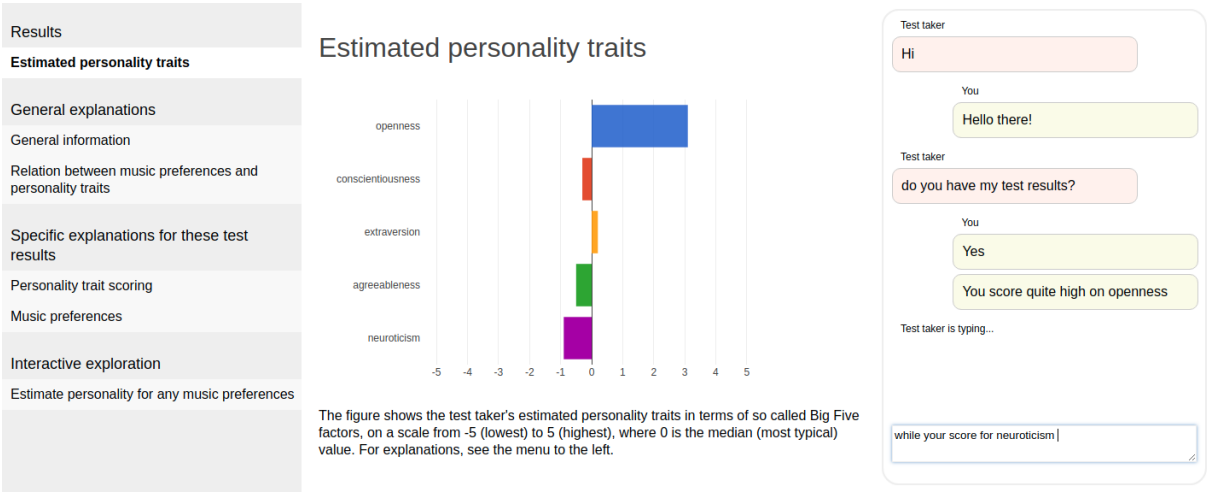


Figure 1: Screenshot of explainer's main view during chat. Explainees only see a chat window (similar to right-most part of explainer's view). Neither the personality prediction or the chat utterances are authentic.

metric test results from the MyPersonality dataset, assembled by Melchiorre and Schedl (2020). Explainers see the log odds of the predictions on a scale from -5 to 5 (see figure 1).

3 Preliminary results

Pilots have been performed with 6 colleagues from the department as participants, resulting in 3 collected dialogues (303 utterances in total). The data encompasses a range of topics including the meaning of labels ("what does agreeableness entail?"), validity of predictions ("conscientiousness is a bit too low I think"), trust ("it is hard to trust these ratings nevertheless"), causation ("I wonder if music influences the personality or if it's only the other way") and the activity as such ("It's a really fun experiment"), as well as different dialogue strategies, exemplified by the two excerpts below (A=explainer, B=explainee):

(1)

A: in terms of the "big five" factors
A: apparently, you are very open
A: almost 5 (out of -5 to 5 where 0 is the median)
B: It's interesting, I wonder what song would give this trait
A: well I actually can tell you something about that I think
A: not which song in particular, but how openness relates to features of the music
B: Oh great I'm interested

(2)

A: um apparently accousticness is positively correlated with neuroticism
B: Haha I'm almost surprised I scored low
B: And openness as well?
A: openness is the opposite with respect to acousticness
A: so I guess if you want to be more open and less neurotic the answer is to develop a preference for acoustic music

These short excerpts demonstrate that explanations given by people for the results provided by the statistical model do not necessarily adhere to the types of explanations usually considered by XAI. In the first excerpt, the explainee seems to target an examplar-based explanation; the explainer offers a correlational explanation instead, which the explainee accepts. The second excerpt exemplifies a logically incomplete explanation (Breitholtz, 2020), drawing on a shared assumption which is not explicitly stated in the dialogues (that being open and non-neurotic is desirable) and that would not necessarily be available to an AI.

4 Future work

In future work, we plan to collect more dialogues with the same setup and perform an analysis of the data. It could also be useful to focus on simpler models – e.g. rule lists or small decision trees – as well as more opaque models such as deep neural nets, with or without the support of a simpler explanation model.

Acknowledgements

This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Howes was additionally supported by VR 2016-0116, Incremental Reasoning in Dialogue (IncReD).

References

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:1909.03012*.

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

Diana C Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *CUI 2021-3rd Conference on Conversational User Interfaces*, pages 1–11.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Anna K Kuhlen and Susan E Brennan. 2013. Language in dialogue: When confederates might be hazardous to your data. *Psychonomic bulletin & review*, 20(1):54–72.

Michał Kuźba and Przemysław Biecek. 2020. What would you ask the machine learning model? Identification of user needs for model explanations based on human-model conversations. In *ECML PKDD 2020 Workshops*, pages 447–459, Cham. Springer International Publishing.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1033–1041.

Alessandro B. Melchiorre and Markus Schedl. 2020. *Personality Correlates of Music Audio Preferences for Modelling Music Listeners*, page 313–317. Association for Computing Machinery, New York, NY, USA.

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

Johanna D Moore and Cécile L Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. Technical report, University of Southern California, Marina Del Rey Information Sciences Institution.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Auste Simkute, Ewa Luger, Bronwyn Jones, Michael Evans, and Rhianne Jones. 2021. Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, 7-8:100017.

Daniel S. Weld and Gagan Bansal. 2018. Intelligible artificial intelligence. *CoRR*, abs/1803.04263.

Evaluation of a Spoken Argumentative Dialogue System for Opinion-Building

Annalena Aicher and **Wolfgang Minker**
Institute for Communications Engineering
Ulm University, Germany
annalena.aicher@uni-ulm.de

Stefan Hillmann and **Thilo Michael** and **Sebastian Möller**
Quality and Usability Lab
TU Berlin, Germany

Stefan Ultes
Mercedes Research & Development
Sindelfingen, Germany

Abstract

Speech interfaces for argumentative dialogue systems (ADS) are rather scarce and quite complex. To provide a more natural and intuitive interface, we include an adaption of a recently introduced natural language understanding (NLU) framework tailored to argumentative tasks into a complete end-to-end ADS. Within this paper, we investigate the influence of two different I/O modalities and discuss issues and problems we encountered in a user study with 202 participants using our ADS.

1 Introduction

The exchange of arguments and conversation with humans via natural language demand for a flexible natural language understanding (NLU), an argumentative dialogue structure, and the integration of commonsense knowledge. The speech-driven argumentative dialogue system (ADS) we introduce in this paper combines these components and enables the user to scrutinize arguments on both sides of a controversial topic. Unlike most approaches to human-machine argumentation (Slonim et al., 2021; Rosenfeld and Kraus, 2016; Le et al., 2018; Rakshit et al., 2017; Chalaguine and Hunter, 2020; Fazzinga et al., 2021) we pursue a cooperative exchange of arguments. Our aim is a system that cooperatively engages the users to explore arguments and to state their preferences in natural language. Therefore, we modified and extended our previously introduced menu-based ADS (Aicher et al., 2021). The speech-based system is evaluated and compared to the robust baseline in terms of naturalness and usability aspects in a crowd-sourcing study with 202 participants.

2 ADS Interface and NLU Framework

The system's graphical user interface (GUI) is illustrated in Figure 1. The interface can either provide a drop-down menu or speech input. In the drop-down system, users can choose their action by clicking. In the speech system, our formerly introduced NLU framework (Abro et al., 2022) processes the spoken user utterance. Its intent classifier uses the BERT Transformer Encoder presented by Devlin et al. (2018) and a bidirectional LSTM classifier.

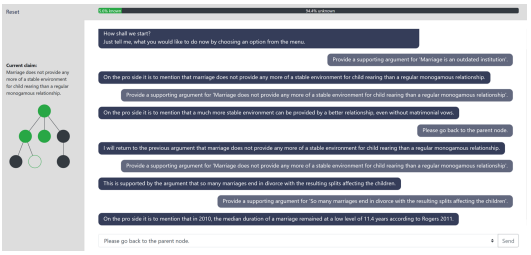


Figure 1: GUI of the menu system with folded drop-down menu. Above the drop-down menu the dialogue history is shown.

In the speech-based system, instead of the drop-down menu, a button labelled "Start Talking" is shown. The button is pressed to start and stop the speech recording. Except for this difference, both systems share the same architecture. The system's generated output is based upon the original textual representation of the argument components on the sample debate topic *Marriage is an outdated institution* taken from *Debatabase* of the idebate.org[1] website.

3 User Study Setting

The study was conducted online via the crowd-sourcing platform "Crowdee" (www.crowdee.com, 12-29th November 2021) with participants from the UK, US and Australia. All 202 participants (menu: 104 [50 female, 54 male], speech: 98 [39 female, 59 male]) were non-experts without a

[1] https://idebate.org/debatabase (last accessed 23th February 2022). Material reproduced from www.iedebate.org with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

topic-specific background. After an introduction to the system (short text and demo video), the users had to listen to enough arguments to build a well-founded opinion. As soon as ten arguments were heard, the end of the interaction could be chosen freely. Afterwards, the participants had to rate the interaction in 40 statements[2] on a 5-point Likert scale (1 = totally disagree, 5 = totally agree).

4 Results and Discussion

In average the interaction with the system last 31.45 minutes (menu: 27.57 speech: 35.34). This significant difference can be explained by the fact that the spoken interaction (speaking and hearing) inherently takes longer than clicking on an option in the drop-down menu and reading the response. Another significant difference is observable in the number of heard arguments (average menu/speech: 22/15). Even though the average time the users of the menu system interacted with the ADS is lower, the number of provided arguments is significantly higher compared to the speech system. 9.6%/17.3% of the menu/speech system participants quit the conversation after hearing the minimum number of 10 arguments (in total: 13.4%). Most of the participants heard between 20-30 arguments of 72 available arguments. Whereas some participants in the menu system listened to even more arguments, only one participant of the speech system did so. The category "Overall Quality" ("What is your overall impression of the system?") is rated on a specific 5-point Likert scale (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad). We perceive a highly significant[3] ($\alpha < 0.01$) difference between both systems, as the menu system with a rating of 3.49 outperformed the speech system rated with 2.66. Altogether, the speech system is significantly outperformed in all categories of the questionnaire. The biggest differences were perceivable in ratings concerning errors which occurred or whether the system provided the expected information. Clearly, this points to a lack in processing the user utterances (errors in the ASR or NLU module). By checking the dialogue logs of the interactions with users in the speech system, we found that about 15% of all speech utterances were processed erroneously. Even though in 70% of these cases the NLU identified the correct intent, the results show that this has had a considerable impact on user perception of the speech system. Furthermore, we noticed inconsistencies in the user behavior, e.g. repetition of requests multiple times and ignoring the system's answer to choose another action. In contrast to the menu system which only displayed the possible actions, the speech users had to figure out what actions they can perform and formulate them. Even though the speech system offered a "Help" button, as well as the "available options" action, only 1.3% of the participants used them. This might be explained by the fact that only 35% of the users spend enough time on the introduction website to read through the explanation and watch the video properly. This is further underpinned by users' feedback, stating that "It was not possible to do what I wanted to do. I repeated myself many times"/"I was stuck in the argument and could not get back.". The results show that the I/O modalities and respective difficulties/problems decrease the rating of the general impression of the system, even in aspects which have no relation to the former. E.g. the incremental approach to present arguments, the sufficiency of different options or the conclusiveness of arguments which are content- but not modality-dependent, are rated significantly worse in the speech than in the menu system. Therefore, it is crucial to solve the identified issues and to introduce a double-staged study setting, which ensures the participants understood how to interact with the system. Even though the introduced speech system does not outperform the menu baseline, we could show that the menu system provides a robust baseline that tends to be rated positively in almost every question. Thus, it suits as a robust baseline to which enhanced spoken ADS versions can be compared to.

5 Conclusion and Future Work

In this work, we evaluated an ADS in two I/O modalities by conducting a crowdsourcing study. Due to an erroneous ASR module and issues in understanding how to communicate with our ADS via speech, we observed that the latter was outperformed significantly by our strong menu baseline. In future work, we will enhance the system's ASR and NLU robustness by training on larger data-sets and including a request for repetition if the intent prediction accuracy falls below a threshold.

[2]Taken from a questionnaire according to ITU-T Recommendation P.851 (P.851, 2003)

[3]To determine whether the difference between the two system means is significant, we used the non-parametric Mann-Whitney U test (McKnight and Najab, 2010) for two independent samples with no specific distribution.

References

Waheed Ahmed Abro, Annalena Aicher, Niklas Rach, Stefan Ultes, Wolfgang Minker, and Guilin Qi. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems*, 242:108318.

Annalena Aicher, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2021. Opinion building based on the argumentative dialogue system bea. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 307–318. Springer.

Lisa A. Chalaguine and A. Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for covid-19 vaccine information. In *Logic and Argumentation*, pages 477–485, Cham.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Patrick E. McKnight and Julius Najab. 2010. *Mann-Whitney U Test*, pages 1–1. American Cancer Society.

ITU-T Recommendation P.851. 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003). International Telecommunication Union.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems*, pages 45–52.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *ECAI'16*, pages 320–328.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, and Lilach Edelstein. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Edge Cases of Discourse Salience in American English Casual Dialogs: A New Window into the Co-Constructed Nature of Social Conversation

Alex Lưu
Brandeis University
alexluu@brandeis.edu

1 Discourse Salience in Social Dialog

[1] According to functional-cognitive literature, **salience** in dialog can be defined as a real-time property of mental entities which accounts for the fact that these mental entities come to be in the foreground of an interlocutor's attention at the given time and therefore are referable from that interlocutor's perspective (cf. Nolan, 2022). The term "**discourse salience**" is adopted to refer to salient content derivable from linguistic cues available in conversational discourse, such as morphosyntactic marking, noun phrase form and definiteness, syntactic role and construction, and discourse structure co-constructed by interlocutors.

I am conducting an empirical study of discourse salience in social conversation, using linguistic annotation and discourse analysis of NEWT-SBCSAE, a publicly accessible corpus of naturally occurring casual dialogs in American English (Du Bois et al., 2000; Riou, 2015; Lưu and Malamud, 2020a). Performed by Linguistics majors and native speakers of English[2], the annotation focuses on the arguments of coherence relations in NEWT-SBCSAE and covers different linguistic aspects characterizing the "main point" of utterances constituting these arguments[3]. The discourse analysis includes statistical examination of annotation results and conversation analysis of target instances filtered based on annotated categories.

This paper focuses on analyzing edge cases of discourse salience identified based on two main aspects of information packaging at the utterance level, namely the given-new ordering of information and syntactic variations for realizing that ordering. The investigation scope is narrowed down to inter-speaker coherence relations, the first choice to explore the co-constructed nature of social dialog.

The examples examined in this paper are formatted as shown in Table 1. The center of each example is the bold utterance, which encapsulates an instance of edge cases, and its surrounding utterances which are connected to it via annotated coherence relations (in parentheses) (Lưu and Malamud, 2020a). The index of each utterance reflects its chronological order (with the increment of 1).

Utterance	Simplified transcript
38-AL (entity)	*Bill comes over with his leatherman toolman or or whatever it is.*
39-AL	***Few minutes he had it undone. [laughter]***
40-AN (concession)	*So she can't use it now though.*

Table 1: A contextualized utterance (in bold) from dialog *SBC043Spoonfuls* in NEWT-SBCSAE between Alice (AL) and her daughter Annette (AN).

2 Edge Cases of Discourse Salience

There is a scholarly consensus that the given-before-new ordering of information in discourse is preferred; and among syntactic variations for realizing that ordering, canonical word order (CWO), e.g. subject-verb-object (SVO) in English, is unmarked (Prince, 1992; Birner, 2012, inter alia). Generally, CWO is felicitous even in the context where it doesn't adhere to the preferred ordering of information, while noncanonical word order (NWO)[4] is felicitous only when it is used to realize the preferred information structure. In this work, new-before-given ordering and NWO characterize the edge cases of discourse salience as illustrated in Table 1. Knowing that the pronoun *it* in the utterance **39-AL** refers to an attaché case in prior discourse, we can conclude this utterance features the new-before-given ordering in the information

[1] This paper's live version is located at https://osf.io/cedvx/.

[2] From North-Eastern US. They were paid $16/hour.

[3] This is grounded in the concept of at-issueness in formal semantics and pragmatics (e.g. Koev, 2018, inter alia).

[4] Including preposing (e.g. topicalization and focus-movement), postposing (e.g. existential and presentational *there*), argument reversal (e.g. inversion and passivization), their combinations, and cleft constructions (*wh-*, *it-*, *th-*).

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

exchange dimension: *few minutes* is newer than *he had it undone*. Moreover, the utterance is a NWO sentence as *few minutes* is preposed.

Among 1920 annotated arguments of inter-speaker coherence relations available in NEWT-SBCSAE there are 14 new-before-given cases (0.73%) and 95 NWO cases (4.95%).

New-before-given Two clear new-before-given categories emerge from conversation analysis:

- dialogic resonance (Du Bois, 2014) (T. 2, 3)
- non-epistemic emphasis (Lưu, 2022) (T. 4)

Utterance	Simplified transcript
2886-K (temporal)	*I left my bag there.*
2890-S	***Now the ghosts'll get it.***
2892-K (entity)	*Ghosts'll get it.*

Table 2: An example in *SBC034Times* between a couple.

The utterance **2890-S** in Table 2 shares a similar sequence of information slots with 2886-K – a triple of an agent, an action, and the bag (referred to by the noun phrase *my bag* and pronoun *it*). While this similarity results in the change of information ordering from given-before-new in 2886-K to new-before-given in **2890-S**, it preserves dialogic resonance at the syntactic level between two arguments of a coordinating coherence relation (temporal).

Utterance	Simplified transcript
225-AL (entity)	*Oh and you know how I get when my heart just beats really fast?*
229-AN	***Cathleen has to wear a heart monitor because of that mom.***
230-AL (entity)	*When did she get that?*
236-AL (entity)	*Would hers do that stop and then get real fast and?*

Table 3: An example in *SBC043Spoonfuls*.

In Table 3 dialogic resonance happens at the social level: the utterance **229-AN** is an interlocutor-decentric move (Lưu and Malamud, 2020b) from topics focusing on the hearer, Alice – the mother, to *Cathleen*. As the immediately preceding discourse of **220-AN** is solely dedicated to how overwhelmed Alice was at work, it is reasonable for Annette, the daughter and the speaker of **229-AN**, to lighten the conversation by switching the social focus to a third person at this moment. The resonance pattern here is someone had some trouble recently.

Different from above examples, the preposed new content in utterance **2582-D** (Table 4) is the speaker's strong self-positioning (*I do know*) and

Utterance	Simplified transcript
2580-P (concession)	*You haven't read the book so you don't know.*
2582-D	***Yeah but I do know it it's an awfully it's it's an awfully presumptuous thing to sit down and write a book about death when you haven't died.***
2583-P	*But.*
2584-P (concession)	*It has it has it has stories in there from from the Zen and.*

Table 4: An example in *SBC005Book* between a couple.

expressive evaluation (*it's an awfully presumptuous thing*), demonstrating the dominance of the normative and affective dimensions in discourse salience. This non-epistemic emphasis allows the speaker to detach from the preferred information ordering.

Noncanonical word order All new-before-given cases examined above, except for the one in Table 1, preserve CWO and therefore confirm its felicity in the context of non-preferred information ordering. Examining the sole exception (**39-AL**), we can observe that it ends with laughter and therefore involves non-epistemic dimensions, similar to the case in Table 4. The difference is **39-AL** doesn't involve strong self-positioning, which is usually realized by *I* in the subject position. Thus, we can argue that the epistemically older information in **39-AL**, *he had it undone [laughter]*, is actually a new focus in the affective dimension. Consequently, NWO expressed in this utterance is still used to realize the preferred given-before-new structure, but in a non-epistemic dimension.

It's worth noting that the minor portion of NWO (4.95%) in annotated data supports CWO as the easiest and preferred way to produce salient content in spontaneous conversation, in which interlocutors constantly faces pressure of real-time interaction.

3 Conclusion

The new findings based on examining the edge cases of discourse salience are directly relevant to social dialog system modeling and evaluation. They confirms the importance of non-epistemic dimensions and relational work interlocutors rely on to co-construct their utterances' meaning. The findings also reveal concrete discourse configurations of these understudied aspects. As a result, this work demonstrates how theoretical work both underpins and arises from the empirical.

Acknowledgements

I am extremely grateful to my annotators, Eben Saveson and Tali Tukachinsky, for their curiosity, diligence and creativity. My deepest gratitude goes to Sophia A. Malamud for her active encouragement and thorough feedback on this paper.

References

Betty J Birner. 2012. *Introduction to Pragmatics*. John Wiley & Sons.

John W. Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410. Publisher: De Gruyter Mouton.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa Barbara corpus of spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Todor Koev. 2018. Notions of at-issueness. *Language and Linguistics Compass*, 12(12):e12306.

Alex Lưu. 2022. Sketching a linguistically-driven reasoning dialog model for social talk. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 153–170, Dublin, Ireland. Association for Computational Linguistics.

Alex Lưu and Sophia A. Malamud. 2020a. Annotating coherence relations for studying topic transitions in social talk. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 174–179, Barcelona, Spain. Association for Computational Linguistics.

Alex Lưu and Sophia A. Malamud. 2020b. Non-topical coherence in social talk: A call for dialogue model enrichment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 118–133, Online. Association for Computational Linguistics.

Brian Nolan. 2022. *Language, Culture and Knowledge in Context: A Functional-Cognitive Approach*. Equinox Publishing Ltd.

Ellen Prince. 1992. The ZPG letter: Subjects, definiteness, and information status. In William C. Mann and Sandra A. Thompson, editors, *Discourse Description: Discourse Analyses of a Fundraising Text*, pages 295–325. Amsterdam: John Benjamins.

Marine Riou. 2015. *The Grammar of Topic Transition in American English Conversation. Topic Transition Design and Management in Typical and Atypical Conversations (Schizophrenia)*. Ph.D. thesis, Université Sorbonne Paris Cité.

The construction of stereotypes through language: the case of evidential markers

Mercedes González Vázquez
Universidade de Vigo

Abstract

The aim of this paper is to link the existence of some cultural stereotypes such as indirectness and tentativeness -attributed to high context cultures and negative politeness cultures- with listeners' interpretation of the use of evidential markers.

1 Introduction

Following Hall (1976), in high-context cultures the interpretation of meaning relies heavily on context as for example Asian and Arabic countries—like Japan, Korea, Iran, Iraq, Saudi Arabia, etc. (Lewis 2006) and English (Grainger and Mills 2016) and Irish (Kallen 2005). In contrast, low-context cultures tend to be explicit, clear and direct with minimal inference (American, German and Ducht cultures, vid. Havertake 1994, Lewis 2006). Concerning negative politeness cultures (Asian, Arabic countries, Scandinavian, English, Irish, etc.; Brown & Levinson 1987; Havertake 1994, Kallen 2005), they avoid intruding on others personal territory - not to impose on the listener - and involve indirectness). On the other hand, evidentiality is the indication of the source of information on which speakers rely when stating something, which is divided into inferential and reportative evidentiality. The inferential expressions *it seems, it appears, it must* convey that speaker accesses the information through an inferential process based on perceptual evidences or based on reasoning from knowledge of the world. The reportative evidentials *it is said*, *reportedly, apparently, allegedly*, etc. show that the propositional content of the speech has been previously uttered by another person or people, in the absence of the actual speaker (vid. Wiemer & Marín-Arrese, 2022).

2 Evidentiality and related notions

As it is known, one of the bases for the construction of cultural stereotypes lies in the different ways individuals interact linguistically. Focusing on verbal interaction, we consider that the indication of the source of information contributes to the creation of cultural stereotypes, since cultures show a different way of using evidential devices, and therefore, a different way of interpreting it. An obvious example of this is the following. In many aboriginal languages (Quichua, Nanti, Western Apache, etc., Nuckolls 2012) evidentiality constitutes an obligatory category that confers a social status on the speaker according to which he/she is recognized as socially well-integrated, and reliable person. If the evidential devices are omitted or misused, the speaker is considered a liar, opaque, and unreliable: he/she speaks for the sake of speaking without any informative basis (Mansfield 2019, Hintz & Hintz 2017). In languages with non-obligatory evidentiality, the omission of this category is the default situation. The appearance of evidentiality is generally interpreted as a manifestation of attenuation and lack of certainty (Mushin 2013) and stereotypes can also emerge. In this paper we will focus on non-evidential languages in general (e.g. European languages).

Based on the theory of information territory (Kamio 1997), it is argued that evidentiality in discourse shows instances of use that are not strictly evidential but rather determined by the epistemic vigilance of territories (Sperber & Wilson 1995, Heritage 2012).

The perception of different territories of information that belong to each participant - territory of the speaker, territory of the

listener, common territory shared by both participants or common ground, and territory beyond the reach of both- entails epistemic rights for the participants. For example, the information obtained through external and internal direct experience, knowledge related to the speaker's field of specialization, and information about people, objects and events close to the speakers belong to the speaker`s territory (Kamio 1997). In this case, the speaker possesses epistemic right over the information and the information usually does not require evidential justification. On the contrary, if the speaker wants to make a statement about someone else's state of mind, he/she must use an evidential form that expresses how the speaker came to know something that is not in their territory of information.

As it is known, crossing the boundary of our own information territory and intruding into the listener's is to claim more epistemic rights for ourselves than we are entitled to, thus infringing the social norms of verbal conduct. Thus, the awareness-raising of other people's territories in conjunction with negative politeness leads to the use of evidential, since they facilitate the speaker to demarcate territories of information in three ways:

1.- The information is outside his/her territory, and consequently, he/she does not consider him/herself to have epistemic primacy:

(1) It seems that there will be good weather at the weekend (inferential evidentiality)

(2) Apparently his death was due to poison (reportative evidentiality)

2.- It delimitates information belonging to the territory of the speaker, but of which he/she is not certain:

(3) Apparently / it is said that/ It seems that my partner cheats on me (reportative/ inferential)

3.- Another case is related to situations in which the propositional content belongs to the listener`s territory. In these situations the listener has epistemic primacy, and therefore, more epistemic rights to make assertive statements than the speaker. The intention of the speaker is to corroborate the information, and to show negative politeness:

(4) Apparently you are going to move to another city (reportative, not lack of certainty)

(5) It seems that you are annoyed this morning (inferential, not lack of certainty)

(6) It is said that you were at the Oscars (reportative)

Nevertheless, there is a difference in speakers' perception of whether a certain piece of information belongs to their territory or not (Kamio 1997), since the way the territory of information is organized in languages is influenced by culture and, consequently, its interpretation changes culturally. The speakers transfer their system of territory of information into other cultures, arising stereotypes.

Thus, the situations in which the speaker perceives the need to resort to evidential cueing is cultural dependent, as the interactants negotiate differently the boundaries of territories of knowledge. A proof of this is the higher use of evidentials in British English and Japanese in comparison with American English (Kamio 1997, Trent 1998, Precht 2003).

3 Conclusion

The fact that evidentiality guide interlocutors to link relevant information and help them discern the different territories makes the message more implicit and open-ended. Implicitness is a characteristic of high-context cultures, which can integrate the use of evidentials as a matter of course. In low-context cultures, where communication is direct and explicit, they are more reluctant to make use of evidentials except in the necessary cases (as a strict source of information or as a strong nuance of doubt). We can conclude that low and high context cultures show a different perception of the territories of information. The latter are very sensible to the territories of information of each participant and handle implicitness as a useful part of the language. Consequently, evidentials are more frequently used. This difference of use in evidential expressions reinforces the cultural stereotype of indirectness and tentativeness attributed to high context cultures.

References

Brown, Penelope and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. New York: Cambridge University Press.

Clift, Rebecca. 2006. Indexing stance: Reported speech as an interactional evidential. *Journal of Sociolinguistics* 10/5, 569–595.

Du Bois, John. 2007. The Stance Triangle. In: Englebretson, R. (Ed.), *Stancetaking in Discourse:Subjectivity, Evaluation, Interaction.* Amsterdam: John Benjamins. 139-182.

Fox, Barbara. 2001. Evidentiality: Authority, Responsibility, and Entitlement in English Conversation. *Journal of Linguistic Anthropology* 11(2), 167-192.

Grainger, Karen and Sara Mills. 2016. *Directness and Indirectness Across Cultures.* New York: Palgrave Macmillan.

Hall, Edward. 1976. *Beyond Culture.* New York: Anchor/Doubleday.

Haverkate, Henk. 1994. *La cortesía verbal. Estudio pragmalingüístico*. Madrid: Gredos.

Heritage, John. 2012. Epistemics in action: action formation and territories of knowledge. *Research on Language & Social Interaction* 45 (1), 1–29.

Hintz, Daniel and Diane Hintz. 2017. The evidential category of mutual knowledge in Quechua. *Lingua* 186, 209-228.

Kallen, Jeffrey L. 2005. Silence and mitigation in Irish English discourse. In: Barron, Anne & Klaus P. Schneider (eds.) *The pragmatics of Irish English.* Berlin: Walter de Gruyter. Pp.47-72.

Kamio, Akio. 1997. *Territory of Information.* John Benjamins, Amsterdam.

Lewis, Richard D. 2006. *When cultures collide: Leading across cultures.* London: Nicholas Brealey International.

Mansfield, John. 2019. Epistemic authority and sociolinguistic stance in an Australian Aboriginal language. *Open Linguistics* 5, 25-48.

Michael, Lev. 2012. Nanti self-quotation: implications for the pragmatics of reported speech and evidentiality. *Pragmatics and Society* 3 (2), 321-357.

Mushin, Ilana. 2013. Making knowledge visible in discourse: Implications for the study of linguistic evidentiality. *Discourse Studies* 15(5), 627–645.

Nuckolls, Janis. 2012. From quotative other to quotative self: Evidential usage in Pastaza Quichua. *Pragmatics and Society* 3(2), 226–242.

Nucholls, Janis and Lev Michael. 2014. *Evidentials in interaction.* Amsterdam & Philadelphia: John Benjamins.

Precht, Kristen. 2003. Stance moods in spoken English: Evidentiality and affect in British and American conversation. *Text* 23(2), 239–257.

Sperber, D. and Wilson, D. 1995. *Relevance: Communication and Cognition*. Oxford: Blackwells

Trent, Noriko.1998. Cross-cultural discourse pragmatics: Speaking about hearsay in English and Japanese. In *Texas Papers of Foreign Language Education* 3 (2), 1-3.

Wiemer, Björn and Marin-Arrese, Juana I. 2022. *Evidential Marking in European Languages:Toward a Unitary Comparative Account.* Berlin /Boston: De Gruyter Mouton.

Appendix

We present examples of evidential units that demarcates territory rather than a lack of certainty in Galician (a Romance language spoken in the autonomous region of Galicia in the North West of Spain, which we consider to be one of the cultures of high context and negative politeness). In languages whose culture does not rank as high-context as Galician (e.g. Spanish) the evidential markers would be omitted since the speaker knows with certainty what he/she is claiming. In Galician the speaker is interested in marking that he/she has accessed this information not by him/herself but from external sources, even if he/she is sure of it. The information marked by the evidential belongs to different territories in Galician and Spanish. The use of evidentiality helps to create stereotypes (e.g. attenuation, implicitness, tentativeness and indirectness) about the language that uses them because of the different interpretation implied by high- and low-context cultures.

Examples for this study have been drawn from the written Corpus: Corpus of Reference of Current Galician (CORGA). http://corpus.cirp.es/corga/buscas

Galician evidential markers: *seica* (`apparently', 'it seems', 'it is said', 'so'), *disque* (`it is said`), *parece* (`it seems`).

Examples

(1) Logo fomos ver teatro, *seica* botaban "O velorio" do grupo Troula, mais cando chegamos, as localidades estaban esgotadas. Perante esta situación decidimos tomar algo no Universal.

Then we went to the theatre, *apparently* they were showing "El velatorio" by the group Fiesta, but when we arrived, the tickets were sold out. In this situation we decided to have a drink at the Universal.

(2) Arrepiouseme todo o corpo, saíronme as bagoas, *seica* chorei, aínda que disimulando como podía. Tan lonxe do meu país aquela música!

My whole body started to shake, my eyes welled up with tears, it seems that/I guess I cried, even though I was trying my best not to. That music so far away from my home country!

(3) *Seica* estiveches na casa de María.

So / apparently you've been to Maria's house.

(4) Xa deixaron de traballa-la terra, logo?
- Pois *disque* si. Xa hai moitisimo tempo.

- Have they already stopped working the fields, then?
- Well, *apparently yes/it seems* so. They already did a long time ago.

(5) -Que, Roxelio, *seica* non hai moito que facer. Imos tomar un vaso?

Then, Roxelio, *it seems/apparently* you have not much to do. Shall we go for a drink?

(6) Eiquí non hai nin restos de don Xaquín, as cousas están todas en orden e, ó menos así ó primeiro visual *non parece faltar* nada.

Here there is no trace of Mr. Xaquín, everything is in order and, at least the first impression is that nothing *seems to be missing.*

Investigating code-switching and disfluencies in bilingual dialogue

Fahima Ayub Khan and **Bill Noble**
Department of Linguistics, Philosophy and Theory of Science
University of Gothenburg
fahima.ayub.khan@gu.se; bill.noble@gu.se

Abstract

This paper investigates the relationship between disfluency and code-switching in bilingual dialogue. We examine a corpus of 41 bilingual (Spanish-English) conversations and test the hypothesis that code-switching can be a response to negative evidence of grounding in the form of disfluencies. We find that there is a statistically significant relation between disfluencies and code-switching. Particularly, disfluencies have a positive effect on within turn code-switching.

1 Introduction

This paper investigates the communicative function of code-switching (switching between languages) in bilingual dialogue, as it relates to disfluencies, such as filled pauses and self-repair.

Disfluency is not just an interruption to the normal flow of conversation. On the contrary, disfluencies have a crucial role to play in facilitating communicative alignment and coordinating interaction (Hlavac, 2011). It is well-established that repair is crucial for aligning speakers and establishing common ground in dialogue (Healey et al., 2018, 2013, 2011). Furthermore, psycholinguistic studies have noted that disfluencies facilitate referential disambiguation, leading to greater interactive efficiency (Bailey and Ferreira, 2007). In bilingual dialogue, code-switching is an additional resource that speakers can draw on to facilitate conversational alignment (Wei and Milroy, 1995; Cromdal and Aronsson, 2000). Furthermore, previous work has demonstrated that code-switches tend to occur in turns with disfluencies (such as "um", "er", etc.,) and in clarification requests (Beatty-Martínez et al., 2020; Kootstra et al., 2020).

However, the mechanisms of interaction between disfluencies and code-switching are not well-studied. Is code-switching a response to communicative problems indicated by disfluency? If so, is it a response to disfluencies produced by other speakers, or also by oneself? More generally, why is it that disfluencies appear with greater frequency in the vicinity of code-switches? In this paper, we test the hypothesis that code-switching is a repair mechanism in bilingual dialogue—i.e., that code-switching can be a response to negative evidence of grounding, such as disfluency or negative feedback.

2 Method

2.1 Data

We investigated this hypothesis using the Bangor Miami corpus (Deuchar, 2010) which is a set of 56 spontaneous conversations between Spanish-English bilinguals living in Miami (USA). Of these we excluded 15 conversations that included only one of the two participant's turns. The final dataset contains 41 dialogues, 40 841 turns, and 254 739 tokens.

The conversations are transcribed in the CHAT format (MacWhinney, 2022), and include token-level language annotation. This makes it possible to pinpoint code-switches both within and between turns.[1] For the purpose of this study, we define intersentential switching as using both languages between turns. Instances where participants code-switch in the same turn has been coded as intrasentential switching in our analysis. In order to determine switch between turns, we coded turns whose language tag was different from the language tag corresponding to the last token of the previous turn. The data contains 4971 turns (12.2%) that switch from the language of the previous turn (i.e., between turn switches) and 2215 turns (5.4%) with within-turn switching. In total, 6539 turns (16%) are code-switched.

The CHAT transcription format also includes

[1] Some tokens, such as proper names, are coded as belonging to both languages. For the purposes of detecting code-switches, we consider these tokens to have the language tag of the previous token in the turn (or the final token in the previous turn, if it is the first token).

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

fine-grained annotations for disfluencies. In total, 6694 turns out of 40 841 (16.4%) contained disfluencies. Under the category of disfluencies, we included repetition, repairs, alteration, filled and unfilled pauses.

2.2 Statistical models

The analysis was done using the lme4-package (Bates et al., 2015) in R Studio (Team, 2021). The data was fitted with mixed-effects logistic regression models to predict the interaction between code-switching and disfluencies.

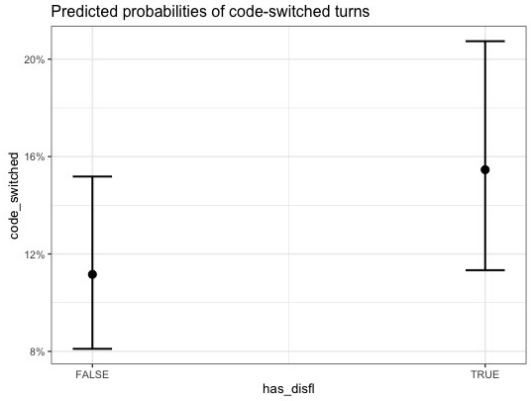
For the first part of the analysis, we built a baseline model with code-switching as the dependent measure. We incrementally added 'turns with disfluencies' variable as the predictor along with 'previous turns with disfluencies' variable as an additional predictor. The speakers and the dialogue ID were included as random effects. In order to check the model fit, the models were compared using a log-likelihood ratio chi-square test. Based on the results of the chi-square distribution, predictors were added or removed in the best-fit model.

For a detailed analysis on the interaction between code-switching and disfluencies, we built models to test the effect of disfluencies on code-switching within and between turns.

3 Results

The model we built for the first analysis yielded a positive and significant effect of disfluencies on code-switching (β = 0.38, 95% CI [0.30, 0.45], p < .001). The best fitting model predicts that speakers code-switch within and between turns after encountering disfluencies in their turn and in the preceding turn. The model's total explanatory power is substantial (conditional R2 = 0.30) and the part related to the fixed effects alone (marginal R2) is of 4.13e-03. The results indicate that speakers code-switched in turns where disfluencies occurred.

The second set of models investigating the type of code-switching that is predicted by disfluencies yielded a significant effect on within turn switching (β = 0.14, 95% CI [0.02, 0.25], p = 0.022). While disfluencies in general have a significant effect on code-switching within a given turn, disfluencies in the previous turn also have a statistically significant effect on within turn switching. The models testing the effect of disfluencies on between turn switching did not yield a significant effect.



4 Discussion

This study is a point of departure for investigating code-switching as an interactive resource to facilitate grounding in bilingual dialogue. We analysed an available bilingual dialogue corpus (Deuchar, 2010) in order to gain some preliminary insights on the effect of disfluencies on code-switching in dialogue. It has to be noted that these results are specific to the context of Spanish-English bilinguals who are fluent in both languages. The results from the analysis indicate a strong relation between disfluencies and code-switching in bilingual dialogue. The results are similar to the findings from previous studies (Hlavac, 2011) on the frequency of pauses and repairs occurring alongside code-switching.

We are extending our analysis to further investigate the effect of each type of disfluency (repairs, pauses, etc.,) on each type of code-switching (within-turn and between-turn). Alternatively, code-switching could also trigger disfluencies since code-switching is cognitively demanding (Green and Abutalebi, 2013). The additional models we built to test this revealed that disfluencies can be predicted by code-switching. To what extent code-switching and disfluencies in bilingual dialogue affect each other can only be investigated further by testing this effect within a controlled experimental setting.

5 Conclusion

The results of this study have confirmed that there is a clear relationship between code-switching and disfluency in spontaneous bilingual interaction. We intend to investigate this further in dialogue-based experiments where we can closely control the effect of disfluencies and the interactive context. In our future work, we will additionally examine the effect of code-switching in interaction across language pairs.

References

Karl GD Bailey and Fernanda Ferreira. 2007. The processing of filled pause disfluencies in the visual world. In *Eye movements*, pages 487–502. Elsevier.

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

Anne L. Beatty-Martínez, Christian A. Navarro-Torres, and Paola E. Dussias. 2020. Codeswitching: A Bilingual Toolkit for Opportunistic Speech Planning. *Frontiers in Psychology*, 11.

Jakob Cromdal and Karin Aronsson. 2000. Footing in bilingual play. *Journal of Sociolinguistics*, 4(3):435–457.

Margaret Deuchar. 2010. BilingBank Spanish-English Bangor Miami Corpus.

David W Green and Jubin Abutalebi. 2013. Language control in bilinguals: The adaptive control hypothesis. *Journal of cognitive psychology*, 25(5):515–530.

Patrick Healey, Mary Lavelle, Christine Howes, Stuart Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker's troubles. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Patrick GT Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, pages 11–13. Citeseer.

Patrick GT Healey, Gregory J Mills, Arash Eshghi, and Christine Howes. 2018. Running repairs: Coordinating meaning in dialogue. *Topics in cognitive science*, 10(2):367–388.

Jim Hlavac. 2011. Hesitation and monitoring phenomena in bilingual speech: A consequence of code-switching or a strategy to facilitate its incorporation? *Journal of Pragmatics*, 43(15):3793–3806.

Gerrit Jan Kootstra, Ton Dijkstra, and Janet G Van Hell. 2020. Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in psychology*, page 1747.

Brian MacWhinney. 2022. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. *Applied Psycholinguistics*.

RStudio Team. 2021. Rstudio: integrated development for r. rstudio, pbc, boston, ma. 2020.

Li Wei and Lesley Milroy. 1995. Conversational code-switching in a chinese community in britain: A sequential analysis. *Journal of Pragmatics*, 23(3):281–299.

On System-Initiated Transitions in a Unified Natural Language Generation Model for Dialogue Systems

Ye Liu[1], **Yung-Ching Yang**[2], **Wolfgang Maier**[3], **Wolfgang Minker**[4] and **Stefan Ultes**[5]
[1,3,5]Mercedes-Benz AG, Sindelfingen, Germany
[1,4]Ulm University, Ulm, Germany, [2]University of Stuttgart, Stuttgart, Germany
[1,3,5]{ye.y.liu,wolfgang.mw.maier,stefan.ultes}@mercedes-benz.com
[2]yangching18@gmail.com, [4]wolfgang.minker@uni-ulm.de

1 Introduction and Motivation

Spoken Dialogue Systems (SDS) have been separately developed under two different categories: task-oriented and chit-chat. The former focuses on achieving functional goals and the latter aims at creating engaging social conversations without special goals. Creating unified models (Lin et al., 2021; Zhao et al., 2021; Young et al., 2022) that can reply to both task-oriented and chit-chat requests is a promising research topic in recent years. It is a desired property of unified models to be sensitive to dialogue mode transitions (switch from chit-chat to task-oriented or from task-oriented to chit-chat) and domain transitions in a task-oriented multi-domain setting (switch from one domain to another domain). The dialogue agent can then proactively guide the transition through generating a transition sentence (red parts in Figure 1).

We will discuss the *system-initiated transitions* of a unified dialogue agent. First, we elaborate the motivation behind this work. If the dialogue agent can track the preceding chit-chat interaction and recognize the potential user intention for requiring some task-oriented service, then the agent can proactively ask if the user needs this task-oriented service. As an example, consider the first part of Figure 1, where the agent realizes that the user wants to visit a "college" and actively guides to task-oriented interaction by saying "Do you want to visit some colleges?". It is beneficial for commercial dialogue systems to offer or sell their service (Chiu et al., 2022) at a right moment. This is also possible in a multi-domain scenario. If the agent is aware that user has the possibility to order other services after completing a task request, the agent can continually promote those services. Like in the third part of Figure 1, the user potentially needs the taxi service after booking a restaurant. This is also a good moment to actively offer this service in a commercial setting. When the human-machine interaction starts with a task-oriented request, the users might have the feeling that they are talking to their friends if the system naturally switches to chit-chat interaction after providing all task-related information (see second example in Figure 1). This can highly improve the user interaction experience.

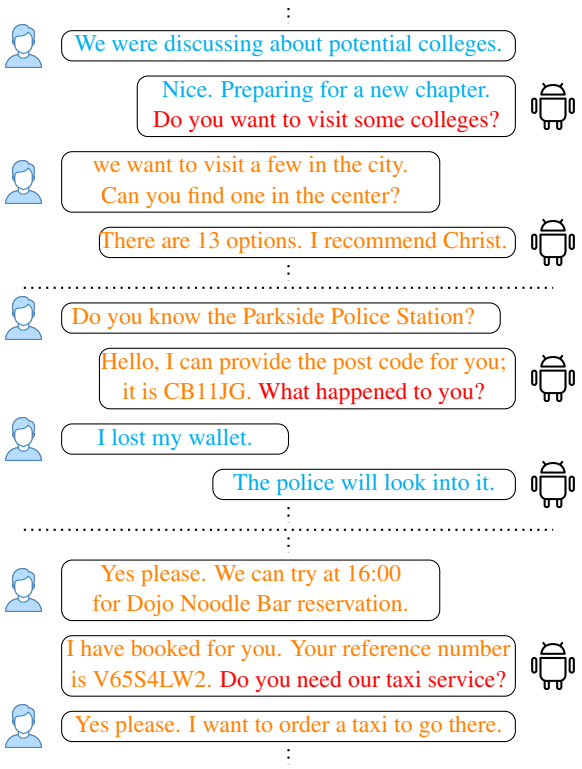


Figure 1: The system-initiated dialogue examples with chit-chat to task-oriented transition, task-oriented to chit-chat transition, and domain transition. The blue parts represent the chit-chat interaction, while the orange parts are task-oriented communication. The newly annotated transition sentences (red parts) demonstrate the system-initiated transitions are controlled by dialogue agent rather than user.

2 Initiative Discussion

(Walker and Whittaker, 1990) describe *initiative* as occasionally "taking the conversational lead".

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

For task-oriented dialogue, the initiative tends to represent "driving the task" (Smith, 1994; Smith and Gordon, 1997). Novick and Sutton (1997) introduce mixed-initiative interaction and describe initiative as a multi-factor concept, which includes choice of task, choice of speaker and choice of outcome. However, the formal definition of the term initiative is still missing from the literature. With the surge in interest in unified (Lin et al., 2021; Zhao et al., 2021; Young et al., 2022) models that can respond to both chit-chat and task-oriented user requests, we explore system-initiated transitions based on a unified model from **three** perspectives as follows (the first two are *dialogue mode* transitions, the third one is a *domain* transition):

- The system-initiated transition from chit-chat to task-oriented as in the first dialogue example shown in Figure 1, where the initiative agent captures the potential task-related information and proactively guides the switch at a proper moment.
- The system-initiated transition from task-oriented to chit-chat as in the second dialogue example shown in Figure 1, in which the system is aware of the completion of a task request and smoothly switches to chit-chat.
- The system-initiated transition from one domain to another in task-oriented interaction, as in the third example shown in Figure 1. Here, the dialogue agent is sensitive to the completion of current task request and proactively switch to another potential task domain.

3 Potential Challenges

Concerning the potential challenges of this work, we have the following questions that need to be precisely discussed and our opinion on these challenges is also elaborated here:

- **When is it a good moment for initiative transition?**[1] If the interaction starts with chit-chat, a good transition moment to switch to task-oriented mode could be when the agent captures some potential task-related information, which could be a task domain, an intent or a slot (such as "college" in the Figure 1). If the interaction starts from task-oriented, a good transition moment could be the completion of the current request, followed by a decision to switch to another domain or to chit-chat interaction.
- **How to guide the generation of a transition sentence?** Transition sentences to chit-chat are hard to control, because they can be diverse and free in style. However, no matter whether it is switching from chit-chat to task-oriented, or from one task domain to another task domain, the system can generate the transition sentence based on relevant information it captures, such as a task domain or a slot.
- **How to evaluate the transition sentence generation?** Firstly, the evaluation on generation tasks is still a challenge in general (Chaganty et al., 2018). Additionally, we argue that the evaluation of transition sentences is even more difficult. One reason is that current publicly available datasets rarely have human annotated transition sentences as a reference to compute automatic metrics, such as BLEU (Papineni et al., 2002) or Meteor (Banerjee and Lavie, 2005). Another reason is that transition sentence generation is different in the three cases mentioned in Section 2, so the evaluation emphasis might be also different.

4 Future Work

In our future work, we will utilize the FusedChat dataset (Young et al., 2022), where human annotated open-domain sentences were prepended and appended to the dialogues of the task-oriented dataset MultiWOZ (Budzianowski et al., 2018; Ye et al., 2021). Hence, every dialogue in FusedChat includes two dialogue modes with the chit-chat and task-oriented parts being interdependent. In addition, many task-oriented dialogues in MultiWOZ include multi-domain interaction. We will first build a unified model that can reply to chit-chat and task-oriented requests with the FusedChat dataset. After that, we will apply the efficient prompt learning (Liu et al., 2021; Li et al., 2022) method to activate the initiative transition of the unified model so it can be sensitive to the timing of transitions and proactively guide them by generating transition sentences.

[1] The question of "when" to switch has also been addressed for pro-activity, which is similar to our initiative switch, by (Nothdurft et al., 2015).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. Salesbot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158.

Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *arXiv preprint arXiv:2202.07371*.

Zhaojiang Lin, Andrea Madotto, Yejin Bang, and Pascale Fung. 2021. The adapter-bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16081–16083.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2015. Finding appropriate interaction strategies for proactive dialogue systems—an open quest. In *Proc. of the 2nd European and the 5th Nordic Symposium on Multimodal Communication 2014*, pages 73–80. LiU Electronic Press.

David G Novick and Stephen Sutton. 1997. What is mixed-initiative interaction. In *Proceedings of the AAAI spring symposium on computational models for mixed initiative interaction*, volume 2, page 12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ronnie W Smith. 1994. Spoken variable initiative dialog: An adaptable natural-language interface. *IEEE Expert*, 9(1):45–50.

Ronnie W Smith and Steven A Gordon. 1997. Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue. *Computational Linguistics*, 23(1):141–168.

Marilyn Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 70–78.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629.

Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu, and Huanhuan Chen. 2021. Unids: A unified dialogue system for chit-chat and task-oriented dialogues. *arXiv e-prints*, pages arXiv–2110.

Assessing the Literal Force Hypothesis in Unconstrained Conversation

Charles Threlkeld and **JP de Ruiter**
Tufts University
{charles.threlkeld, jp.deruiter}@tufts.edu

Abstract

Speech acts are the social acts we perform with our linguistic utterances. Identifying the speech act of an utterance, however, has always been an elusive challenge. But linguistic theory does provide us with the well-defined concept of sentence type. The widely criticized *Literal Force Hypothesis* (LFH) states that the speech act of a sentence can be derived from its sentence type. In this paper, we test the Literal Force Hypothesis in unconstrained conversation. We conclude that while it is far from perfect, there is substantial empirical support for using it as a heuristic.

1 Introduction

Speech acts are critical to understanding language. A speech act describes "the sense in which utterances are not mere meaning-bearers, but rather in a very real sense do things, that is perform actions (Levinson, 2016)." The Literal Force Hypothesis states that (performatives aside) sentences have one-to-one correspondence between sentence type and speech act (Gazdar, 1981). A modern assessment can be found in (Meibauer, 2019).

Sentence types are well-defined linguistic structures, and the three major types listed in the LFH—declaratives, interrogatives, and imperatives—are present in most or all languages (Sadock and Zwicky, 1985). However, the LFH has many detractors. Searle's formulation of speech acts includes a chapter on indirect speech acts — utterances whose function differs from appearance based on the context (Searle, 1975). Levinson (1983) notes that the LFH causes strange semantic and syntactic problems in standard theory. Cummins and de Ruiter (2014) state that utterances have a many-to-many mapping to speech act, a strong refutation of the LFH.

Speech act practice has moved beyond the LFH with detailed dialogue act schemas that move well beyond sentence type, such as DAMSL (Core and Allen, 1997) or the ISO 246172 standard (Bunt et al., 2016). These schemas have furthered the field but are not without their detractors. Traum (2000) examines questions that schemas must answer, and how different answers will provide different lenses for different questions. Bunt et al. (2017) shows that standards must be revised as the science evolves.

Assigning speech acts to utterances in sentences is also fraught. Cordon and Lakoff (1975) suggest re-appraisal if literal interpretations are problematic. Searle (1975) suggests a selection process based on context. Prosody (Shriberg et al., 1998) and dialogue structure (Schegloff and Sacks, 1973; Clark, 1996) also offer clues. Anomalous utterances are detected in planning models for speech act attribution (Cohen and Perrault, 1979; Brenner and Kruijff-Korbayová, 2008; Engesser et al., 2017).

All of this work makes it clear that the LFH is not sufficient for a robust analysis of all language use, but we are not aware of any work directly testing the LFH in open conversation. Other studies look at information-seeking contexts (Beun, 1990), such as the TRAINS corpus (Heeman and Allen, 1995). Indirect speech acts are as high as 50% in these contexts, showing that the LFH is extremely inappropriate. In this work, we seek similar metrics for open conversation to learn if these low numbers of direct speech acts are pervasive or context-dependent.

2 Methods

To test the LFH, we are limiting ourselves to an analysis of sentences of the three major types—declarative, interrogative, and imperative—each of which has well-defined syntactic properties in English (Sadock and Zwicky, 1985). Following

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

Sadock (2012), tag-questions are included as interrogatives.

Following the LFH predictions, we are limiting ourselves to the speech acts listed—statement, question, and request. We recognize that this is an impoverished list, especially compared to the work referenced in the introduction, but this allows for a clean interpretation of the LFH.

We use the *next-turn proof procedure* from Conversation Analysis for labeling speech acts (Hutchby and Wooffitt, 2008). The next-turn proof procedure defines the speech act of an utterance according to how it was received. A question gets an answer (or further interrogation) (Stivers and Rossano, 2010); a request gets fulfilled or rejected (Searle and Sadock, 1976); and a statement lacks either of these qualities. This approach may not be suitable for all speech act work, but it does provide a clean distinction between form and function for labeling sentence type and speech act. This is particularly important for work looking at indirect speech acts, where we need a well-defined notion of directness, which is explicated here by the LFH.

Previous work has shown that utterances may have more than one speech act (Grice, 1975; Bunt, 2011). However, if an interrogative goes unanswered or an imperative ignored, and it is not *marked* in the dialogue, we can only speculate that the direct speech act was intended. Therefore, we have simplified our schema to a forced-choice methodology.

For an open dialogue corpus, we have chosen to use the Conversation Analysis British National Corpus (CABNC) (Albert et al., 2015), which is a set of transcriptions and linked audio recordings of open British conversation. Our intention is to find a representative sample of unconstrained, conversation dialogue—the setting in which language evolved (Enfield and Levinson, 2006). We used the audio recordings in tandem with the transcriptions for annotating, so prosody and intonation were available to the transcribers. We tagged 1002 utterances in eight conversations. Inter-rater reliability was 92% accurate ($\kappa = 0.89$).

3 Results

In our sample, we found that the LFH held for 92% of sentences, substantially higher than previous work in constrained contexts. This may be in part due to our small set of speech acts, but we believe it is in large part due to the unconstrained context of the dialogue studied. In constrained context, the speech act can be inferred regardless of syntactic structure, but we find here that syntactic structure is often a good guide to speech act interpretation in open dialogue.

	Declarative	Interrogative	Imperative	Fragment	Total
Statement	**450**	8	3	0	461
Question	22	**89**	0	1	112
Request	4	3	**15**	2	24
Other	0	8	0	397	405
Total	476	108	18	400	1002

Figure 1: **Sentence type/speech act pairs as raw count of all utterances**. Sentence type is by column and speech act by row. Numbers in bold are predicted by the LFH. Of 602 sentences, 554 speech acts are correctly predicted by the LFH and the remaining 48 are not.

If the high number of direct speech acts found in our sample indicates that indirect speech acts are more likely in certain contexts, we can use our indirect speech act findings as exploratory research what these contexts are likely to be. We found that declarative questions were Labov B-events (Labov and Fanshel, 1977), in which the speaker lacked epistemic authority over the statement. Declarative requests were found in ritualized contexts like shopping or eating. Interrogative statements were often exclamations, tag questions, or rhetorical questions. Interrogative requests were rare, despite their prominence in robotics work (e.g., (Williams et al., 2018)), but their context was similar—sales situations. We also found interrogatives in self-talk which did not fit anywhere in our schema. Finally, we found imperatives used as exhortations like "let's hope so!"

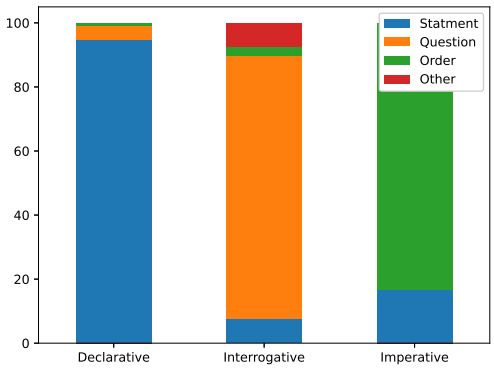


Figure 2: **Speech act as portion of sentence type**. The LFH predicts three solid columns of blue orange and green.

References

Saul Albert, LE De Ruiter, and JP De Ruiter. 2015. Cabnc: the jeffersonian transcription of the spoken british national corpus.

Robbert-Jan Beun. 1990. The recognition of dutch declarative questions. *Journal of Pragmatics*, 14(1):39–56.

Michael Brenner and Ivana Kruijff-Korbayová. 2008. A continual multiagent planning approach to situated dialogue. *Proceedings of LONDIAL-2008*, pages 67–74.

Harry Bunt. 2011. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222–245.

Harry Bunt, Volha Petukhova, and Alex Chengyu Fang. 2017. Revisiting the iso standard for dialogue act annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2016. Dialogue act annotation with the iso 24617-2 standard. *Multimodal Interaction with W3C Standards*, page 109–135.

H.H. Clark. 1996. *Using Language*. Cambridge University Press.

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

David Cordon and George Lakoff. 1975. Conversational postulates. In *Speech acts*, page 83–106. BRILL.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA.

Chris Cummins and Jan P de Ruiter. 2014. Computational approaches to the pragmatics problem. *Language and Linguistics Compass*, 8(4):133–143.

Nicholas J Enfield and Stephen C Levinson. 2006. *Roots of Human Sociality: Culture, Cognition and Interaction*. Berg Publishers, New York, NY.

Thorsten Engesser, Thomas Bolander, Robert Mattmüller, and Bernhard Nebel. 2017. Cooperative epistemic multi-agent planning for implicit coordination. *arXiv preprint arXiv:1703.02196*.

Gerald Gazdar. 1981. Speech act assignment. In A. Joshi, Bruce H. Weber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 64–83. Cambridge University Press.

H.P. Grice. 1975. Logic and convesation. *Syntax and Semantics*, 3:41–58.

Peter A Heeman and James F Allen. 1995. The trains 93 dialogues. Technical report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE.

Ian Hutchby and Robin Wooffitt. 2008. *Conversation Analysis*. Polity.

William Labov and David Fanshel. 1977. *Therapeutic discourse: Psychotherapy as conversation*. Academic Press.

Stephen C. Levinson. 1983. *Pragmatics. Cambridge textbooks in linguistics*. Cambridge University Press, Shaftesbury Road Cambridge CB2 8BS UK.

Stephen C. Levinson. 2016. Speech acts. *Oxford Handbooks Online*.

Jörg Meibauer. 2019. What is an indirect speech act?: Reconsidering the literal force hypothesis. *Pragmatics & Cognition*, 26(1):61–84.

Jerrold M Sadock and Arnold M Zwicky. 1985. Speech act distinctions in syntax. *Language typology and syntactic description*, 1:155–196.

Jerry Sadock. 2012. Formal features of questions. *Questions*, page 103–122.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4).

John R Searle. 1975. Indirect speech acts. In *Speech acts*, page 59–82. BRILL.

John R. Searle and Jerrold M. Sadock. 1976. Toward a linguistic theory of speech acts. *Language*, 52(4):966.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):443–492.

Tanya Stivers and Federico Rossano. 2010. Mobilizing response. *Research on Language and Social Interaction*.

David R Traum. 2000. 20 questions on dialogue act taxonomies. *Journal of semantics*, 17(1):7–30.

Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. 2018. "thank you for sharing that interesting fact!". *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*.

A Appendix

A.1 Tagging Replication

The author tagged the entire corpus, and their tags were used for the analyses shown in this paper. A subset of the data was also tagged by a colleague who is an expert in speech acts. The table shown here shows the agreement between the author and their colleague. From this table, we calculate a 92% agreement and $\kappa = 0.89$.

	Statement	Question	Request	Fragment
Statement	124	1	0	3
Question	5	20	0	0
Request	2	1	5	0

Table 1: This table shows the speech act tagging by the author (rows) and replicator (columns).

ARCIDUCA: Annotating Reference and Coreference In Dialogue Using Conversational Agents in games

Massimo Poesio
Queen Mary University
m.poesio@qmul.ac.uk

Richard Bartle
University of Essex
rabartle@essex.ac.uk

Jon Chamberlain
University of Essex
jchamb@essex.ac.uk

Julian Hough
Queen Mary University
j.hough@qmul.ac.uk

Chris Madge
Queen Mary University
c.j.madge@qmul.ac.uk

Diego Perez-Llebana
Queen Mary University
diego.perez@qmul.ac.uk

Matt Purver
Queen Mary University
m.purver@qmul.ac.uk

Juntao Yu
University of Essex
j.yu@essex.ac.uk

Abstract

The objective of ARCIDUCA is to address the twin challenge of developing conversational agents (CAs) able to deal with coreference and reference, and of creating datasets for training such agents, by having CAs generate through interaction the needed training data, which can then be used to improve those agents as well as train agents for other domains. A core hypothesis of the project is that the most effective way to motivate enough individuals to participate in such interactions is by embedding these interactions in online games-with-a-purpose.

1 Introduction

The development of architectures such as the encoder/decoder model (Sutskever et al., 2014) and the Transformer (Vaswani et al., 2017) has brought about an explosion of interest in neural architectures for conversational agents (CAs) (Vinyals and Le, 2015; Bordes et al., 2017; Zhang et al., 2018; Dinan et al., 2019b; Gao et al., 2019; Ram et al., 2018; Dinan et al., 2019a). CA research has since shifted towards CAs capable of engaging in more complex and task-oriented dialogue such as restaurant booking (Bordes et al., 2017) or question answering (Dhingra et al., 2017) . The results on these tasks show that CAs carrying out more complex tasks require the ability to carry out more in-depth interpretation (Quan et al., 2019; Roller et al., 2020). Achieving this requires, on the one hand, architectures capable of carrying out such aspects of interpretation, typically incorporating models of dialogue memory and representations of task-specific knowledge (Sukhbaatar et al., 2015; Dinan et al., 2019b). On the other end, training such models requires appropriate resources. Recently, a number of datasets have become available for end-to-end training of task-oriented CAs; these include the datasets available through ParlAi,[1] Amazon's MultiDOGO (Peskov et al., 2019) and Facebook's Dialogue Dodecathlon (Shuster et al., 2020). However, none of these datasets is also annotated with information about the semantic and discourse interpretation of utterances required to train modules for these tasks. The objective of ARCIDUCA is to develop conversational agents (CAs) able to deal with coreference and reference, and of creating datasets for training such agents, by having the CAs themselves generate through interaction the needed training data, which can then be used to improve those agents as well as train agents for other domains.

[1] https://parl.ai/docs/tasks.html

2 The approach

Datasets and Architectures for Coreference in Dialogue Coreference is prevalent even in the shortest conversations (Müller, 2008; Quan et al., 2019; Grobol, 2020). However, current neural architectures for conversational agents mostly do not resolve coreference. Such CAs can only react appropriately when generating the correct response does not require understanding coreference. Part of the problem is that despite impressive recent improvements (Lee et al., 2017; Joshi et al., 2019), coreference research is still mostly focused on written text. This research gap is largely due to a lack of resources. Training a coreference resolver on written text and domain-adapting it to dialogue has proven ineffective, as coreference in dialogue involves different phenomena and is more involved than coreference in text (Müller, 2008; Grobol, 2020). But the largest annotated corpus of coreference in dialogue, the TRAINS subset of our own ARRAU corpus (Uryupina et al., 2020), is too small to train a high performance coreference resolver for CAs. One objective of the project is to create more substantial datasets to study the problem. Also, there is a need for CA architectures including specific modules that enable them to interpret coreference. Some such architectures have recently appeared, such as GECOR (Quan et al., 2019), based on a

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

copying architecture that solves coreference as an incomplete utterance restoration task. (Quan et al., 2019) showed that adding a coreference resolver to a task-oriented CA can substantially improve performance. In the project we will experiment with such architectures.

Games with a Purpose Games with a Purpose (GWAPs) (von Ahn, 2006) have emerged as an alternative to traditional micro-task crowdsourcing (Snow et al., 2008). GWAPs, particularly when run over large periods, can collect large amounts of annotations: e.g., our own *Phrase Detectives* (Poesio et al., 2013), designed to collect labels for coreference, accumulated over 5.7 million coreference judgments from more than 60,000 players over the last fifteen years; the third release of the corpus has now 400,000 markables, twice the number of ONTONOTES. But there is a fundamental difference between conversation and written text: the latter is designed to be read by third parties, whereas, e.g., (Clark and Schober, 1989) have shown that overhearers to a conversation only acquire a partial understanding of what was said.

Games and AI In recent years, games have become one of the most widely used platforms to test progress on machine learning-based AI agent theories (Silver et al., 2016). This progress became visible when DeepMind AlphaGo (Silver et al., 2016) mastered the GO game using a combination of Monte Carlo Tree Search and Deep Learning, but progress since has been accelerated through competitions such as General Video Game AI (Perez et al., 2019) and the development of platforms for rapid experimentation such as MALMO (Johnson et al., 2016) or Unity/ML (Juliani et al., 2018).

Collecting conversational data through conversational learning in games The dominant paradigm for CAs training discussed above (pretraining against an annotated corpus, followed by fine-tuning via reinforcement learning through interaction with other agents) is also the approach used in Game AI, which recently led to an exciting synergy between the two areas of AI, whereby Game AI platforms would be used to train conversational agents as well. One example of this synergy is the MALMO project at Microsoft (Johnson et al., 2016), a platform for training agents in Minecraft which was extended to allow training of conversational agents (Allison et al., 2018; Szlam et al., 2019). More recently, Hockenmaier's group developed an extension of MALMO to allow conversational agents to learn to interact, and used the extension to introduce the Minecraft Collaborative Game Task (Narayan-Chen et al., 2019). In parallel with this, Facebook launched project LIGHT (Urbanek et al., 2019)–an open platform for collecting conversations in a very rich textual fantasy game with extensive crowdsourced resources entirely described in natural language. In ARCIDUCA, we aim to train conversational agents able to interpret coreference and reference by embedding them in LIGHT and the Minecraft Collaborative Game.

Collecting judgments through clarification questions The obvious way to enable a CA to acquire information about interpretation is by making it able to ask **clarification questions** (CQs) as to that interpretation (Purver et al., 2003). As far as we know, this has not yet been attempted for coreference, or for CAs. The one proposal along these lines we are aware of (Thomason et al., 2019) was developed to learn grounded reference for robots. What we propose to do is to adopt a similar strategy for improving conversational agents in games' ability to interpret both references and coreference, but also recording these judgments in the form of an annotated corpus.

3 Progress so far

The project officially started in February 2022, but work started beginning of 2021 with the preparation of the CODI-CRAC 2021 shared task on anaphora resolution in dialogue (Khosla et al., 2021), a second edition of which is currently running. One of the outcomes of this work is the creation of the CODI-CRAC corpus of anaphoric reference in dialogue, covering four well-known domains including AMI, LIGHT, PERSUASION and SWITCHBOARD, and is currently the largest such dataset for English. A second outcome of the shared task has been the development of the Universal Anaphora scorer (Yu et al., 2022), currently being revised to make it more suitable to score coreference in dialogue, e.g., by allowing for discontinuous markables Next work was fine-tuning of a coreference resolver for the LIGHT domain and its incorporation in a conversational agent for the LIGHT domain based on the poly-encoder architecture from (Humeau et al., 2020).

Acknowledgements

ARCIDUCA is funded by EPSRC, EP/W001632/1 The creation of the CODI-CRAC corpus was funded in part by the DALI project (ERC project 695662), in part by HITS Heidelberg (Michael Strube), in part by funding from CMU (Carolyn Rose and Lori Levin).

References

Fraser Allison, Ewa Luger, and Katja Hofmann. 2018. How players speak to an intelligent game character using natural language messages. *TDGRA*, 4(2).

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of ICLR*.

Herbert H. Clark and Michael F. Schober. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21:211–232.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proc. of ACL*, pages 484–495. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, JasonWilliams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019a. The second conversational intelligence challenge (convai2). ArXiv preprint arXiv:1902.00098.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proc. of ICLR*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural approaches to Conversational AI*, volume 13 of *Foundations and Trends in Information Retrieval*. Now.

Loïc Grobol. 2020. *Coreference resolution for spoken French*. Ph.D. thesis, Université Sorbonne Nouvelle.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proc. of ICLR*.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The Malmo platform for artificial intelligence experimentation. In *Proc. of IJCAI*, pages 4246–4247.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Arthur Juliani et al. 2018. Unity: A General Platform for Intelligent Agents. *arXiv preprint arXiv:1809.02627*.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The codi-crac 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proc. of the CODI/CRAC Shared Task Workshop*.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proc. of EMNLP*.

M.-C. Müller. 2008. *Fully Automatic Resolution of It, This And That in Unrestricted Multy-Party Dialog*. Ph.D. thesis, Universität Tübingen.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proc. of the 57th Annual Meeting of the ACL*, pages 5405–5415.

Diego Perez, Simon M. Lucas, Raluca D. Gaina, Julian Togelius, Ahmed Khalifa, and Jialin Liu. 2019. *General Video Game AI*. Morgan Claypool.

Denis Peskov, Nancy Clarke, Jason Krone, Brigitta Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proc. of EMNLP*. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: ... *ACM Transactions on Intelligent Interactive Systems*, 3(1).

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse & Dialogue*, pages 235–255. Kluwer.

J. Quan, D. Xiong, B. Webber, and C. Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for ... In *Proc. of EMNLP*, Hong Kong. ACL.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigru. 2018. Conversational AI: The science behind the Alexa Prize. ArXiv abs/1801.03604.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. 2020. Open-domain conversational agents:... ArXiv preprint arXiv:2006.12442.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded cas. In *Proc. of the ACL*. Association for Computational Linguistics.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and others. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Arthur Szlam, Jonathan Gray, Kavya Srinet, Yacine Jernite, Armand Joulin, Gabriel Synnaeve, Douwe Kiela, Haonan Yu, Zhuoyuan Chen, Siddharth Goyal, Demi Guo, Danielle Rothermel, C. Lawrence Zitnick, and Jason Weston. 2019. Why build an assistant in minecraft? ArXiv: 1907.09273.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. 2019. Improving grounded natural language understanding through human-robot dialog. In *Proc. of ICRA*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, Samuel Humeau, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. ArXiv preprint arXiv:1903.03094.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. ArXiv:1706.03762.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the Deep Learning Workshop at ICLR*, Lille.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Juntao Yu, Nafise Moosavi, Silviu Paun, Sopan Khosla, Sameer Pradhan, and Massimo Poesio. 2022. The universal anaphora scorer 1.0. In *Proc. of LREC*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Dialogue Policies for Confusion Mitigation in Situated HRI

Na Li, Robert Ross
School of Computer Science
Technological University Dublin
{na.li, robert.ross }@tudublin.ie

Abstract

Confusion is a mental state triggered by cognitive disequilibrium that can occur in many types of task-oriented interaction, including Human-Robot Interaction (HRI). People may become confused while interacting with robots due to communicative or even task-centred challenges. To build a smooth and engaging HRI, it is insufficient for an agent to simply detect confusion; instead, the system should aim to mitigate the situation. In light of this, in this paper, we present our approach to a linguistic design of dialogue policies to build a dialogue framework to alleviate interlocutor confusion. We also outline our sketch and discuss challenges with respect to its operationalisation.

1 Introduction

Confusion is a type of dynamic mental state, which can not only lead to negative conditions, *i.e.*, frustration, boredom or subsequent disengagement in a task or a conversation, but can also be associated with positive conditions as a user seeks to overcome initial confusion (D'Mello et al., 2014; Li et al., 2021). In mainstream human-computer interaction (HCI) studies, a number of studies have investigated confusion state effects in the context of online learning and driver assistance (Kumar et al., 2019; Grafsgaard et al., 2011; Zhou et al., 2019). One prominent model of confusion is from Lodge et al. (2018) who pointed to a zone of optimal confusion (ZOC) which is productive confusion, where learners are self-motivated to overcome their confusion state; but also pointed to a zone of sub-optimal confusion (ZOSOC) where learners could not resolve the disequilibrium which in turn leads to confusion persisting such that the confusion becomes unproductive. Similarly, D'Mello et al. (2014) described three bi-directional transitions, *i.e.* confusion-engagement, confusion-frustration and frustration-boredom transitions to explain confusion dynamics. Finally, Arguel and Lane (2015) presented two thresholds (T_a and T_b) bounding levels of confusion potential in learning. Between the two thresholds is the confusion stage, and if the level of confusion is less than T_a, then the learners should be fully engaged, whereas if the confusion level is over T_b the confusion is not mitigated leading to learners becoming bored.

However, little work has focused on confusion detection and modelling in general conversational interactions or human-robot interaction (HRI). Given this gap, in our research, we aim to detect, model, and in time mitigate confusion states (*i.e.* productive confusion, unproductive confusion). For this work, we focus on four confusion induction types, *i.e.*, complex information, contradictory information, insufficient information, and false feedback (Lehman et al., 2012, 2013; Silvia, 2010).

Although our work to date has focused on confusion (Li et al., 2021; Li and Ross, 2022), modelling and detection, it is also essential that the dialogue agent is capable of mitigating user confusion and helping participants reengage in the ongoing task-oriented interaction. Our model is based on seven dialogue act types that are used to implement strategies for confusion mitigation. In light of this need, in the paper, we sketch out our initial approach to design a dialogue policy for task-oriented interaction that can be used to mitigate users confusion states if identified. The model consists of a general dialogue policy and two specific policies for different confusion induction situations. While HRI includes verbal and nonverbal interactions (Bartneck et al., 2020), in this initial work, our outline dialogue policies are restricted to linguistic interactions.

2 Act and Policy Outline

As the basis of the policy combining the specific case study of confusion mitigation, we first outline a sort of dialogue act types corresponding to a

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

general dialogue policy, and then two sub-policies for two confusion states mitigation are produced. Therefore, we start by introducing the following seven key dialogue act types and highlight their relevance to the mitigation as follows:

1. **Restatement**: The agent repeats the information or question.

2. **Feedback request**: The agent asks for the participant's feedback and response.

3. **Information extension**: The agent provides more information to expand on the information or question already raised.

4. **Information supplement**: The agent provides comprehensive information or questions in different ways for participants to quickly understand easily.

5. **Response correction**: The agent provides the appropriate response in order to avoid confusion states on the participant.

6. **Confirmation**: The agent admits that the information or question has one or more issues leading to the participant being confused.

7. **Subject change**: The agent changes straightforward questions or other topics.

We applied the seven types of dialogue act to first design a general dialogue policy based on a number of communicative rules (see Table 1). Figure 1 illustrates the operating dialogue policy as a control flow process, with each step corresponding to one of the detailed elements of the outline rules in Table 1. In this control flow policy, each step makes it possible to help users who are confused transfer to a non-confusion state. If after any one step, the user's confusion still cannot be mitigated, then the agent will move to the next step.

Based on this general framework policy, we have developed a set of sub-policies to apply in the specific cases of productive and unproductive confusion in the case of the four confusion induction types mentioned earlier. The first of these dialogue sub-policies (see Table 2) includes the dialogue act types and corresponding communication rules to reduce productive confusion according to the induction of a specific confusion method. The second sub-policy (see Table 3) addresses the case where the participant has reached an unproductive

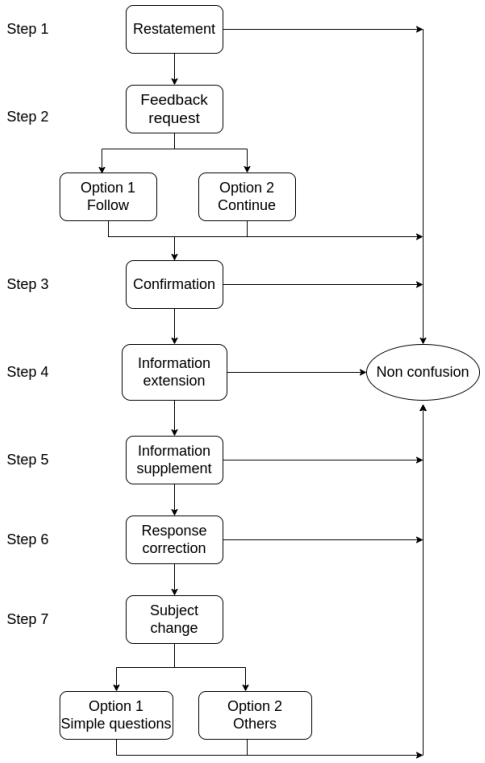


Figure 1: General policy process of confusion mitigation

confusion state, where they may be frustrated or even want to drop the conversation. Therefore, this sub-policy helps the participant reengages in interacting with the agent from their unproductive confusion state. The three detail policies in Table 1, Table 2 and Table 3 are mentioned early, *i.e.* general dialogue policy, and two sub-policies for mitigating productive and unproductive confusion are attached to GitHub [1].

3 Discussion & Outlook

Although this short paper simply provides a sketch of our approach, we are building on this sketch to implement a physical test for those policies based on a wizard-of-oz study (Riek, 2012) using physical situated robots integrating our existing platform. We expect that this work can drive a true formalisation and evaluation of these policies. Therefore, our goal is to fully operationalise this policy, but this, of course, is non-trivial. While we could aim to formalise this model through an appropriate formalisation, such as type theory with records (TTR), a Machine Learning (ML) driven approach would be more suitable for a robust system construction. Ultimately, our goal is to develop a hybrid policy

[1] Table 1, 2, 3: https://github.com/lindalibjchn/dialoguepolicy.git

that can have general structures to accommodate the user state, but is driven by a probabilistic framework.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Amaël Arguel and Rod Lane. 2015. Fostering deep understanding in geography by inducing and managing confusion: An online learning approach. *ASCILITE 2015 - Australasian Society for Computers in Learning and Tertiary Education, Conference Proceedings*, (November):374–378.

Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *References*. Cambridge University Press.

Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170.

Joseph F Grafsgaard, Kristy Elizabeth Boyer, and James C Lester. 2011. Predicting Facial Indicators of Confusion with Hidden Markov Models. Technical report.

Harsh Kumar, Mayank Sethia, Himanshu Thakur, Ishita Agrawal, and Swarnalatha P. 2019. Electroencephalogram with Machine Learning for Estimation of Mental Confusion Level. *International Journal of Engineering and Advanced Technology*, 9(2):761–765.

Blair Lehman, Sidney D'Mello, and Art Graesser. 2012. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, 15(3):184–194. Emotions in online learning environments.

Blair A. Lehman, Sidney K. D'Mello, and Arthur C. Graesser. 2013. Who benefits from confusion induction during learning? an individual differences cluster analysis. In *AIED*.

Na Li, John D Kelleher, and Robert Ross. 2021. Detecting interlocutor confusion in situated human-avatar dialogue: A pilot study. In *25th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021) University of Potsdam, Germany*.

Na Li and Robert Ross. 2022. Transferring studies across embodiments: A case study in confusion detection. In *1st workshop (MMAI2022) that is a part of the conference on Hybrid Human-Artificial Intelligence 2022, Amsterdam, Netherlands*.

Jason M. Lodge, Gregor Kennedy, Lori Lockyer, Amael Arguel, and Mariya Pachman. 2018. Understanding Difficulties and Resulting Confusion in Learning: An Integrative Review. *Frontiers in Education*, 3.

L. Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. In *HRI 2012*.

P. Silvia. 2010. Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, 4:75–80.

Yun Zhou, Tao Xu, Shaoqi Li, and Ruifeng Shi. 2019. Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game. *Universal Access in the Information Society*, 18(3):551–563.

Interactivism in Spoken Dialogue Systems

Teresa Rodríguez Muñoz[*], **Emily Ip**[*], **Guanyu Huang**[*], and **Roger K. Moore**[*]

[*]Department of Computer Science, The University of Sheffield
[*]{trodriguezmunoz1, eyjip1, ghuang10, r.k.moore}@sheffield.ac.uk

Abstract

The interactivism model introduces a dynamic approach to language, communication and cognition. In this work, we explore this fundamental theory in the context of dialogue modelling for spoken dialogue systems (SDS). To extend such a theoretical framework, we present a set of design principles which adhere to central psycholinguistic and communication theories to achieve interactivism in SDS. From these, key ideas are linked to constitute the basis of our proposed design principles.

Keywords: Spoken Dialogue System, interactivism, incremental dialogue, transactional model.

1 Introduction

In recent years, with the exponential growth of speech technologies such as Siri and Alexa, users have grown accustomed to the rigid dialogue schemes these devices offer. Thus, current human-robot interactions (HRI) are far from being conversational (Moore et al., 2016). To optimise the effectiveness of HRI dialogues, researchers have worked on the accuracy of Automatic Speech Recognition (ASR), the naturalness of Text-to-Speech (TTS) modules and alternative dialogue frameworks. Incremental dialogue systems are one such alternative to attaining natural timing in conversation (Schlangen and Skantze, 2011).

However, the quality of spoken interactions goes beyond increasing ASR accuracy and delivering timely responses. One must also have a better understanding of dialogue as a process, which is not linear, but rather, transactional (Pierce and Corey, 2009). Moreover, dialogue involves continuous, bi-directional interactions between the conversational agent, the contextual environment and the interlocutor via verbal and non-verbal signals (Moore, 2016). Hence, in this work, we suggest that high-performing SDS must embrace interactivism to demonstrate situational and social awareness.

2 Background Theories

2.1 Interactivism: Dialogue as a Process

Interactivism has favoured frameworks of process over models of substance (Bickhard, 2009). This idea may translate to dialogue modelling, as dialogue is the process whereby ideas are exchanged among multiple social actors. It is also suggested that dialogue modelling is only one part of the interconnected modular system of an 'intelligent' agent, which coexists in a given environment (Maturana and Varela, 1987). Thus, the system behaves in an enactivist manner as a continuously and autonomously self-producing autopoietic entity (Moore, 2016). Therefore, future SDS design efforts should include *autopoiesis* (Maturana and Varela, 1987) in the form of self-monitoring of the system's output (i.e., utterances produced by the agent) and its own current status. This has been attempted in incremental dialogue frameworks (Skantze and Schlangen, 2009; Schlangen and Skantze, 2011), where self-monitoring feedback loops between the Contextualiser and Dialogue Modelling modules are employed to self-monitor, self-repair and monitor the user's speech production and non-verbal feedback signals.

Consequentially, the proposed interactivist framework presents implications for language and its use in human interaction. The framework is inherently social and interactional. Thus, the conversational agent needs to be imbued with the knowledge of these conventions to respond appropriately. Furthermore, future SDS cannot exclusively take in human responses in isolation; they must establish an awareness of the environment as it evolves through such an interaction and makes changes to itself accordingly (Moore, 2016).

2.2 Transactional Model of Communication

As an interactive process, spoken dialogue could be seen as transactional (Pierce and Corey, 2009).

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

In comparison with linear and interactive dialogue models, the transactional model is the most dynamic. It considers dialogue as a cooperative process in which interlocutors exchange messages simultaneously. The dialogue is built upon shared experiences in culture, language and/or environment, allowing one to use less speech or even a single sound to achieve a successful interaction (Hawkins, 2003). We propose that the transactional model may be the most preferred paradigm to achieve incremental dialogue. The conversational agent needs to be attentive and adaptive, which requires it to be able to adjust its language behaviours according to changes in the user and the environment.

How can we make this adjustment happen? The inner workings of the brain reveal that thought is not linear; it is a process in which we constantly produce and shape ideas (Clark, 2014). While the brain receives information from external resources, it tries to piece things together in a bottom-up way. It also attempts to guess incoming sensory data on the basis of what it knows about how the world is likely to be in a top-down manner. This Predictive Processing allows the agent to anticipate which actions to take, and to adjust its prediction upon its perception of embodied and environmental information (Clark, 2015). Hence, a predictive function should be employed to enrich SDS design.

2.3 Adjustment of Dialogue Behaviours

In tandem with these *intrapersonal* adjustments within the system, *interpersonal* dynamics evolve as well. Entrainment describes how interlocutors become more similar to each other in their speech throughout a conversation (Levitan, 2013), as speakers' conversational behaviour tends to be influenced by that of the other interlocutor. For this reason, effort-based models (Lindblom, 1990; Moore and Nicolao, 2017) have been developed to account for humans' regulatory behaviour in everyday speech. These models consider what the speaker and the listener(s) share in common to adjust that effort. This closely relates to the Theory of Mind, which involves the ability to discern the mental states, including emotions, knowledge and beliefs of oneself and others (Woodruff and Premack, 1978). A conversational agent with such capabilities would be able to adjust to the user and build a rapport. This type of closed-loop dialogue system facilitates adaptive behaviours which are emergent, and not choreographed.

Furthermore, Gricean pragmatics can be considered to achieve more interactive behaviours in SDS. To more closely emulate human dialogue, Grice's maxims dictate that one should be as informative, concise, and relevant to the discussion as possible (Grice, 1975). Moreover, ostensive communication (i.e., communication involving the expression and recognition of intentions via verbal or non-verbal cues) should be further explored in SDS design (Scott-Phillips, 2017). This would allow the conversational agent to be aware of the multimodal protocols to open and close channels of communication and engagement.

3 Design Principles for Transactional SDS

Based on the theoretical background discussed, several design principles have been suggested below:

(i) **Conversational agents must have an incremental dialogue framework**: incremental SDS employ self-monitoring feedback loops to perform revisions on the system's output (either covertly or overtly) and determine whether an utterance was spoken, interrupted by the user or revoked as a failed hypothesis.

(ii) **Agents should adjust their communicative effort in dialogue**: if the conversational agent can identify the user's abilities and adjust itself, then it could achieve autonomous, progressive learning of a user incrementally.

(iii) **Conversational agents must be aware of the context of the conversation and have memory of past interactions with a user**. Short- and long-term information needs are essential in natural conversation.

(iv) Further design of **multi-party SDS should consider ostensive behaviour** when engaging with different social actors.

4 Conclusion and Future Work

This work has briefly discussed interactivism and other central psycholinguistic and communication theories to improve the performance of current SDS. We have identified key design principles that the community may employ to design novel conversational agents, founded on the interactivist theoretical framework. Future work will involve the study of incremental dialogue systems and adapting them to align with the principles identified in this work.

Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [EP/S023062/1]. Moreover, this work was partially supported by The Fulbright University of Sheffield Postgraduate Award.

References

Mark H Bickhard. 2009. Interactivism: A manifesto. *New Ideas in Psychology*, 27(1):85–95.

Andy Clark. 2014. Perceiving as predicting. *Perception and its modalities*, pages 23–43.

Andy Clark. 2015. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Sarah Hawkins. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of phonetics*, 31(3-4):373–405.

Rivka Levitan. 2013. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 84–90.

Björn Lindblom. 1990. Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling*, pages 403–439. Springer.

Humberto R Maturana and Francisco J Varela. 1987. *The tree of knowledge: The biological roots of human understanding*. New Science Library/Shambhala Publications.

Roger K Moore. 2016. Introducing a pictographic language for envisioning a rich variety of enactive systems with different degrees of complexity. *International Journal of Advanced Robotic Systems*, 13(2):74.

Roger K Moore, Hui Li, and Shih-Hao Liao. 2016. Progress and prospects for spoken language technology: What ordinary people think. In *INTERSPEECH*, pages 3007–3011. San Francisco, CA.

Roger K Moore and Mauro Nicolao. 2017. Toward a needs-based architecture for 'intelligent'communicative agents: Speaking with intention. *Frontiers in Robotics and AI*, 4:66.

Teresa Pierce and Amy M Corey. 2009. *The evolution of human communication: From theory to practice*. EtrePress.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

Thomas C Scott-Phillips. 2017. Pragmatics and the aims of language evolution. *Psychonomic Bulletin & Review*, 24(1):186–189.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 745–753.

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

Contingency in Child-Caregiver Naturalistic Conversation: Evidence for Mutual Influence

Charlie Hallart*
Aix-Marseille University
charlisahallart@gmail.com

Morgane Peirolo*
Aix-Marseille University
morgane.peirolo@gmail.com

Zihan Xu
National University of Singapore
zhxu98@gmail.com

Abdellah Fourtassi
Aix-Marseille University
abdellah.fourtassi@gmail.com

Abstract

To be able to hold conversations, children need to learn contingency, i.e., the ability to contribute to a dialog with relevant utterances. We study this skill in the context of child-caregiver naturalistic interactions. While much of previous work has focused on the caregiver or on the child, here we study contingency in the dyad as a whole, allowing for a deeper understanding of how both children and caregivers influence the course of the dialog.

1 Introduction

How do children learn to become competent conversational partners? The current study focuses on the development of one skill that is at the core of the very definition of conversation: *Contingency*, which we can be understood, in broad terms, as the ability of children to contribute with utterances that connect with the interlocutor's previous turn and with the topic of the ongoing exchange more generally, allowing for a coherent back and forth between the interlocutors (e.g., Slomkowski and Dunn, 1996).

Previous related work has either focused the caregiver's contingency (with respect to the child's behavior/utterance) (see review in Masek et al., 2021) or on the child's contingency (Bloom et al., 1976; Hale and Tager-Flusberg, 2005; Nadig et al., 2010; Pagmar et al., 2022). The novelty of the current work is that it studies the development of early dialog contingency by investigating how both the child's and caregiver's contingent behaviors (or lack thereof) influence each other in naturalistic interactions.

2 Method

2.1 Data

We used data from the French "Paris Corpus" (Morgenstern and Parisse, 2012), publicly available[1] on CHILDES repository (MacWhinney, 2000). The corpus is made of longitudinal recordings (and their transcriptions) of children spontaneously interacting with their caregivers at home. The participants were videotaped (by a researcher) once a month, over a developmental period ranging from 1 to 5 years of age. Based on the quality of the recordings, we studied the data of two female children (Anae and Madeleine) and two males (Adrien and Theophile). We sampled, for each child, 6 transcripts. We made sure these picked transcripts spanned the entire developmental range of the corpus. We ended up with a total of 24 transcripts, each lasting around 1 hour.

2.2 Coding

Question, Response, Follow-up (QRF)

We focus on parts of the dialog that are initiated with a question. The reason is that questions are frequent in child-caregiver dialogues, making the Question, Response, Follow-up sequence (hereafter QRF) a rather time-stable micro-structure, within which we can study children's contingency starting from young age (Chouinard et al., 2007). Besides, researchers have suggested that questions are a way caregivers initiate children to the exercise of contingency (Foster, 1986). Our data yielded a total of 402 child-initiated QRF units and 2,815 caregiver-initiated QRF units (across all 24 transcripts).

Contingency coding

The coding proceeded in two steps. First, we coded the sequences using a fine-grained coding scheme based on the literature on child-initiated QRF (e.g., Kurkul and Corriveau, 2018), while introducing slight adjustments to capture *both* child- and caregiver-initiated QRFs. Inter-annotation agreement based on a sample of about 20% of the data, coded independently by two annotators, led to Cohen's kappa values of 0.8 for child-initiated QRF

[1]The corpus link: https://phonbank.talkbank.org/access/French/Paris.html

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

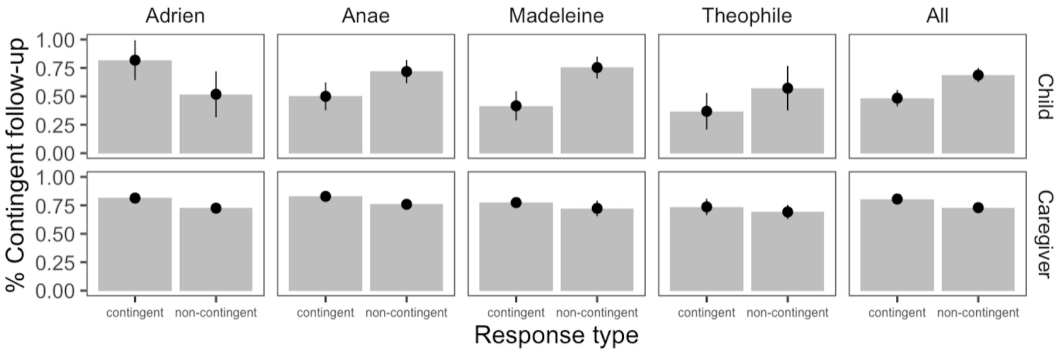


Figure 1: Percent of contingent follow-ups per response contingency status (summed over all ages and transcripts for maximal statistical power). Dots and ranges represent the means and % 95 confidence intervals.

data and 0.9 for caregiver-initiated QRF data, both reflecting "strong" agreement (McHugh, 2012).

Second, we classified the fine-grained categories given for responses and follow-ups into contingent vs. non-contingent, as follows. A response was considered non-contingent if it was classified by the annotators as: no answer given, irrelevant, unsatisfactory, or unintelligible. As for the follow-up, we chose to judge its contingency with respect to the question asked, not with respect to the response. It was considered non-contingent if it was classified by the annotators as: no follow-up given, changing the topic of the question, or ambiguous such as when the follow-up is not explicitly communicative (e.g., laugh) or does not add specific informational content (e.g., 'hum').

3 Results and Discussion

To investigate how caregivers' response contingency (or lack thereof) influences the child's follow-up (and vice-versa), we compared follow-ups after contingent vs. non-contingent responses. The results are shown in Figure 1.

For children's (top row), we found that *more* contingent follow-ups were given following *non-contingent* responses from caregivers.[2] This effect was consistent among all children except for one. This finding suggests that children expect their questions to elicit responses and they expect these responses to be contingent. When this is not the case, i.e., when the caregiver's response is not contingent, children are more likely to follow up contingently, mostly by suggesting an answer to their own question (34% of total follow-ups vs. only 15% in the contingent case) or by persisting via re-asking the same question (11.2% of their total follow-ups vs. only 3.2% in the contingent case) (See also Frazier et al., 2009).

For adults (bottom row), we found — interestingly — the opposite pattern: More contingent follow-ups were given following *contingent* responses from children.[3] This pattern was consistent among all caregivers. It reflects the fact that caregivers are adapting to the children's responses, often with the purpose of keeping the conversation alive. Indeed, when the child's response is contingent, they follow up more on their original question to extend the exchange and/or provide expressions of agreement ("yes, that's right" repeating the child's utterance, etc.). When the child's response is not contingent, their slightly lower contingent follow-ups indicate that they do not necessarily persist, as children do, by bringing the conversation back to the original question (although they often do; given that the percentage is still quite high). However, they seem to also be happy to switch to the child's new focus of attention or initiate a new, perhaps, more engaging topic of discussion.

Limitations and Future work

This paper investigated an aspect of mutual influence in child-caregiver conversations. The limitation, however, is that hand annotation allowed for the study of only a small sample of children. Besides, the annotation relied primarily on verbal data. In future research, we will extend this work both via automatic labeling to test the scalability of the findings (e.g., Cervone and Riccardi, 2020; Nikolaus et al., 2021) and via using corpora that allow for the study of multimodal signaling (e.g., Bodur et al., 2021, 2022).

[2]We verified this observation statistically by fitting a mixed-effects logistic regression. The numbers are not shown due to space constraints.

[3]We verified this observation statistically by fitting a mixed-effects logistic regression.

Acknowledgments

The authors of this work have been supported by funding from the Institute of Language Communication and the Brain (ANR-16-CONV-0002) and the MACOMIC project (ANR-21-CE28-0005-01).

References

Lois Bloom, Lorraine Rocissano, and Lois Hood. 1976. Adult-child discourse: Developmental interaction between information processing and linguistic knowledge. *Cognitive Psychology*, 8(4):521–552.

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. Chico: A multimodal corpus for the study of child conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 158–163.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2022. Backchannel behavior in child-caregiver video calls. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this Dialogue Coherent? Learning from Dialogue Acts and Entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174, 1st virtual meeting. Association for Computational Linguistics.

Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. 2007. Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, pages i–129.

Susan H Foster. 1986. Learning discourse topic management in the preschool years. *Journal of Child Language*, 13(2):231–250.

Brandy N Frazier, Susan A Gelman, and Henry M Wellman. 2009. Preschoolers' search for explanatory information within adult–child conversation. *Child development*, 80(6):1592–1611.

Courtney M Hale and Helen Tager-Flusberg. 2005. Social communication in children with autism: The relationship between theory of mind and discourse development. *Autism*, 9(2):157–178.

Katelyn E Kurkul and Kathleen H Corriveau. 2018. Question, explanation, follow-up: A mechanism for learning from others? *Child Development*, 89(1):280–294.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Lillian R Masek, Brianna TM McMillan, Sarah J Paterson, Catherine S Tamis-LeMonda, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. 2021. Where language meets attention: How contingent interactions promote learning. *Developmental Review*, 60:100961.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Aliyah Morgenstern and Christophe Parisse. 2012. The paris corpus. *Journal of French language studies*, 22(1):7–12.

Aparna Nadig, Iris Lee, Leher Singh, Kyle Bosshart, and Sally Ozonoff. 2010. How does the topic of conversation affect verbal exchange and eye gaze? a comparison between typical development and high-functioning autism. *Neuropsychologia*, 48(9):2730–2739.

Mitja Nikolaus, Juliette Maes, Jeremy Auguste, Laurent Prevot, and Abdellah Fourtassi. 2021. Large-scale study of speech acts' development using automatic labelling. In *Proceedings of the 43rd annual meeting of the cognitive science society*.

David Pagmar, Kirsten Abbot-Smith, and Danielle Matthews. 2022. Predictors of children's conversational contingency. *Language Development Research*, 2(1):139–179.

Cheryl Slomkowski and Judy Dunn. 1996. Young children's understanding of other people's beliefs and feelings and their connected communication with friends. *Developmental psychology*, 32(3):442.

An Approach to Model Self-imposed Filter Bubbles

Annalena Aicher and **Wolfgang Minker**
Institute for Communication Engineering
Ulm University, Germany
annalena.aicher@uni-ulm.de

Stefan Ultes
Mercedes Research & Development
Sindelfingen, Germany

Abstract

. When people are confronted with an overwhelming amount of information, they tend to filter out all the parts of the available information that do not fit their existing beliefs or opinions. Within this paper, we propose the first model to describe this "self-imposed filter bubble" (SFB) during argumentative information seeking. Based upon this model, argumentative dialogue systems (ADS) shall be able to learn and adapt their dialogue strategy to overcome this SFB in cooperation with the user.

1 Introduction

Especially when searching for information online, users tend to select claims that adhere to their beliefs and to ignore dissenting information, which coins the terms self-imposed filter bubbles (SFB) (Ekström, 2021) and echo chambers (Quattrociocchi et al., 2016). These phenomena belong to the generic term *confirmation bias* which is typically used in psychological literature (Nickerson, 1998). Allahverdyan and Galstyan (2014) describe confirmation bias as the tendency to acquire or evaluate new information in a way that is consistent with one's preexisting beliefs.

To resolve the confirmation bias of a user in decision making processes Huang et al. (2012) propose the usage of computer-mediated counter-argument. Furthermore, Schwind and Buder (2012) regard preference-inconsistent recommendations as a promising approach to trigger critical thinking. Still, if too many counter-arguments are introduced this could lead to unwanted effects negative emotional consequences (annoyance, confusion) (Huang et al., 2012). Consequently, Huang et al. (2012) stress the need for an intelligent system which is able to adapt the frequency, timing and choice of the counter-arguments. To provide such a system, it is crucial to develop and find a model, which can be adapted to the user. The goal of this paper is to present such an abstract model for a user's individual self-imposed filter bubble. It is based on our previous work (Aicher et al., 2022) and consists of the four dimensions *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)*, and *False Knowledge (FK)* and makes it possible to assess the probability of a user being caught in a self-imposed filter bubble with regard to a certain topic. To the best of our knowledge, our approach is the first existing model of a user's SFB which is furthermore suitable to be implemented in an argumentative dialogue system.

2 Self-imposed Filter Bubble Model

As previously mentioned we focus on four dimensions in our model[1]. Their choice is examined in detail in (Aicher et al., 2022) and builds upon findings in well-established state-of-the-art literature (Petty et al., 2009).

2.1 SFB-Model Dimensions

The **RUE** describes the critical-thinking and open-mindedness demonstrated by the user. It takes into account the polarity and number of arguments he/she has heard. This can be mapped onto two actions of the user by asking for more information, either on the pro or con side of the topic of the discussion. Thus, it can be interpreted as a weighting how balanced the user is exploring the topic. Due to the limited scope of this paper we refer to our previous work (Aicher et al., 2021) where its calculation is described in detail.

The **PR** refers to the user's individual assessment of how relevant a subtopic is with regard to the topic of the discussion. We assume that the bigger the PR of a certain subtopic is, the higher is the user's interest and motivation to explore arguments belonging to it.

The **TK** serves as a measure for the information gain and is defined as the new information the user

[1]Please note, that we do not claim the dimensions or our model to be complete but a first approach to model SFBs.

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

is provided with by talking to the system. It can be determined by comparing the total information provided by the system and the information, which is already known to the user. We aim for the user to explore as much information as possible, as this increases the chance to explore other aspects and viewpoints. Thus, the bigger the TK of the users, the more unlikely they find themselves in an SFB.

The **FK** describes the incorrect information a user has on a certain topic[2]. If the user is misinformed on certain aspects, it increases the probability of being stuck in an SFB and reluctant towards contradicting information and viewpoints.

2.2 SFB-Model

Using the dimensions in Subsection 2.1, we define an SFB-vector $\overrightarrow{SFB}$. It has its origin in the origin of the coordinate system and its end is the position of the user in the four-dimensional space at the current state of the interaction.

$$\overrightarrow{SFB} = (PR, RUE, TK, FK)^T. \quad (1)$$

The SFB is described by a four-dimensional body

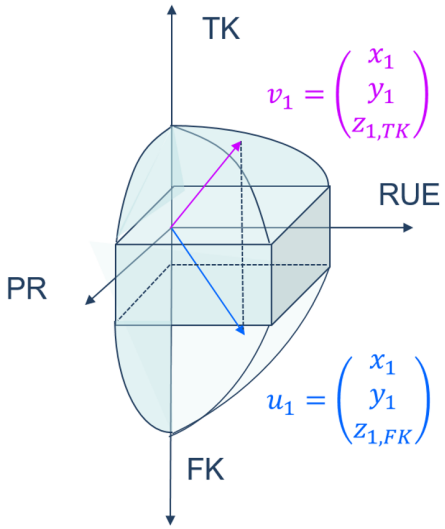


Figure 1: Schematic sketch of an SFB-vector and SFB. For better illustration the four-dimensional SFB-vector is displayed in two split components only differing in their z_1 component. Whereas the blue vector displays TK in the z_1-component, the violet one displays FK. The x_1 component depicts RUE and y_1 PR. The blue filled areas denote the SFB.

describing the probability with which users find themselves within an SFB. Obviously, it is very difficult to determine an exact limit up to which point users are still in their SFB and from which point on-wards they no longer are. The smaller the SFB-vector, the higher the probability that the user is inside the SFB. The longer the SFB vector and the more it extends beyond the SFB, the lower the probability that the user is within the SFB. In Figure 1 an exemplary sketch of this vector and the respective SFB are shown. As a four-dimensional vector cannot be displayed, for better illustration, it was split in two different z_1-components TK and FK. Please note that this sketch is for illustrative purposes only and it is very difficult to determine the "real" shape of the SFB. Therefore the light blue coloured areas indicate a high probability of being inside the bubble, while the non-coloured areas indicate a low probability, without defining the exact boundary of the bubble. To detect and "break" the user's SFB in an ongoing interaction, the model can be adapted dynamically during the interaction. To estimate the success of breaking the SFB the position of the initial (before the interaction) and final (after the interaction) SFB-vector with respect to the SFB are considered.

3 Conclusion and Future Work

In this work, we introduced a novel model for a user's self-imposed filter bubble, consisting of four dimensions: *Reflective User Engagement, Personal Relevance, True Knowledge* and *False Knowledge* (but not limited thereto). To the best of our knowledge this model represents the first approach to estimate the probability that users find themselves within an SFB. To break the user's SFB it is important not to force new information onto the user but to find a more subtle way to weave in information that is not requested (Huang et al., 2012). Our SFB model shall help to identify suitable points of reference (e.g. the most decisive dimensions strengthening the bubble) which can be used as starting point to break the user's SFB in an engaging cooperative argumentative dialogue. In future work, our model will be implemented in a suitable (cooperative) ADS and evaluated in a user study. Therefore, we will investigate how the change and behaviour of each dimension can be tracked in detail during an ongoing interaction using explicit and implicit methods. Furthermore, other potential dimensions shall be explored, such as user trust, communication styles and a virtual agent interface.

[2]Without loss of generality, the information in the system's database is assumed to be correct and consequently, information contradicting the former to be incorrect.

References

Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2021. Determination of reflective user engagement in argumentative dialogue systems.

Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2022. Towards modelling self-imposed filter bubbles in argumentative dialogue systems. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 4126–4134.

Armen E Allahverdyan and Aram Galstyan. 2014. Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557.

Axel Ekström. 2021. The self-imposed filter bubble hypothesis. Student Paper.

Hsieh-Hong Huang, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku. 2012. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, 53(3):438–447.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Richard E Petty, Pablo Briñol, and Joseph R Priester. 2009. Mass media attitude change: Implications of the elaboration likelihood model of persuasion. In *Media effects*, pages 141–180. Routledge.

Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on facebook. *Available at SSRN 2795110*.

Christina Schwind and Jürgen Buder. 2012. Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior*, 28(6):2280–2290.

Mutual gaze detection and estimation: towards human-robot interaction

Vidya Somashekarappa, Christine Howes and Asad Sayeed
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{vidya.somashekarappa,christine.howes,asad.sayeed}@gu.se

Abstract

Mutual gaze is an important part of social interaction and the perception of others emotional states and essential for establishing joint attention. It has been described as "the most powerful mode of developing a communicative link between humans". Although gaze estimation and prediction using machine learning and computer vision is extensively studied in images and videos for automation, there is a lack of research in dialogue and interaction. In this study, we estimate gaze using a PyTorch-based model with the goal of implementing human-like mutual gaze in a robot.

1 Introduction

Eye gaze supports and augments other social behaviours such as speech and gesture, and mental states or cognitive effort can substantially influence gaze. Since speech is a dominant mode of communication in human interactions, it is not feasible to separate gaze from speech in face-to-face human-human dialogue, and we should not do so for human-robot dialogue either. Researchers have shown that gaze improves speech-based interactions, e.g., disambiguating object references, maintaining engagement, conversation and narration, guiding attention, managing partners, influencing turn-taking (Kaiser et al., 2003; Rapp et al., 2021; Somashekarappa et al., 2021)

1.1 Mutual Gaze in Human Interaction

Mutual gaze occurs from birth when infants gaze at their caregivers. The field of vision of the newborns is approximately the distance required to make eye contact when held by an adult (Stern et al., 1985) and they prefer to look at faces over stimuli that engage them in mutual attention.

A study investigated if mutual gaze would induce feeling of romantic love. Subjects who gazed at their partners' eyes and whose partner was gazing back reported significantly higher feelings of affection, dispositional love and liking (Farroni et al., 2002).

1.2 Mutual Gaze in Human-Robot Interaction

In everyday situations, gaze is not only reactive, but also anticipates and predictes others' behaviour. In such scenarios, gaze is highly informative about intentions and upcoming decisions. An investigation into whether a humanoid robot's mutual or averted gaze influenced how people strategically reason in social decision making, after playing a strategic game with the robot iCub, revealed that participants were slower to respond when iCub established mutual attention before the decision. When people are sensitive to the mutual gaze of an artificial agent, they feel more engaged with the robot (Belkaid et al., 2021).

Robot gaze acts as a strong social signal for humans, modulating response times and decision threshold, promoting neural synchronization, and influencing choice strategies and sensitivity to outcomes. This has strong implications for robotics and clinical applications for all contexts involving human-robot interactions.

2 Aims of the study

- Estimate mutual gaze using neural networks
- Investigate effect of mutual gaze on agreement and disagreement in interaction.
- Understand the uncanny valley effect caused by eerie mutual attention.

3 Gaze Estimation

Gaze estimation aims to predict where the person is looking at by estimating the horizontal and vertical coordinates of the gaze target on a 2-D screen. Deep learning has revolutionised many computer

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

Figure 1: P1 gaze on P2

Figure 2: P2 gaze on P1

vision tasks; however, there is still a lack of guidelines for designing algorithms for gaze estimation in interaction. The GHI corpus (Lavia et al., 2018; Somashekarappa et al., 2020) has speech and gaze annotations in dyadic dialogues which was used in this study for automatic gaze detection.

The main approaches for gaze estimation that currently exist are deep learning-based (Cheng et al., 2021), headmap activated multimodal gaze estimation (Sinha et al., 2021), robust CNN model (Abdelrahman et al., 2022), and U-Net style multistream gaze estimation (synthetic to real transfer learning) (Mahmud et al., 2022). The benchmark datasets are MPIIGaze, Eyediap and UTMultiview.

3.1 Eye gaze classifier

The dataset contains 24 videos, capturing the frontal view of each participant, thus containing two different videos for each session. The videos were recorded at 30 high-definition frames per second.

The vector features were extracted from each video by the PyTorch implementation of MPII face gaze for AlexNet and ResNet14 [1]. For facial landmark detection, a pretrained dlib model was used. The processed video provided landmarks, head pose, projected points of the 3D face model, and a face bounding box. Every frame of the video containing gaze estimation coordinates was then extracted and time stamped. The mutual gaze incident is determined based on the overlapping averted gaze.

[1] https://github.com/hysts/pytorch mpiigaze

3.2 Mutual gaze during agreement and disagreement

Consistent with previous research, we noted that the participants looked at their partner more when listening than speaking. The magnitude of this listening-speaking difference depended on agreement condition, disagreement (but not agreement) exacerbated the maintaining mutual gaze, particularly by averting gaze.

4 Discussion

It is tempting to assume that perfectly matching robot gaze behaviors to human gaze behaviors will elicit identical responses from people, but this is not always the case. Several studies suggest that gaze from robots is interpreted differently than gaze from humans. In general, it is difficult to compare robot gaze to human gaze directly, because while robot gaze can be infinitely controlled, human gaze tends to have small, unpredictable variations.

Once the conversation has begun, conversational fluidity is managed as much by the absence of mutual gaze as by its presence. Virtual agents using gaze aversions for these conversational functions are more successful at regulating the conversational flow and elicit greater disclosure from people than agents that do not perform gaze aversions or perform gaze aversions at inappropriate times (Andrist et al., 2013). Expressive robots could take advantage of these fine-grained gaze behaviors to convey mental states—for example, when they are thinking, when they are waiting for a response, or when they are experiencing difficulty—in a natural and human-like way. For future work, we will conduct a blink estimation study as well as an interactive human-robot experiment.

Acknowledgements

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

References

Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. 2022. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*.

Sean Andrist, Bilge Mutlu, and Michael Gleicher. 2013. Conversational gaze aversion for virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 249–262. Springer.

Marwen Belkaid, Kyveli Kompatsiari, Davide De Tommaso, Ingrid Zablith, and Agnieszka Wykowska. 2021. Mutual gaze with a robot affects human neural activity and delays decision-making processes. *Science Robotics*, 6(58):eabc5044.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2021. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*.

Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H Johnson. 2002. Eye contact detection in humans from birth. *Proceedings of the National academy of sciences*, 99(14):9602–9605.

Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. 2003. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 12–19.

Lisa Lavia, Harry J. Witchel, Francesco Aletta, Jochen Steffens, André Fiebig, Jian Kang, Christine Howes, and Patrick G. T. Healey. 2018. Non-participant observation methods for soundscape design and urban planning. In Francesco Aletta and Jieling Xiao, editors, *Handbook of Research on Perception-Driven Approaches to Urban Assessment and Design*. IGI Global.

Zunayed Mahmud, Paul Hungler, and Ali Etemad. 2022. Multistream gaze estimation with anatomical eye region isolation by synthetic to real transfer learning. *arXiv preprint arXiv:2206.09256*.

Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630.

Neelabh Sinha, Michal Balazia, and François Bremond. 2021. Flame: Facial landmark heatmap activated multimodal gaze estimation. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 759–765.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2021. A deep gaze into social and referential interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Daniel Stern, Lynne Hofer, Wendy Haft, John Dore, Tiffany M Field, and Nathan A Fox. 1985. Social perception in infants.

Which stress is on PRPs?

Maryam Mohammadi
Konstanz University
maryam.mohammadi@uni-konstanz.de

Abstract

This paper is an empirical study investigating on the prosodic patterns of polar response particles (PRPs) in Farsi, where PRPs are ambiguous in response to negative questions. I present novel data showing that while negative answers to positive questions lack prosodic stress (in line with data from Goodhue and Wagner 2018), such responses bear stress when the question obligatorily implicates bias. I claim that two types of stress are used on PRPs in order to either disambiguate the reading in contrast to the alternatives or to express the conflict between what is believed by the speaker (the bias implicature) and the addressee (the answer proposition). I propose that the semantics-pragmatics of each stress can explain the data.

1 Introduction

Polar response particles (PRPs) have been the subject of variety of studies in semantics and pragmatics (Krifka 2013, Roelofsen and Farkas 2015 among others), specially when they are ambiguous between two readings: *polarity*-reading, which marks a response as being either positive or negative (superscripted as $^{Pos/Neg}$) and *(dis)agreement*-reading, which expresses agreement or disagreement with an antecedent (superscripted as $^{Agr/DAgr}$). The ambiguity occurs in languages, like Farsi, in which the same particles can be used in both readings, in the sense that *âre* 'Yes' and *na* 'No' with either of their readings generate the same proposition in response to positive questions (1), however, they result in two different propositions in response to negative questions (2).

(1) A: Ali mehmuni raft?
Ali party went
'Did Ali go to the party?
B1: *âre*$^{Pos/Agr}$. meaning 'he did.'
B2: # *âre*$^{Pos/Agr}$. meaning 'he didn't.'
B3: # *na*$^{Neg/DAgr}$. meaning 'he did.'
B4: *na*$^{Neg/DAgr}$. meaning 'he didn't.'

(2) A: Ali mehmuni na-raft?
Ali party NEG-went
'Did Ali not go to the party?
B1: *âre*^{Pos}. meaning 'he did.'
B2: *âre*^{Agr}. meaning 'he didn't.'
B3: *na*^{DAgr}. meaning 'he did.'
B4: *na*^{Neg}. meaning 'he didn't.'

On the other hand, prosodic stress, mentioned as *rejecting accent*, *verum focus* or *contradiction contour* in different studies (see Goodhue and Wagner 2018), is frequently prescribed for positive answers to negative questions (2.B1, B3), which leads properly to opposition answers and disambiguates the reading. Although (in some languages like Farsi) verum accent and contrastive focus (CF) are prosodically homophones, they are semantically different (Romero and Han 2004, Bill and Koev 2021). The experimental work presented here provides an investigation into how the semantics of prosodic stress, i.e. verum focus and CF separately, can describe the presence/absence of stress on PRPs in response to positive and negative polar questions (PPQs and NPQs respectively).

2 Experimental data

Two experiments were conducted to find the prosodic patterns of PRPs in affirmation and opposition answers. In Experiment 1, I used PPQs (1) and NPQs (2), while in Experiment 2, I considered bias as in (3) and (4), in which a biased particle *dige* obligatorily expresses speaker's expectation towards the uttered proposition in the question:

(3) Ali mehmuni raft dige?
Ali party went DIGE
'Did Ali go to the party?'
⇝ The speaker expects that Ali went.

(4) Ali mehmuni na-raft dige?
Ali party NEG-went DIGE
'Did Ali not go to the party?'
⇝ The speaker expects that Ali didn't go.

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

The data were recorded from 36 Farsi native speakers, reading 18 stimuli. Fifteen data points of f0 trajectory from each particle were automatically extracted (in PRAAT). The pitch track of PRPs are illustrated by GAMMs (from 8370 measurement points in each experiment). In response to PPQs and NPQs, the result of Exp. 1 indicates prosodic saliency as a significant difference in f0 magnitude excursion 26Hz (179-205Hz) on both particles in oppositions, as compared to affirmations 9Hz (184-193Hz). However, considering question polarity (Fig. 1), the data doesn't show significant saliency in PPQs, where PRPs in both oppositions and affirmations have almost 10Hz rising, while in NPQs the f0 excursion in affirmations and oppositions are largely different, 8Hz and 35Hz respectively.

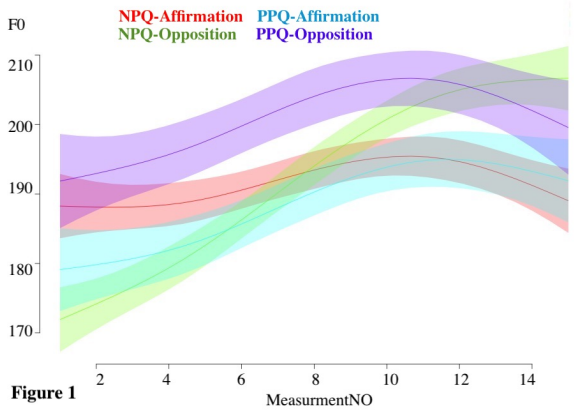


Figure 1

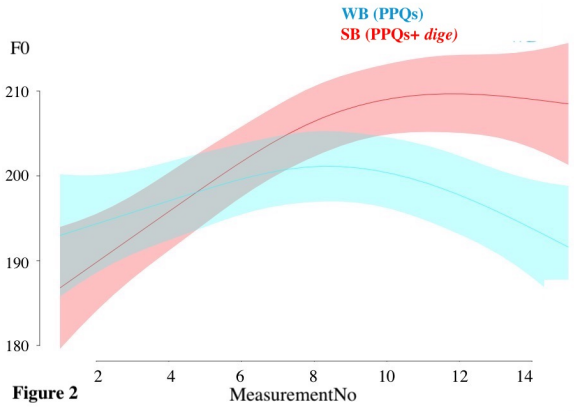


Figure 2

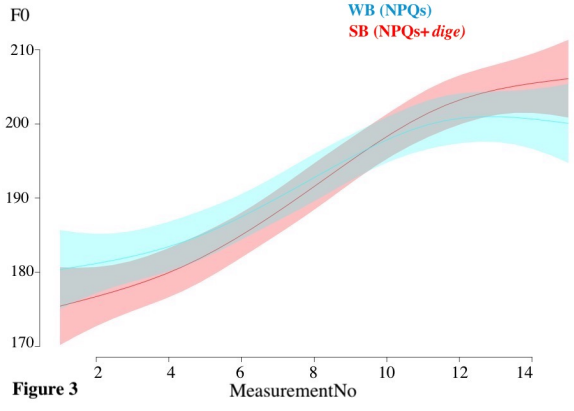


Figure 3

Experiment 2 investigates in opposition answers to strongly biased (SB) questions, where bias was obligatorily expressed by *dige*, as well as to simple questions, where one can accommodate bias weakly (WB). Interestingly, the polarity of the questions did not affect the prosodic saliency of oppositions in response to SB PPQs and SB NPQs. Whilst Figure 3 shows a slight difference in NPQs with respect to the bias strength (f0 magnitude excursion WB=21Hz (180-201Hz) and SB=30Hz (175-205Hz)), there is a significant difference between SB and WB forms of PPQs (f0 magnitude excursion WB=7Hz (193-200Hz) and SB=22Hz (187-209Hz)) as in Figure 2. The data indicates bias affects the prosody.

3 Discussion

I propose that the prosodic stress on PRPs in response to WB and SB questions are different types with different jobs. That is the stress on PRPs in response to WB NPQs is CF, which triggers a set of alternatives of the same particle with different readings (following Rooth 1992), where the set of alternatives in CF contains various possible replacements in the similar domain of the focused expression). Note that in my analysis, PRPs are lexically ambiguous, thus, e.g. for $[na]_F$ we have a set of $\{na^{Neg}, na^{DAgr}\}$, that generates $\{p, \neg p\}$. Hence, the opposition answer is derived via the compositional semantics of CF (à la Rooth) in order to disambiguate the reading. Besides, in response to PPQs the suggested set with either of the readings of *na*, equates a singleton set, $\{\neg p\}$, which contrasts with nothing and predicts truly the absence of CF (in line with data in Exp. 1). In turn, with a set of $\{\hat{a}re, na\}$, one would expect CF in oppositions to PPQs too, which was absent in our data.

On the other hand, verum focus, as the overt realization of VERUM epistemic operator, presupposes the existence of conflicting evidence about the prejacent (Romero and Han 2004, Bill and Koev 2021). In SB questions (both NPQs/PPQs), the bias implicature in the question conflicts with the addressee's belief (the opposition answer), which licenses verum accent on PRPs. Note that the bias in SB is necessarily there and cannot be canceled, while it is optional and cancelable in WB forms. Therefore, the (higher) stress on oppositions to SB NPQs and PPQs is verum focus, indicating the conflict, while CF on oppositions to WB NPQs disambiguates the answer.

References

Cory Bill and Todor Koev. 2021. Verum accent IS VERUM, but not always focus. In *Proceedings of the Linguistic Society of America, [S.l.], v. 6, n. 1*, page 188–202.

Daniel Goodhue and Michael Wagner. 2018. Intonation, yes and no. *Glossa: a journal of general linguistics*, 3.

Manfred Krifka. 2013. Response particles as propositional anaphors. *Semantics and Linguistic Theory*, 23:1.

Floris Roelofsen and Donka F. Farkas. 2015. Polarity particle responses as a window onto the interpretation of questions and assertions. *Language*, 91(2):359–414.

Maribel Romero and Chung-Hye Han. 2004. On negative yes/no questions. *Linguistics and Philosophy*, 27(5):609–658.

M. Rooth. 1992. A theory of focus interpretation. *Nat Lang Seman*, 1:75–116.

Developing a Dataset for Classifying Intents and Sentiments from Judicial Conversations

Palash Nandi[†], **Pinaki Karkun**[†], **Chitra Maji**[†], **Adrija Karmakar**[‡§]
Protyush Jana[‡§], **Arunima Roy**[¶§] and **Dipankar Das**[†]
{sondhanil1, pinaki.karkun, chitra.maji2308, adrijakarmakar2, protyush4711,
arunimaroy1111, dipankar.dipnil2005}@gmail.com
[†]Jadavpur University, Kolkata, West Bengal
[‡]Maulana Abul Kalam Azad University of Technology, West Bengal
[¶]University of Engineering and Management, Kolkata, West Bengal

Abstract

In the present work, we developed a dataset annotated with intents and sentiments at the utterance level. The dataset consists of 430 legal conversations between the user and automated assistant with a total of 2854 utterances (user: 1440, assistant: 1414). The intent annotation follows an ontology provided by experts whereas the sentiment of each user utterance has been evaluated on a scale of -5 to +5. The motivation for including sentiment along with intent was to aid in the generation of an appropriate response. We explored different machine learning (ML) and deep learning (DL) models to accomplish two major tasks: Intent Classification (IC) and Sentiment Classification (SC) to evaluate the usability of the dataset. The results and outcomes were satisfactory for both tasks.

Keywords: intent, sentiment, classification, judicial dataset

1 Introduction

Consultation with a legal expert turns into a necessity to overcome legal issues which can be time-consuming as well as economically challenging. Moreover, serving a large pool of clients simultaneously can be a tiresome job for a legal consultant. A conversational assistant that is able to analyze the client's perspective and suggest accordingly, can be a solution to it. To the best of our knowledge, any large corpus in the legal context is not available to train such an assistant. Thus, we present a conversational dataset in the legal context annotated with intents and sentiments at the utterance level. We also conducted a comparative study of different ML and DL models on the task of IC and SC for assessing usability. The dataset consists of 430 legal conversations with a total of 2854 utterances (user: 1440, assistant: 1414). Each user utterance may fall under multiple intent class out of 29 predefined classes proposed by experts and marked with a sentiment score within a range of -5 to +5 based on annotators' perception. We also carried out a comparative study and error analysis for different models for both intent and sentiment classification. In the case of Intent Classification (IC), Rasa DIET achieves the highest precision of 0.896, recall 0.944, F1-score 0.921, respectively, and outperforms other models. Besides, for Sentiment Classification (SC), RNN performs better in all cases of non-sampling, undersampling, and oversampling in comparison to all other models.

2 Related Work

IC and SC have been in the interest of researchers for a long time. In initial days, lexicon (Kang and Kim, 2003) (Lee et al., 2005), statistical (Liu et al., 2006) or rule (Jansen et al., 2008) based models were considered. In the next decade, authors applied neural models for the same purpose. (Xu and Sarikaya, 2013) had used CNN followed by triangular CRF, (Mesnil et al., 2013) used bi-directional RNN followed by basic CRF where as (Yao et al., 2014) used a modified deep LSTM followed by CRF and softmax for better understanding of the context. (Qin et al., 2019) opted for self-attentive encoder to produce context-aware representation which extracts and summarizes features for IC at sentence and the token level. Recently (Chen et al., 2019) have fine tuned a BERT model for both IC and SC task.

3 Dataset

A total of 2854 utterances were collected [2] from an online legal forum [3] . The raw data was in the form of a sequence of user-posted legal issues and corresponding advice from legal experts (in Indian

§ Equal contribution in the research work as a part of an internship at Jadavpur University.

[2]https://www.crummy.com/software/BeautifulSoup
[3]https://www.kaanoon.com/

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

Speaker	Statement
User	My husband is abusing me for years.
Legal Expert 1	1. File for divorce 2. Apply for maintenance.
Legal Expert 2	1. Make a police complaint 2. Send him legal notices 3. File divorce (optional)

Table 1: Sample of the scrapped corpus

legal context). Table 1 represents a sample of the raw scrapped corpus.

A total number of 430 different legal cases were collected. Later, the raw dataset was converted into a conversational format. Initially, different advice from different legal experts was analyzed to identify the direction and chronology of the events. Each of the events is represented as a pair of an issue followed by corresponding legal advice. Each of the important pairs was concatenated to form different conversational storylines. Finally, the informative ones are chosen to be included in the dataset. Table 2 represents the possible conversations w.r.t the raw text of Table 1 but only the third conversation was considered suitable.

The dataset consists of 29 intents proposed by

Id	Conversation
1	**User**: My husband is abusing me for years. **Bot**: You can file for divorce.
2	**User**: My husband is abusing me for years. **Bot**: File for divorce or opt for mutual settlement.
3	**User**: My husband is abusing me for years. **Bot**: File a written complaint at the police station. **User**: We have tried to solve this mutually but failed. **Bot**: Then file a divorce case on the ground of mental cruelty. **User**: But how will I survive if i divorce him? **Bot**: File a maintenance case too.

Table 2: The possible conversations flow w.r.t. the raw text mentioned in Table 1

experts and each user's utterance is tagged with a sentiment score between -5 to +5. In case of utterances of the agent, the annotation has been limited to intent only.

4 System Description

The system is assessed with two tasks - IC and SC. Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Logistic Regression (LR) were employed for both IC and SC whereas Stochastic Gradient Descent Classifier (SGDC), Multi-class BERT (MBERT), Rasa DIET (RDIET) [4] were used for only IC. Lexicon based model (LBM), Random Forest (RF), Convolution Neural Network (CNN), and Recurrent Neural Network (RNN) were used for SC task exclusively.

5 Experimental Results & Observations

Table 3 represents the experimental outcomes of IC. As observed, MNB performs lowest with F1-score of 0.17 for IC. SGDC, SVM, and LR perform similarly but MBERT and RDIET outperform rest of the models. The RDIET or MBERT uses transformer-based approach that aids in better performance.

	Precision	Recall	F1-score
MNB	0.15	0.26	0.17
SGDC	0.38	0.38	0.37
SVM	0.36	0.42	0.37
LR	0.38	0.42	0.39
MBRT	0.59	0.49	0.53
RDIET	**0.89**	**0.94**	**0.92**

Table 3: Experimental result for intent classification models

For SC, the presence of neutral sentiment is highest followed by negative and positive. To eliminate the bias, a separate study was done on under-sampled and over-sampled data along with the original one. In all of the cases, CNN is able to score similar to RNN but RNN performs best.

6 Conclusions

This paper aimed to develop a conversational dataset in the legal domain and investigate the usability through IC and SC. As observed, transformer-based models perform best because of better contextual understanding. In the future, we will undoubtedly focus on increasing the amount of training data (including devanagari and code-mixed regional Indian languages) and explore other transformer-based models.

[4] https://rasa.com/

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266.

In-Ho Kang and GilChang Kim. 2003. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71.

Uichin Lee, Zhenyu Liu, and Junghoo Cho. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 391–400.

Yiqun Liu, Min Zhang, Liyun Ru, and Shaoping Ma. 2006. Automatic query type identification based on click through information. In *Asia information retrieval symposium*, pages 593–600. Springer.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv preprint arXiv:1909.02188*.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194. IEEE.