SemDial 2020

WatchDial

Proceedings of the 24th Workshop On the Semantics and Pragmatics of Dialogue

Held virtually at Brandeis University, Waltham, Massachusetts
July 18-19 2020

Sophia Malamud, James Pustejovsky & Jonathan Ginzburg (eds.)

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.semdial.org/

Co-presidents: Ellen Breitholtz and Julian Hough

Anthologists: Christine Howes and Casey Kennington

Webmaster: Janosch Haber

WatchDial Website

https://www.brandeis.edu/nasslli2020/watchdial-main/

WatchDial Sponsors

Brandeis Linguistics and Computational Linguistics Programs
Brandeis Department of Philosophy
Brandeis Department of Computer Science

WatchDial Endorsements

Preface

This year WatchDial, hosted at Brandeis University in Waltham (known as ”Watch City”), marked a first as the first virtual SemDial. Despite the challenges of the global COVID-19 pandemic, the workshop went ahead with great success and high attendance, using Zoom conference software, in addition to Slack-based conversations and SpatialChat for virtual coffee breaks.

WatchDial, in tune with the SemDial workshop series as a whole, offers a unique cross section of dialogue research including experimental studies, corpus studies, and computational and formal models.

This year we received 23 full paper submissions, 21 of which were accepted after a peer-review process, during which each submission was reviewed by a panel of three experts. We note that the review scores for the accepted papers are entirely in line with prior years. The demo/breaking news session hosts 7 additional submissions that came in response to a call for demos and research in progress. All accepted full papers and demo/breaking news abstracts are included in this volume.

The WatchDial programme features three keynote presentations by Bonnie Webber of University of Edinburgh, by Craige Roberts of Ohio State and Rutgers University, and a joint presentation by Claire Bonial, Stephanie Lukin, and Clare Voss of the U.S. Army Research Laboratory. We thank them for participating in SemDial and are honoured to have them at the workshop. Abstracts of their contributions are also included in this volume.

WatchDial has received generous financial support from Brandeis University, specifically, the Linguistics program, the Computational Linguistics program, the Philosophy department, and the department of Computer Science. We have also been given an endorsement by the ACL Special Interest Group SigDial.

We would like to thank SemDial co-Presidents, Ellen Breitholtz and Julian Hough, for their guidance and support. We are grateful to Casey Kennington for his continuing stewardship on the challenge of making SemDial proceedings more indexable and to Chris Howes for her continuing archaeology of past proceedings. We would also like to extend our thanks to our Programme Committee members for their very detailed and helpful reviews.

We would also like to thank our local organizers at Brandeis University, including: administrative support from Julie Seeger, Justin Penney, and Michael Golitsyn; student organizers Fox Baudelaire, Dominique O’Donnell, Kenneth Lai, Alex Lu, Hayley Ross; tech support team lead Dimokritos Stamatakis, and tech support team member Angus L’Herrou; and webmaster Victoria Steger. We thank student moderators Che Qu, Lizzie Liang, Vicki Ng, and Luke Parker Gopnik. 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 to science during a global pandemic are a remarkable achievement, and to the participants who attended live sessions and viewed the recordings from all around the globe. Without you, there would be no WatchDial.

Sophia Malamud, James Pustejovsky, and Jonathan Ginzburg

Waltham, MA

July 2020

Programme Committee

Sophia Malamud (chair)	Brandeis University
James Pustejovsky (chair)	Brandeis University
Jonathan Ginzburg (chair)	Université Paris-Diderot (Paris 7)
Maria Aloni	University of Amsterdam
Ron Artstein	Institute for Creative Technologies, USC
Timo Baumann	Universität Hamburg
Luciana Benotti	Universidad Nacional de Córdoba
Anton Benz	Leibniz-Zentrum Allgemeine Sprachwissenschaft
Nate Blaylock	Cerence Inc.
Monica Bucciarelli	University of Turin
Eve V. Clark	Stanford University
Mathilde Dargnat	Nancy University and ATILF-CNRS
Paul Dekker	University of Amsterdam
Simon Dobnik	University of Gothenburg
Lucia Donatelli	Saarland University
Kallirroi Georgila	University of Southern California
Julian Hough	Queen Mary University of London
Christine Howes	University of Gothenburg
Joris Hulstijn	Tilburg University
Julie Hunter	Universitat Pompeu Fabra and Université Paul Sabatier
Elsi Kaiser	University of Southern California
Casey Kennington	Boise State University
Alistair Knott	University of Otago
Alex Lascarides	University of Edinburgh
Pierre Lison	Norwegian Computing Center
Danielle Matthews	University of Sheffield
Gregory Mills	University of Groningen
Philippe Muller	Universitè Paul Sabatier
Valeria de Paiva	Samsung Research America and University of Birmingham
Paul Piwek	The Open University
Massimo Poesio	Queen Mary University of London
Matthew Purver	Queen Mary University of London
Mark Steedman	University of Edinburgh
Matthew Stone	Rutgers University
Takenobu Tokunaga	Tokyo Institute of Technology
David Traum	Institute for Creative Technologies, USC
Enric Vallduví	Universitat Pompeu Fabra
Nigel Ward	University of Texas at El Paso
Grégoire Winterstein	Université du Québec à Montréal
Mats Wirén	TeliaSonera and University of Stokholm
Henk Zeevat	ILLC, University of Amsterdam

Local Organizing Committee

Sophia Malamud (chair)	Brandeis University
James Pustejovsky (chair)	Brandeis University
Victoria Steger (webmaster)	Brandeis University
Dimokritos Stamatakis (tech support lead)	Brandeis University
Angus L'Herrou (tech support)	Brandeis University
Julie Seeger (admin lead)	Brandeis University
Michael Golitsyn (admin)	Brandeis University
Justin Penney (admin)	Brandeis University
Fox Baudelaire	Brandeis University
Dominique O'Donnell	Brandeis University
Kenneth Lai	Brandeis University
Alex Lu	Brandeis University
Hayley Ross	Brandeis University

Table of Contents

Invited Talks

What does it mean to say a discourse relation is 'implicit'? 2
Bonnie Webber

Where to next? Situated Dialogue with Robots for Search and Navigation 4
Claire Bonial, Stephanie Lukin, and Clare R. Voss

Perspective and Indexicality: A NeoLocalist approach 5
Craige Roberts

Paper Presentations

Discourse Coherence, Reference Grounding and Goal Oriented Dialogue 7
Baber Khalid, Malihe Alikhani, Michael Fellner, Brian McMahan and Matthew Stone

Investigating the Semantic Wave in Tutorial Dialogues: An Annotation Scheme and Corpus Study on Analogy Components 19
Jorge Del-Bosque-Trevino, Julian Hough and Matthew Purver

Let's do that first! A Comparative Analysis of Instruction-Giving in Human-Human and Human-Robot Situated Dialogue 31
Matthew Marge, Felix Gervits, Gordon Briggs, Matthias Scheutz and Antonio Roque

Pro-Active Systems and Influenceable Users: Simulating Pro-Activity in Task-oriented Dialogues 41
Vevake Balaraman and Bernardo Magnini

On Laughter and Forgetting and Reconversing: A Neurologically-Inspired Model of Conversational Context 53
Jonathan Ginzburg and Andy Lücking

Different Games in Dialogue: Combining Character and Conversational Types in Strategic Choice 63
Alafate Abulimiti

An Outline of a Model of Situated Discourse Representation and Processing 71
David Schlangen

Semantics with Feeling: Emotions for Abstract Embedding, Affect for Concrete Grounding 81
Daniele Moro, Gerardo Caracas, David McNeill and Casey Kennington

Overcoming Distances – The Usage of Calling Contours to Initiate Dialogues 91
Jakob Maché

Comparing Models of Speakers' and Listeners' Head Nods 103
Tom Gurion, Patrick Healey and Julian Hough

Commitments in German Tag Questions: An Experimental Study 112
Yulia Clausen and Tatjana Scheffler

The Contribution of Dialogue Act Labels for Convergence Studies in Natural Conversations . . . 128
Simone Fuscone, Benoit Favre and Laurent Prévot

Towards the Score of Communication 136
Andy Lücking and Jonathan Ginzburg

Classification of Low-Agreement Pronouns Through Collaborative Dialogue: A Proof of Concept 150
Janosch Haber and Massimo Poesio

Communicable Reasons: How Children Learn Topoi Through Dialogue 160
Ellen Breitholtz and Christine Howes

Extensions Are Indeterminate if Intensions Are Classifiers 170
Staffan Larsson

There Is As Yet Insufficient Data for a Meaningful Answer 180
Maria Boritchev and Maxime Amblard

Dialogue Informing Syntax/Semantics: The Case of Afterthoughts 195
Stergios Chatzikyriakidis

Affordance Competition in Dialogue: The Case of Syntactic Universals 204
Eleni Gregoromichelaki, Stergios Chatzikyriakidis, Arash Eshghi, Julian Hough, Christine Howes, Ruth Kempson, Jieun Kiaer, Matthew Purver, Mehrnoosh Sadrzadeh and Graham White

A Multi-level Framework for Understanding Spoken Dialogue Using Topic Detection 220
Andrew Valenti, Ravenna Thielstrom, Michael Gold, Felix Gervits and Matthias Scheutz

How to Say Things with Hand Gestures 230
Sandy Ciroux

Demo and Breaking News Presentations

Towards Situation- and User-Adaptive Voice Output: Classifying Driver Personality in Context . . 242
Daniela Stier, Wolfgang Minker, Ulrich Heid

A Corpus of Swedish Conversations with a Librarian 245
Magdalena Sandahl, Simon Dobnik, Mehdi Ghanimifard

Detecting Urgency in Speech with Personalised Acoustic Features 248
Jakob Landesberger, Ute Ehrlich

Topic Modelling for a Virtual Librarian Assistant Tool 251
Linnea Strand, Robert Rhys Thomas, Simon Dobnik, Mehdi Ghanimifard

A Statistical and Machine Learning Investigation into the Effect of Demographic Context on Dialogue Phenomena 255
David Hopes, Julian Hough

LDM: A Linear Dialogue Manager 258
Vladislav Maraev, Jean-Philippe Bernardy, Jonathan Ginzburg

From Answer to Question: Coherence Analysis with Rezonator 261
John W. Du Bois, Terry DuBois, Georgio Klironomos, Brady Moore

Invited Talks

What does it mean to say a discourse relation is 'implicit'?

Bonnie Webber
School of Informatics, University of Edinburgh, Edinburgh
bonnie@inf.ed.ac.uk

The term 'implicit (discourse) relation' was created to contrast with 'explicit discourse relation'. The latter was defined as a relation containing an explicit discourse connective drawn from well-defined syntactic classes. (For English, these were coordinating or subordinating conjunctions, or discourse adverbials). Explicit connectives were taken to signal (1) that a relation holds between two text spans (usually sentences or clauses) interpretable as abstract objects (Asher, 1993) and (2) that the relation has one or more particular senses. In contrast, if adjacent sentences were not related by an explicit connective, then an 'implicit discourse relation' was taken to hold if a relation could be inferred between them, that could be made explicit by inserting a connective. The connective so inserted was referred to as an 'implicit connective'.

Implicit discourse relations were defined in terms of adjacent sentences because of limited resources available for annotation (Prasad et al., 2008). In time, some 624 of these implicit discourse relations were seen to have some other word or phrase conveying the sense of the relation and were renamed 'AltLex' relations (Prasad et al., 2010). Similarly, in around 5100 of these relations, the right-hand span was seen to convey information about an entity mentioned in the left-hand span, and were renamed 'Entity' relations or 'EntRels'. In all these cases, the LOCUS of the relation was clear (i.e, in between the two spans) and one or more SENSES could be inferred to hold between them.

The 2008 release of the Penn Discourse TreeBank allowed researchers to experiment with inducing 'shallow parsers' for both explicit and implicit relations, culminating in Shared Tasks at both the 2015 and 2016 Conferences on Natural Language Learning (CoNLL).

At the same time, work on the Penn TreeBank began to move from annotating discourse relations primarily between SENTENCES to annotating relations between CLAUSES within a single sentence (Prasad et al., 2018). This forced us to confront two issues:

- There was no longer a specific LOCUS for implicit relations. Now one needed to decide what signals the LOCUS of an implicit relation, when that locus was no longer the position between adjacent sentences, and when, unlike in the RST-DT corpus (Carlson et al., 2003), it was not assumed that a text could first be partitioned into a sequence of elementary discourse units (EDUs). - what should be taken to signal the SENSE(S) of an implicit relation specified with respect to a given LOCUS?

How these questions are answered has implicatioms for the design of future discourse parsers, as well as informing our understanding of discourse relations.

References

Nicholas Asher. 1993. *Reference to abstract objects in English.* D. Reidel, Dordrecht.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*, pages 2961–2968.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Coling 2010: Posters*, pages 1023–1031.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the pdtb: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97.

Where to next?
Situated Dialogue with Robots for Search and Navigation

Claire Bonial[1], Stephanie Lukin[2], and Clare R. Voss[1]
[1]U.S. Army Research Laboratory, Adelphi, MD
[2]U.S. Army Research Laboratory, Playa Vista, CA
claire.n.bonial.civ@mail.mil

Hey Robot, can you take a picture of the brown object behind you?

This instruction exemplifies several of the challenges we are faced with in developing a human-robot dialogue system in the search and navigation domain—How does the robot know whether this is a question of ability or a command? What implicit sub-actions (such asturning around) might the robot have to take in order to fulfill this instruction? How can the robot ground the unconstrained or underspecified language ("brown object") to its surrounding environment? We describe our research efforts, beginning with how to collect human-robot dialogue data when lacking technology that supports such two-way dialogue. We provide an overview of two types of annotation devised and applied to this data: one that captures the dialogue structure, or the relationship between two utterances, and one that captures both the illocutionary force and the propositional content. We will demo our existing system and highlight ongoing challenges, including grounding the language in the physical environment.

Perspective and Indexicality: A NeoLocalist approach

Craige Roberts
The Ohio State University
roberts.21@osu.edu

Localism is a very old idea among grammarians, stretching back at least to the Byzantine era, reflected in the work of von Humboldt and others in the 19th century, and emerging again in the work of Hjelmslev (1935, 1937) and a lively British tradition seen in the work of Anderson (1971) and Lyons (1977), as well as in more modern incarnation in the work of Talmy, Langacker, and Jackendoff. (Lyons, 1977, p. 718) characterizes it thus:

"[Localism is] the hypothesis that spatial expressions are more basic, grammatically and semantically, than various kinds of non-spatial expressions (...). Spatial expressions are linguistically more basic, according to the localists, in that they serve as structural templates, as it were, for other expressions; and the reason why this should be so, it is plausibly suggested by psychologists, is that spatial organization is of central importance in human cognition."

Localism is often realized in some form of lexical decomposition, for example in the componential analysis of Dwight Bolinger (1975) or in that of Ray Jackendoff. But I offer a rather different approach, based not on decomposition but on presupposition. I argue that cross-linguistically what spatial expressions, indexicals, epistemic modals and a wide variety of other types of expressions have in common semantically is that they all presuppose a point of view with concomitant perspective on the relevant domain—actual space, and the space of epistemic possibilities being two prominent examples. The presupposed perspective, resolved contextually, then has a bearing on the resulting truth conditions of the expression in which the perspectival expression occurs. Moreover, there is evidence in all these domains that interlocutors track the relevant points of view in the discourse in which they're engaged, and that point of view can shift dynamically in the course of interpretation of a single utterance, for example from that of the speaker or the interlocutors collectively to that of the agent of an embedding attitude. This simple, general approach yields insight into a wide range of prima facie unrelated phenomena, including not only the semantics of perspectival locative expressions and deictic motion verbs like come, but the de se character of indexicals (including shifting indexicals), puzzles about epistemic modals, the interpretation of Free Indirect Style, and that of Pottsian Conventional Implicature triggers—appositives and non-restrictive relative clauses, emotive epithets, etc. I'll primarily illustrate this thesis with indexicals like English I and its shifting counterparts in other languages and by discussing how it sheds light on puzzles about epistemic modality.

References

John Anderson. 1971. The grammar of case: Towards a localist theory.

Dwight Bolinger. 1975. *Aspects of Language.* Harcourt Brace Jovanovich.

Louis Hjelmslev. 1935. *La catégorie des cas: étude de grammaire générale.* Universitetsforlaget.

Louis Hjelmslev. 1937. *La catégorie des cas.* Universitetsforlaget.

John Lyons. 1977. Semantics (2 volumes). *Cambridge University Press.*

Oral Presentations

Discourse Coherence, Reference Grounding and Goal Oriented Dialogue

Baber Khalid[†] **Malihe Alikhani**[†*] **Michael Fellner**[*] **Brian McMahan**[†] **Matthew Stone**[†*]
[†]Rutgers Computer Science Department
[*]Rutgers Center for Cognitive Science
{firstname.lastname}@rutgers.edu

Abstract

Prior approaches to realizing mixed-initiative human–computer referential communication have adopted information-state or collaborative problem-solving approaches. In this paper, we argue for a new approach, inspired by coherence-based models of discourse such as SDRT (Asher and Lascarides, 2003), in which utterances attach to an evolving discourse structure and the associated knowledge graph of speaker commitments serves as an interface to real-world reasoning and conversational strategy. As first steps towards implementing the approach, we describe a simple dialogue system in a referential communication domain that accumulates constraints across discourse, interprets them using a learned probabilistic model, and plans clarification using reinforcement learning.

1 Introduction

Referential communication tasks have long served benchmark problems for human–computer dialogue systems (Heeman and Hirst, 1995), and they remain a topic of current investigation (Monroe et al., 2017). One reason is the flexibility and robustness of human joint activity in referential communication (Clark and Wilkes-Gibbs, 1986). Another is the complexity of bootstrapping and reconciling wide-coverage models of semantics in dialogue, on the one hand, and effective models of collaboration under uncertainty, on the other.

Computational work on dialogue has generally integrated collaboration with semantic processing using act-based models (Traum, 1994; Larsson and Traum, 2000; Rich et al., 2001; DeVault, 2008; Galescu et al., 2018). On these approaches, problem-solving actions are recognized indirectly from utterance content using intention recognition, and thereby integrated into a hierarchical model of the dialogue's open goals and plans. While these approaches are powerful and successful, it remains challenging to scale up the necessary models of dialogue structure and intention recognition. For example, system builders have to map out an inventory of transitions that interlocutors can use to update the problem-solving state and have to describe how those transitions are signaled by utterances with specific kinds of form and meaning.

This paper advocates an alternative approach to link the theory of collaborative activity to the practice of dialogue system design, inspired by coherence-based approaches to dialogue semantics such as SDRT (Asher and Lascarides, 2003; Lascarides and Asher, 2009). On a coherence-based approach, collaboration plays out at the level of the discourse as a whole, rather than at the level of the individual utterance. In particular, semantic interpretation attaches utterance content directly into an evolving discourse structure, from which the commitments of interlocutors can be derived as a knowledge graph; it is coherence-based interpretation, rather than plan recognition, that builds a hierarchical structure for the dialogue. Collaborative reasoning comes in later, to connect the coherent dialogue content to its implications for interlocutors' goals, plans and problem solving. In particular, semantic evaluation operations, trained by reinforcement learning or supervised learning from generic resources, ground the knowledge graph in the task context and underwrite system strategy.

We begin this paper with a high-level perspective on this approach, highlighting the relevance of discourse-level interpretation for collaborative reasoning and sketching the possible benefits of eliminating step-by-step models of the task contributions of utterances in favor of holistic task-based evaluations of discourse content. As a first step in implementing this approach, we describe a proof-of-concept system that distinguishes arbitrary color patches from alternatives, as in Monroe

et al. (2017). We emphasize our extensive (but not complete) use of self-supervised and reinforcement learning methods in constructing the system, and report preliminary results showing the effectiveness of a baseline version of the system in interaction with crowd workers. This experience suggests that a coherence-based approach offers a promising alternative direction for systematic efforts to scale up computational approaches to flexible, mixed-initiative dialogue.

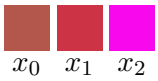
2 Motivation

The key linguistic insight of our work is that referential communication depends on semantic grounding operations that apply at the level of the discourse as a whole, operations that draw not only on the explicit commitments of interlocutors but also on those signalled implicitly. We begin by showing how these insights are captured by coherence approaches to discourse representation, particularly SDRT (Asher and Lascarides, 2003; Lascarides and Asher, 2009), and motivating their importance for referential communication tasks. We argue further that coherence-based approaches create novel, and potentially attractive engineering trade-offs in capturing these phenomena in interactive systems.

As the empirical starting point for our work, we use the Colors in Context (CIC) dataset of Monroe et al. (2017). They asked participants to talk about items in a visual display using a free-form chat interface. On each round of interaction, one human subject, designated the director, was provided privately with a target item from a randomized display and tasked with instructing the other human subject, designated the matcher, to click on the correct item. Examples from this dataset are shown in Figure 1–4. In line with the longstanding findings of empirical work on referential communication since Clark and Wilkes-Gibbs (1986), both director and matcher assume responsibility in these dialogues to make sure that the team members arrive at a mutually satisfactory description of the target patch and complete the task successfully.

2.1 Coherence

Coherence theory (Hobbs, 1979, 1985; Kehler, 2001; Asher and Lascarides, 2003) characterizes conversation through a small number of qualitative relations that link each utterance to an evolving representation of the dialogue. Linking an utterance to a related segment can create a sibling (coordination) or a child (subordination), generating a hierarchical structure. In some approaches to discourse coherence, including SDRT, contributions can have multiple attachments, so discourse structure is a complex graph, not just a tree.[1] This seems to be particularly important for situated conversation (Hunter et al., 2018), including talk about visual contexts like the displays of the CIC dataset.



director:	the very dull red
matcher:	lighter or darker
director:	lighter

Figure 1: An example from the Colors in Context (CIC) dataset (Monroe et al., 2017) of the director and matcher coordinating so that the matcher can click on the correct color patch (x_0).

Coherence relations come in a number of qualitatively different varieties. Some relations signal inferences at the level of domain entities: *Narrative* is a coordinating relation involving descriptions of successive events; *Expansion* is a subordinating relation involving the extended description of a single entity. Some relations are best described in terms of propositions: *Parallel* is a coordinating relation that presents two propositions as analogous; *Explanation* is a subordinating relation connecting a fact to the generalization that explains it. A final class of relations involve meta-level relationships that involve the goals or circumstances of utterances. *Correction* in dialogue is a coordinating relationship of this kind; *Clarification Request* is a subordinating one. Coherence theories postulate substantial divergences between these three different classes of relations. However, relations of any kind suffice to make the conversation coherent and to underwrite inferences about the content of the conversation as a whole, including the commitments of the different speakers.

Schlangen and Lascarides (2003) argue for the suitability of using coherence to characterize the interpretation of the fragmentary utterances—such as those in Figure 1—that are ubiquitous in referential communication, and Stone and Lascarides (2010) show how representations of coherence and speaker commitment can underwrite probabilistic inferences about interlocutors' shared understand-

[1] Hobbs (1985) and Kehler (2001) seem to focus on tree-like discourse structures, for example.

ing in dialogue. We begin by reviewing, extending and deepening those formal accounts. Alternative approaches to these issues are of course possible (Ginzburg, 2012, among others), and we are certainly not in a position to argue against other approaches at this stage of our work. However, we do argue that coherence theory offers an elegant, integrated framework that addresses the structural, semantic and inferential challenges of referential communication in a way that extends to other kinds of dialogue interaction as well.

2.2 Coherence and Dialogue Structure

Coherence theory offers a simple and natural way to represent the STRUCTURE of interactions such as that in Figure 1. Let's look at this interaction in detail. It's useful to segment the utterances into basic moves that establish reference, commit to propositions and contribute open questions. This lets us describe the dialogue as an action u_1 that refers to the target with the description *the very dull red*; a question u_2 whether the target is *lighter* and a question u_3 whether the target is *darker*, combined via an alternation relation into segment σ_4 contributing a complex question; and a contribution u_5 that the target is *lighter*. These fit together into a hierarchical structure:

u_1	discourse initial
σ_4	*Alternative*(u_2, u_3)
	subordinate to u_1
	Clarification(u_1, σ_4)
u_5	coordinate to σ_4
	Answer(σ_4, u_5)
	and *Answer*(u_2, u_5)

The hierarchical structure here follows from the definitions of the relations involved—*Clarification*, *Alternative*, *Answer*—and is built incrementally as moves attach to the discourse. Nevertheless, it supports the key inferences you'd expect for a hierarchical discourse structure, including the possibility for moves to attach high and return to an ongoing topic after a subsidiary inquiry, and the incoherence that interlocutors typically feel when topics are disjointed or interleaved. Moreover, it delivers these inferences as part of general model of dialogue that can be relevant for other tasks such as planning successful dialogue. For example, recognizing terms of the matcher's question *lighter or darker* are parallel alternatives (which might be specified or learned as a typical feature of effective clarification questions).

2.3 Coherence and Semantic Reference

Coherence theory also offers a sophisticated approach to the SEMANTICS of referential communication dialogues. Like all referential communication, the dialogue of Figure 1 is about its contextual situation: the director's utterances work to identify patch x_0. It's common to think of such situated reference in terms of the compositional semantics and pragmatics of individual expressions, as in Kripke's (1978) philosophical account of speaker reference or the words-as-classifiers paradigm of Kennington and Schlangen (2015). However, researchers as early as Luperfoy (1992) have argued that the management of situated reference in the presence of anaphoric relations and partial information requires representing and reasoning about the content of the discourse as a whole.

Figure 2 shows how the problem comes up in the CIC dataset. In cases such as that shown in Figure 2, directors frame difficult reference problems not just by describing the target but also by giving contrasting descriptions of relevant alternatives. Subsequent utterance can identify objects in context by exploiting these established expressions.

Coherence theory might describe this dialogue in a fine-grained way via the composition of four moves by the director (for now we ignore the contribution of *i see*). The director begins with three descriptions that introduce discourse referents into an evolving discourse model: Such inferences are the bread-and-butter of the dynamic approaches to discourse semantics on which coherence approaches such as SDRT are based; see Dekker (2012). We will represent this as three fragments u_1, u_2 and u_3, related by a *Parallel* relation into a segment σ_4. This segment is subordinate to the follow up *the pinkish one* and attaches to it via a *Background* relation; *the pinkish one* also attaches to u_3 by a *Presupposition Resolution* relation through which the definite noun phrase is resolved anaphorically to the discourse referent introduced by u_3. The result is an integrated representation of the speaker's

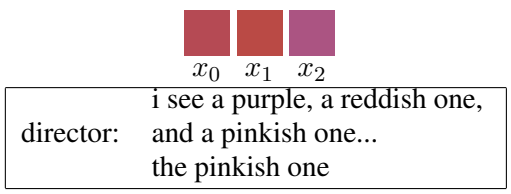


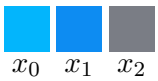
Figure 2: Semantic grounding needs to apply at the level of the discourse to handle anaphoric relations.

commitments in the conversation that features just three discourse referents, each of which is constrained to pick out a different entity in the visual scene, and one of which is designated as giving the target referent for the task.

Note how coreference constraints, disjoint reference constraints, and visual grounding work together at the level of the discourse in interpreting the director's contribution. Evidently, the description *pinkish one* on its own leaves substantial uncertainty about the intended referent; indeed, x_2 might be a more likely match than x_0. However, the implicit constraint that the target can be distinguished from another patch that's *purple* and another that's *reddish* significantly reduces the possibility of misinterpretation. Any constraint-satisfaction approach to ambiguity resolution will handle this, whether logical (Mellish, 1985) or probabilistic (Gorniak and Roy, 2005). But what's to prevent the original uncertainties to resurface when we interpret a new token of the description *the pinkish one*? The natural way is to follow the coherence-based account, and invoke a discourse-level inference that interprets a second occurrence as dependent on (and therefore referentially identical to) the first, in virtue of an anaphoric link.

2.4 Coherence and Common Ground

Coherence itself is vital for calculating an integrated representation of dialogue content (Lascarides and Asher, 2009). Content is often contributed and acknowledged implicitly, in virtue of the coherence relations that attach utterances to the discourse. This is easily regimented in coherence theory. Lascarides and Asher (2009) give a formal model that exploits coherence to identify the commitments that interloctors have made over the course of a coherent discourse. Such calculations can be crucial to tracking mutual understanding in dialogue (Stone and Lascarides, 2010), as we can see even in the CIC domain.



director:	sky blue
matcher:	Is it a darker shade or lighter shade of blue
director:	more greenish

Figure 3: Clarification questions implicate the matcher's implicit acceptance of the possibility of grounding the director's description in the visual scene.

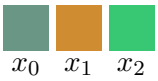
Take clarification requests, as in Figures 1 and 3. Even though clarification requests flag particular aspects of interpretation as incomplete, they implicitly accept the descriptions they target as appropriate to the visual scene. Thus, the matcher of Figure 1 seems willing to accept x_0 or x_1 as a very dull red; that of Figure 3 seems willing to accept x_0 or x_1 as a sky blue. In particular, the dialogue of Figure 3 illustrates how these implicit commitments can inform interlocutors' subsequent coordination: the utterance *more greenish* must be understood to select from the sky blue candidates that have been mutually agreed; x_3 may well be the *more greenish* patch in the display as a whole.

Figure 4 shows what happens when no such implicit commitments can be inferred. Since the matcher's response completely rejects the director's initial description, the director proceeds with an alternative referential strategy—exclusion of the two other patches—that stands independently of the content contributed by the initial utterance.

2.5 Implementing Coherence

In this paper, we explore coherence not just as a theoretical model for dialogue but as an architecture for implementing dialogue systems. Our overarching vision is that utterances in dialogue should be interpreted with a grammatically-specified meaning that's resolved against the linguistic context and attached to an evolving discourse structure. This meaning need not specify abstract updates to the state of the interaction, such as grounding acts or problem-solving acts; it just specifies what's required to track the content of the discourse. It is discourse-level inferences that link speakers' commitments to the real-world task context and underwrite the prediction of the outcomes of conversational strategies based on system experience.

We can contrast this architecture with problem-



director:	blue
matcher:	weird i don't see a blue
director:	weird, hmmm...ok, not orange or green

Figure 4: Implicit commitments contrast with outright rejections. Here, the director's followup strategy of excluding distractors seems uninformed by expectations or conclusions the matcher might have drawn from the initial utterance.

solving approaches that capture semantic grounding and contributions to collaborative activity by reasoning about individual utterances. Compared to a discourse-level approach, an utterance-level architecture is likely to require a finer-grained integration of system capabilities, correspondingly finer-grained data resources—and, given current methods—substantially more handcrafted representation and data annotation.

3 System Architecture

Realizing the vision we have sketched in Section 2 represents a long-term project for dialogue systems design. In particular, it requires us to adapt linguistic processing models so that utterance interpretations contribute to a structured logical form for conversation; it requires us to adapt reasoning and planning frameworks so they apply at the level of discourse content. Will it be possible to make the changes involved?

We think it will. In the rest of this paper, we argue for the practical feasibility of a coherence-based approach to dialogue architecture, using our initial experience implementing a dialogue system using the design principles articulated in Section 2 that successfully plays the referential game investigated by Monroe et al. (2017). As described in Section 3.1, we used the development dialogues of the CIC dataset to design linguistic processing models that interpret utterances as coherent updates to discourse-level representations of logical form. Then, as described in Section 3.2, we combined machine-learned models of lexical semantics and referential interpretation with rule-based models of compositional semantics to characterize the referential ambiguities of these discourse representations. Finally, as described in Section 3.3, we characterize system abilities in terms of a set of high-level strategies for contributing to conversation, and implement a reinforcement-learning approach to selecting those strategies based on their discourse-level effects. These three interacting reasoning models work together to operationalize our approach to referential communication as an empirically-based, discourse-level collaborative process.

3.1 Handling Utterances

Our approach to utterance interpretation is based on a handcrafted context free grammar that compositionally interprets utterances in terms of corresponding formal contributions to dialogue.

The basic units of content in the system are color predications drawn from the XKCD lexicon presented in McMahan and Stone (2015). Using the inference methods described in Section 3.2, we can interpret these predications descriptively, simply as saying what a color is like, or referentially, as identifying the unique patch that fits its content. Compositional operations allow these predications to be combined by logical operations into complex descriptions. Contributions to logical form can introduce discourse referents and characterize them in terms of descriptions; these contributions are defined recursively in terms of coherence relations, which are attached into the overarching discourse structure. Given a discourse structure, we can traverse it to determine the commitments of the director, which involve a set of discourse referents, a sequence of descriptive characterizations of those discourse referents, and inequalities among discourse referents. One of these discourse referents is distinguished as representing the target of the referential communication task.

Our logical form language offers the following moves at the level of discourse: *Identify*, *Distinguish*, *Ask_Clarification*, *Confirmation*, *Rejection*. *Identify*(x, D) commits the speaker that description D identifies discourse referent x; for example, *the green one* corresponds to a move *Identify(T,green)* (for the distinguished target T). *Distinguish*(x, y) commits the speaker that discourse referents x and y are different; for example *not the dark blue* corresponds to two related moves *Distinguish(T, A)*; *Identify(A,dark blue)*. In this simple implementation, *Ask_Clarification* proposes a description of the target object, to which nobody is committed. *Confirmation* (*Rejection*) commit the speaker that this description fits (does not fit) the target.

Every utterance contributes a node which attaches to existing node(s) in the discourse graph. Initially, the discourse graph consists of a single node introducing the target discourse referent. With each new description of target by the director we add a new node in the graph which attaches to most recent node by the director. Whenever a new clarification question comes, it adds a new node in the graph which attaches to the initial node. A confirmation or rejection attaches to the most recent node representing a clarification question and the most recent node by the director.

We handle the idiosyncratic syntax and tele-

graphic structure of color words, acknowledgments and other frequent contributions in CIC utterances using a domain-specific syntactic specification, developed by analyzing the development set of the CIC corpus.[2] We illustrate our approach with a small excerpt from our context-free grammar (CFG) in Table 1.

Category	Expansions	
$\langle S \rangle$	CP	$NegP$
$\langle CP \rangle$	$ADJ : CLR$	CLR
$\langle ADJ \rangle$	$grassy$	$super$
$\langle NegP \rangle$	$NEG : CP$	$NEG : ADJ$
$\langle NEG \rangle$	not	
$\langle CLR \rangle$	Lux Lexicon	

Table 1: A small portion of the CFG developed for our dialogue system's interpretation engine (':' denotes string concatenation).

We extend our basic grammar to a probabilistic-CFG (PCFG) by automatically inducing the relevant weights from the CIC train split. In particular, we obtain all parse trees of the 15285 utterances in the CIC train split, determine the commitments the parse trees contribute to the dialogue, and then weight the trees by the probability that those commitments converge on the actual target of the interaction using the models of Section 3.2. The resulting weighted expansions are used to estimate the probabilities associated with our CFG rules.

We use Earley's algorithm for CFG parsing and A* search to obtain the most likely parse tree given a PCFG. To handle incomplete parses, we used the *first-set* of grammar variables to find the next position where parser could resume to find a sub-parse tree.

Table 2 shows the coverage of our parser over the train split of CIC dataset. The performance shows that our lexicon and grammar rules were able to capture the vast majority of words and structures used by human interlocutors in this setting.

To evaluate parser accuracy, we looked at the first utterances in each dialogue of the CIC test set, and focused on just those cases where the parser returned at least one complete parse. We took the most likely parse tree and computed its semantic interpretation and evaluation. If the actual target of that interaction had the largest posterior probability of the candidate patches, we counted this as a successful item. Overall, the success rate was 88.9%. Even though this evaluation method does not confirm the correctness of the parse itself, we believe that it provides strong evidence for the broad accuracy of our parser.

[2]For example, when analyzing the data, it became apparent that the fact that whether a given occurrence of color name figured syntactically as either a noun or an adjective had no bearing on the meaning derived for the utterance as a whole.

Parse Status	Rate
Complete Parse	.8000
One NOPP	.1329
Two NOPPs	.0353
Three or More NOPPs	.0020
No Parse	.0283

Table 2: Parsing coverage on CIC training data, including complete parses, non-overlapping partial parses spanning segments of the input (NOPPs), and complete parse failure.

3.2 Semantic Evaluation

Given these representations of discourse content, we define a semantic evaluation operation to link that discourse content to the visual world. The input is the structure of speaker commitments. The output is a posterior distribution over candidate color patches for the target. (Although in principle our logical form language allows us to compute posteriors over color patches for any of the discourse referents evoked in the dialogue, we focus here on the typical case where we only need to track the distinguished discourse referent for the target.) In other words, sematnic evaluation defines a probability distribution

$$P(x_i|w^t, w^{t-1}, ..., w^0, C)$$

where w^t represents the descriptive constraint on the target associated with the tth speaker commitment, x_i ranges over the candidate color patches in the scene, and C indicates the dependency of the calculation on a specific visual context, in this case given by three color patches.

For atomic descriptions from the XKCD lexicon, we use the models developed in McMahan and Stone (2020), which are available at `https://go.rutgers.edu/ugycm1b0`. These models are learned in two steps directly from naturally-occurring data. The first step is to train a model of color term meaning from Randall Munroe's crowd-sourced collection of free text descriptions of color patches, as curated by McMahan and Stone (2015).

The next step is to use cognitive modeling and a latent variable model to estimate the prevalence of different speaker strategies for using color meanings to construct identifying descriptions from the CIC dataset. The result is a probability distribution

$$P(w_k|x_i, C)$$

describing the likelihood that a human speaker will use the term w_k to identify a target x_i in the context C. This is readily converted to $P(x_i|w_k, C)$ by Bayes's theorem (on the assumption that each of the targets is equally likely *a priori*).

To extend atomic probabilities to complex predications, we use logical definitions. For example, for a conjoined predication $w^{t1} \wedge w^{t2}$ we have:

$$P(x^i|w^{t1} \wedge w^{t2}) = P(x^i|w^{t1})P(x_i|w^{t2})$$

For disjunction, we get $P(x_i|w^{t1} \vee w^{t2}) =$

$$1 - ((1 - P(x_i|w^{t1}))(1 - P(x_i|w^{t2})))$$

by de Morgan's rule.

Finally, we treat the commitments of the speaker as providing independent evidence about the identity of the target. Again under the assumption that all targets are equally likely *a priori*, this gives

$$P(x_i|w^t, w^{t-1}, ..., w^0, C) \propto \prod_{i=0}^{t} P(x_i|w^i, C)$$

3.3 Learning Data-Driven Decisions

Act-based approaches to collaboration plan contributions to conversation based on formal models of the preconditions and effects of moves in conversation: the system can reason backward from conversational goals to the actions it might use to achieve them. In our approach, the formal effects of conversational actions are defined exclusively by the commitments of interlocutors, as represented in discourse structure, which offers a comparatively thin notion of the dynamics and goals of conversation. We therefore need a different approach to reason about what actions to take next. Our approach is reinforcement learning, which is based on characterizing the outcomes that count as successes in dialogue and anticipating the course of conversational interaction.

A reinforcement learner represents its decision making as a Markov Decision Process (MDP), defined by states S, actions A, a transition function T, rewards R, and discount factor γ. It iteratively improves its estimates of the expected payoffs for an action a in a state s based on experience, and ultimately selects the action whose estimated payoff is the highest. In particular, using data from the Colors in Context dataset, we train a reinforcement learner to approximate the state-action value function under a decision-making policy π, otherwise known as the Q-function, which is given as:

$$Q^\pi(s, a) = r(a) + \gamma * E[max_{a' \in A} Q^\pi(s', a')]$$

We use a feature-based representation of the dialogue state and function approximation to drive a uniform, general estimate of the Q-function.

Developing this approach involves establishing formal models for the actions, reward, state and effects of conversational actions. Our system chooses from two possible actions in each turn, *Ask_Clarification* and *Select*. The description associated with the *Ask_Clarification* move is determined by the referential goal communication model of McMahan and Stone (2020) applied to the most likely candidate referent; we generate the most likely expression not previously used in the dialogue. The *Select* action, meanwhile, picks the most likely color patch μ under the posterior defined in Section 3.2.

We represent the state s^t using features so as to record the history of the dialogue and the current information available about the target:

$$s_i = \{\forall x_i \; P(x_i|w^t, ..., w^0, C) : [a^0, ..., a^t]\}$$

The prior actions in the dialogue are represented as a vector using a one-hot encoding.

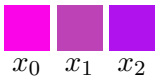
The reward function r is designed to create a tradeoff between dialogue length and task success; good dialogues are both short and effective. In particular, if the correct target is t

$$r(\textit{Select}) = 0.3 \text{ if } \mu = t$$
$$r(\textit{Select}) = -0.5 \text{ if } \mu \neq t$$
$$r(\textit{Ask_Clarification}) = -0.1$$

After 15 turns the dialogue ends in failure with a reward of -1 and no further actions are possible.

In order to anticipate the long-term effects of system choices, we model the responses that the system elicits. This is known as a user simulation. Initially, the simulation identifies the target using a sampled description controlled by a parameter P_x which represents the probability the user will succeed in generating an identifying description. With probability P_x, the user presents a description

that's merely true of the target patch using the semantic model of McMahan and Stone (2020), but with probability $1 - P_x$ the user's description is sampled from their model of identifying descriptions. If the system clarifies, the user simulation confirms if the most likely referent of the distribution is the target, it responds with a *Confirmation*; otherwise it responds with a *Rejection* and a new sampled description.



director:	Pink
matcher:	The hot pink?
director:	Hot pink
matcher:	[SELECT]

Figure 5: An example conversation of a human director interacting with our system using a baseline matcher policy.

4 Analysis and Results

In Section 3, we presented our initial approach to characterizing semantic evaluation and collaborative reasoning at the level of discourse structure. We've seen that we can characterize utterances effectively in this framework. Now we show further that the semantic evaluation operations and collaborative reasoning models are effective.

4.1 Human Evaluation

We ran a human evaluation experiment with a simple rule-based baseline system that uses the semantic evaluation models of Section 3.2 to decide on dialogue strategy. This baseline system went for a *Select* action whenever the probability of most likely referent was greater than 0.95 and chose to ask a clarification request otherwise.

Protocol. This study was conducted with the approval of our human subjects review committee. We recruited 50 subjects through Amazon Mechanical Turk. Participants were all US citizens, gave written consent, and were compensated at an estimated rate of USD 15 an hour. Each subject played 4 games with the system out of which two were selected from the CLOSE condition of Monroe et al. (2017), one from their FAR condition and one from their SPLIT condition. (According to our sensitivity power analysis, with a sample size of 176 conversations we would be able to detect effect sizes as small as 0.820 with a power and significance level of 90%.) After each trial, subjects were asked to rate the performance of the system on a scale of 0 to 5 and leave us feedback.

Results. The overall rate of successful trials is 72.6% and the average overall rating of the system is 3.7. In understanding the effectiveness of discourse-level semantic operations, we are particularly interested in the course of dialogues where information accumulates over time. There are actually 92 cases where the system simply chooses to select on the first turn, picking the correct target in 76 of them (over 82%). In the 123 trials where the system asks for clarification after the first turn, just 78 led to successful identification—a rather disappointing success rate of just over 63%. However, the problem seems to be the users' interpretation of system utterances, not the system's integration of content across the discourse. System clarification questions include descriptions that the utterance-level model of McMahan and Stone (2020) predicts will identify a particular color patch. If that patch is in fact the target of this round, we'd expect users to confirm; if that patch is not the target of this round, we'd expect users to reject. However, users' answers align with the answers we would expect them to give based on the true target only 48% of the time. Truly effective systems will need to ask clarification questions that lead to more reliable answers.

Nevertheless, Figure 5 shows an example conversation of the user with our system that shows how the use of clarification does sometimes lead to a natural and effective conversation in this task. We are releasing the data collected for this evaluation.[3]

4.2 RL Experiments

We ran computer experiments in our RL framework to demonstrate that our approach converges to sensible policies for contributing to a collaborative interaction.

Method. We estimated action outcomes using the deep Q learning (DQN) algorithm of Mnih et al. (2015). The Q function is approximated using linear function θ_p called POLICY NET. Another linear function θ_t also known as TARGET NET is used to predict the future value of a state s. A delta δ is calculated as below and the gradient loss of this

[3] https://github.com/baber-sos/Discourse-Coherence-Reference-Grounding-and-Goal-Oriented-Dialogue

error is back-propagated to learn the Q function.

$$\delta = Q^{\pi}_{\theta_p}(s,a) - (r(a) + \gamma * max_{a' \in A} Q^{\pi}_{\theta_t}(s',a'))$$

Training is done by sampling a mini-batch of size 24 from an experience replay memory of size 200 with a circular buffer, calculating the δ using the Huber loss and then back-propagating using the Adam optimizer. Also, instead of updating θ_t at each iteration using back-propagation, it gets updated by copying the weights of θ_p every *20* iterations. We set the discount factor $\gamma = 1$, learning rate $\alpha = 0.001$ and weight decay parameter $\lambda = 0.01$ during the training process.We train our RL framework for 4000 dialogue simulations on color samples from CIC train split. The model is then tested on 400 simulations over unseen color samples.

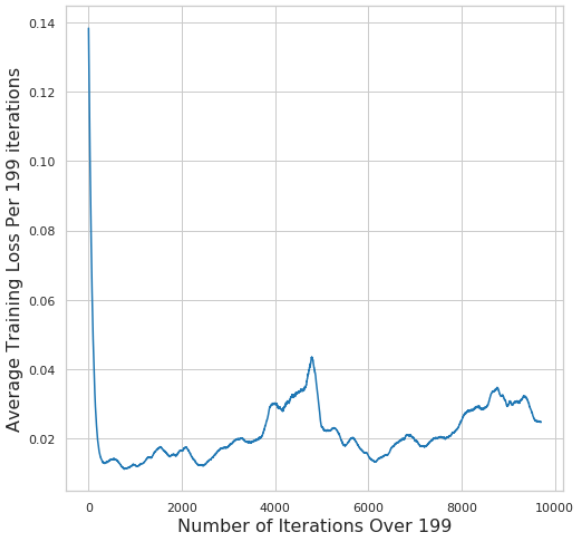


Figure 6: Progression of training loss over 4000 simulated conversations with color samples for CIC dataset train split.

Results. We first report experiments with $P_x = 0.7$, which show that the DQN algorithm is effective at optimizing the expected results of system decisions during training. Figure 6 shows the trend of the training loss and Figure 7 presents the reward progression during training. Both show the system consistently improving over time. Table 3 shows the success rate over test data after each epoch. Average loss over testing data after training is 0.038. Both measures indicate that the model is generalizing well to unseen examples. Ultimately, the simulated success rate is 92% for unseen color patches at test time.

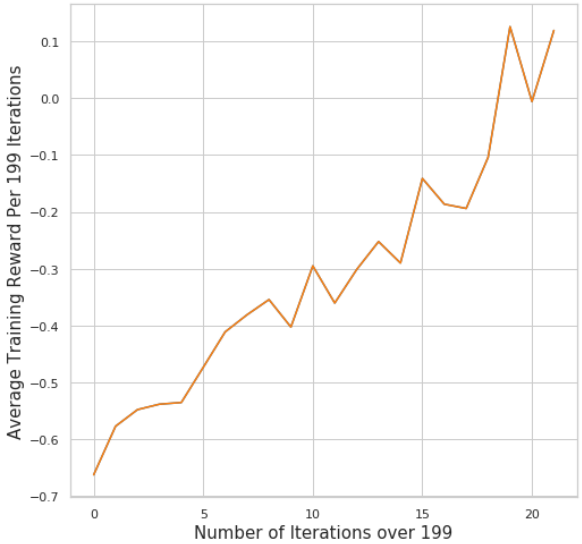


Figure 7: Trend line of average training reward.

	Successes	Failures
1st Epoch	378	22
2nd Epoch	369	31

Table 3: Success vs Failures over Test Set when following the learned policy π. The minor difference in the number of successes is because of imprecision in simulation.

Our next experiments contrast the policies learned with $P_x = 0.7$ against policies learned with $P_x = 0.4$. The differents show that the RL setup is able to learn strategies that respond appropriately to differences in simulated user behavior. Figure 8 shows histograms of model choices at the first turn, as a function of the probability value associated with the most likely referent for the target. Recall that the options are *Selection*, which commits the system to a specific referent, and *Clarification*, which aims to get more information. We can see that the model trained with $P_x = 0.7$—with users that are less successful in identifying the target in the first turn—chooses to ask clarification questions more often than the model trained with $P_x = 0.4$. Thus as it was expected, the model learns to select the color patch only when it is almost certain that it can match the director's description with the target.

We observe that the model trained with $P_x = 0.7$ learned to ask 66.7% more clarification questions at probability values ≥ 0.8. Similarly, it tends to select a patch 5.5% less at the probability values ≥ 0.8. This also showcases how P_x controls the trade-off between clarification and selection.

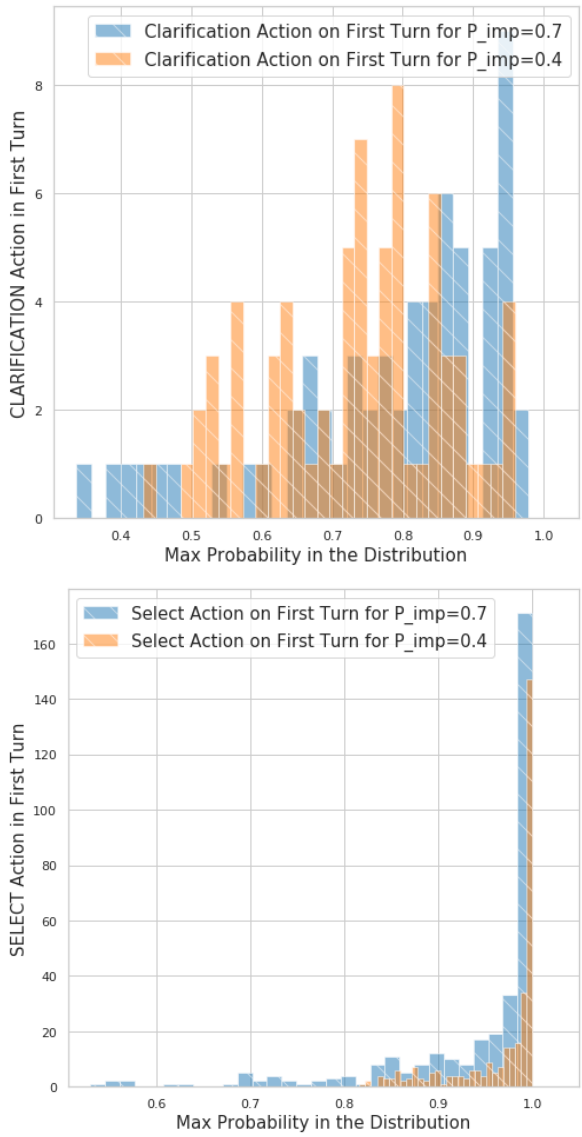


Figure 8: A comparison of the number of *Clarification* actions and *Select* actions chosen by the models.

5 Related Work

The role of the dialogue manager in dialogue systems is to interact in a natural way to help the users complete the task that the system is designed to support (Walker, 2000). Reinforcement learning has been traditionally successful for learning optimized yet flexible strategies in dialogue (Frampton and Lemon, 2006; Manuvinakurike et al., 2017; Walker, 2000; Zhang et al., 2018). To automatically learn dialogue management strategies from dialogue data, these models rely on simulated users, as we do; see Schatzmann et al. (2006) for a review. To learn interaction policies, we follow the approach of Fang et al. (2013); She and Chai (2017), who use a simulated human model built from the dataset (they learn grounded verb semantics interactively in situated human–robot dialogue).

Several projects have studied deep RL for continuous state (and action) spaces, generally in robotics applications (Lillicrap et al., 2016, among others). Nevertheless, to the best of our knowledge, our work is the first that studies the effectiveness of applying deep RL models for learning clarification strategies in dialogue. In fact, however, because these methods make decisions based on the values of continuous parameters, they offer similar flexibility to approaches to dialogue management based on partially-observable MDPs, which are widely used in dialogue management (Williams and Young, 2007; Young et al., 2013). A comparison of partially observable MDPs to deep RL in this domain is left for future work.

Our findings provide new insights into ways that interlocutors employ disambiguation strategies while suggesting the potential discourse approaches in collaborative problem solving in dialogue. The closest work to ours is Schlangen (2004), who presents a model of possible causes for requesting clarifications in dialogue while relating models of communication (Clark, 1996; Allwood, 2000) to the discourse semantics of Asher and Lascarides (2003). Our work goes beyond this approach by providing tools for scaling it up, including by learning semantic models and dialogue strategies.

6 Conclusion

In this paper, we have described the rationale and design principles for using coherence in mixed-initiative referential communication dialogues, and described a preliminary implementation. We showed that we could engineer significant coverage of natural human–human dialogue utterances and strategies, drawing systematically on a range of machine learning methods, and how we could fine-tune the system's purposeful behavior using reinforcement learning drawing on cognitively inspired user simulations. At the same time, we suggested that the system's representations and architecture promise better coverage of anaphorically-mediated real-world reference and dialogues that mix agreement, partial agreement and explicit disagreement.

Some gaps issues remain to offer a complete and convincing demonstration of our framework. We need better performance evaluations, along with

closer links between the reinforcement learning models and the system's interaction with people, ultimately naive users with an intrinsic interest in communicating with the system. We need to substantiate the capacity of the architecture to support mixed initiative by learning strategies where the system acts as the director, not just as the matcher. We need to explore more systematic processing of discourse in context, including a robust and general treatment of anaphora and attachment. And we need to start explicitly drawing on general methods for parsing and DRS construction. Though we would have preferred to use domain-independent tools and resources, it remains a general challenge to build high-precision linguistic resources for natural dialogue. These diverse open problems offer exciting opportunities for future research.

Acknowledgement

The research presented here is supported by NSF Awards IIS-1526723 and CCF-19349243. We thank the Semdial reviewers for helpful comments, and the Mechanical Turk annotators for their contributions.

References

Jens Allwood. 2000. An activity based approach to pragmatics. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue: Studies in computational pragmatics*, pages 47–80. John Benjamins Publishing Company.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

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

Paul J. E. Dekker. 2012. *Dynamic Semantics*. Number 91 in Studies in Linguistics and Philosophy. Springer, Berlin.

David DeVault. 2008. *Contribution Tracking: Participating in Task-Oriented Dialogue under Uncertainty*. Ph.D. thesis, Rutgers University, New Brunswick, NJ.

Rui Fang, Changsong Liu, Lanbo She, and Joyce Chai. 2013. Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 392–402.

Matthew Frampton and Oliver Lemon. 2006. Learning more effective dialogue strategies using limited dialogue move features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 185–192. Association for Computational Linguistics.

Lucian Galescu, Choh Man Teng, James Allen, and Ian Perera. 2018. Cogent: A generic dialogue system shell based on a collaborative problem solving model. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 400–409, Melbourne, Australia. Association for Computational Linguistics.

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

Peter Gorniak and Deb Roy. 2005. Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, ICMI 05, page 138143, New York, NY, USA. Association for Computing Machinery.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, Center for the Study of Language and Information, Stanford University.

Julia Hunter, Nicholas Asher, and Alex Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*.

Andrew Kehler. 2001. *Coherence, Reference and the Theory of Grammar*. CSLI, Stanford.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental 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 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

Saul Kripke. 1978. Speaker's reference and semantic reference. In Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, editors, *Contemporary Perspectives in the Philosophy of Language*, pages 6–27. University of Minnesota Press, Minneapolis.

Staffan Larsson and David Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Susann Luperfoy. 1992. The representation of multimodal user interface dialogues using discourse pegs. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Newark, Delaware, USA. Association for Computational Linguistics.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 331–341.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Brian McMahan and Matthew Stone. 2020. Analyzing speaker strategy in referential communication. In *The 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2020)*.

Christopher S. Mellish. 1985. *Computer Interpretation of Natural Language Descriptions*. Ellis Horwood, Chichester, UK.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Will 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*, 5:325–338.

C. Rich, C. L. Sidner, and N. Lesh. 2001. COLLAGEN: applying collaborative discourse theory to human-computer interaction. *AI Magazine*, 22:15–25.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review*, 21(2):97–126.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143.

David Schlangen and Alex Lascarides. 2003. The interpretation of non-sentential utterances in dialogue. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 62–71.

Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1634–1644.

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

David R Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation.* Ph.D. thesis, University of Rochester.

Marilyn A Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Steve Young, Milica Gai, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Jiaping Zhang, Tiancheng Zhao, and Zhou Yu. 2018. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 140–150.

Investigating the Semantic Wave in Tutorial Dialogues: An Annotation Scheme and Corpus Study on Analogy Components

Jorge Del-Bosque-Trevino[1], Julian Hough, Matthew Purver[2]
Computational Linguistics Lab
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
[1]QMUL Education Lab [2]Jožef Stefan Institute
j.delbosque, j.hough, m.purver@qmul.ac.uk

Abstract

We introduce an annotation scheme and corpus study to investigate the use of base and target components of analogies in tutorial dialogues. We present the development of the scheme and test its final form on a corpus of one-to-one tutorial dialogues on computer science, for which we achieve over 0.77 multi-rater inter-annotator agreement. We then annotate data from the same corpus to investigate the use of semantic wave structures from Legitimation Code Theory in tutoring, and we find a regular adherence to semantic wave structures in explanations which use analogies. We further identified different semantic wave shapes and show their distributions. We conclude that semantic waves and the novel characterisation of analogical explanations in tutorial dialogues reported in this investigation can be useful tools for both the analysis of human tutorial dialogue and future implementation of tutorial dialogue systems.

1 Introduction

We present an empirical study of analogy stemming from the goal of building a spoken dialogue system for computer science tutoring capable of explaining concepts using analogies. While there has been work investigating the use of analogy in tutoring, it is currently insufficiently detailed to build a system with the ability to generate analogies in a interactively natural way; in fact, in general there is an insufficient understanding of how people interact using analogies. In this paper we focus on the sequential unfolding of analogies by tutors on an utterance-by-utterance incremental basis. The paper investigates how tutors go up and down the level of abstraction during their explanations– a structure known as *semantic waves* (Maton, 2013) – with the motivation that discovering how this is done sequentially over the dialogue could eventually be transferred to an artificial tutoring agent.

The rest of the paper is as follows: in Section 2 we explain the theoretical and empirical foundations of explanations, analogies, tutorial dialogues and semantic waves; Section 3 then outlines the first principal contribution of this paper, which is the development of an annotation scheme of *base* and *target* components in analogies within tutoring dialogues which achieves a high inter-annotator agreement for three annotators; Section 4 then presents a corpus study on dialogues annotated using this verified scheme to establish the patterns of base and target annotations to check the extent to which semantic wave teaching is deployed by tutors in dialogue and the distribution of different types of semantic wave, followed by concluding on the findings in Section 5.

2 Background

2.1 Analogies in Explanations

People continuously search, create and evaluate explanations (Keil, 2006; Thagard, 1989) and our explanatory capacity is similar to our ability to reason analogically (Hummel et al., 2014). Analogy is habitually interpreted as a cognitive process involving a *target* domain and a *base* (or 'source') domain, the former being the one that is being explicated and the latter functioning as a different but structurally similar domain used to communicate the concept (Gentner, 1983; Gick and Holyoak, 1983). An example of the base and target components of an analogy in an utterance can be seen in (1) where the base is underlined and the target is in bold.

(1) um the **stack** is a lot like a <u>Lego set</u>, okay?

Analogies are used extensively in explanations in instructional texts (Barbella and Forbus, 2011) and in one-to-one tutoring sessions to explain new

concepts to students (Holyoak et al., 2001). Reasoning with analogies is conceptualised as mapping a single source to a single target (Hummel et al., 2014).

2.2 Tutorial Dialogues

Human one-to-one tutoring has been shown to be a very effective form of instruction (Chi et al., 2001) and is considered one of the most effective methods of helping students to learn. However, there are a number of variables which could either improve or impede learning gains during a tutoring session, including the domain, tutor, tutee and session structure features (Hacker et al., 2009). Under the category of structure-related variables, a number of pedagogical strategies have been tested empirically for their efficacy, including direct procedural instructions, direct declarative instructions, positive feedback, negative feedback, worked-out examples and analogies (Di Eugenio et al., 2013, 2009; Alizadeh et al., 2015). Evidence on the pedagogical efficacy of using analogical explanations shows that their presence in combination with specific dialogue acts correlate positively with learning gain (Alizadeh et al., 2015). However to our knowledge a statistical study on the use of base and target components of analogies within human one-to-one tutorial dialogues showing the structural nuances of how tutors unfold analogies over time has not yet been researched.

2.3 Semantic Waves

The annotation scheme and corpus study we present here aim to uncover patterns of base and target component utilisation in the tutoring dialogues in terms of their adherence to the the structure of *semantic waves*. This concept is part of the Legitimation Code Theory (LCT) (Maton, 2013) and provides an explanatory framework of what constitutes an effective explanation (Waite et al., 2019). According to semantic waves, the complexity of meanings fluctuates in terms of *semantic density* and *semantic gravity*. *Semantic density* is a continuum that ranges from the use of common words utilised with their ordinary meaning at the lowest density to the use of specialized brief terms or symbols at the highest density. *Semantic gravity* contrasts between abstract concepts and real world examples (Maton, 2011). For a learning episode to adhere to the semantic waves construct, it should start with high density and low gravity, descend to low density and high gravity and ascend back

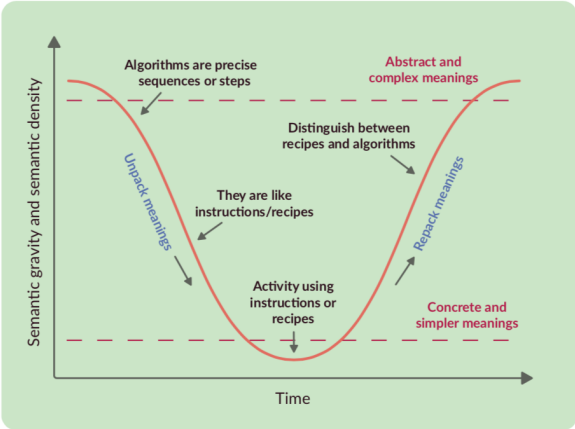


Figure 1: The Semantic Wave technique

to the initial state (Curzon et al., 2018), as illustrated in the diagram in Fig 1[1]. In Fig. 1 initially the concept of algorithms is presented abstractly as precise sequences or steps, then comparison to instructions or recipes is made to unpack the meaning as the semantic density is reduced and a more concrete or simpler base concept is used, followed by a repacking of meanings to go back to the abstract and complex meaning originally presented. In the analogical explanations we study here, we consider the base component to be used at the trough of the wave, and the target component to be at the two peaks. The annotation scheme and corpus study described below seek to answer the question as to whether analogical explanations adhere to the semantic wave structure in tutoring dialogues empirically, and therefore whether the theoretical construct is useful for modelling human-human tutoring dialogue and for designing tutorial dialogues systems.

3 Analogy Annotation Scheme

3.1 Corpora

We used 3 different corpora for developing an analogy annotation scheme, which are the British National Corpus (BNC1994), the Basic Electricity and Electronics Corpus (BEEC) and Computer Science Tutorial Dialogues (CSTD). We selected subcorpora from these three sources which have the following characteristics:

[1]From the National Centre of Computing Pedagogy Quick Read 'Improving Explanations and learning activities in computing using semantic waves', https://raspberrypi-education.s3-eu-west-1.amazonaws.com/Quick+Reads/Pedagogy+Quick+Read+6+-+Semantic+Waves.pdf

British National Corpus (BNC1994) We used one career orientation dialogue of 700 utterances from the British National Corpus (Burnard, 2000), obtained with the SCoRE tool (Purver, 2001).

Basic Electricity and Electronics Corpus (BEEC) We used a dialogue excerpt of 15 utterances and one entire dialogue of 292 utterances of tutorial dialogue of the BEEC corpus (Litman et al., 2009) also obtained with SCoRE.

Computer Science Tutorial Dialogues (CSTD) The largest and principal corpus we use is that of tutorial dialogues on computer science data structures collected in the late 2000's (Di Eugenio et al., 2009) consisting of 54 one-to-one tutoring sessions on the topics of linked lists, stacks and binary search trees. The corpus contains a total of 35,609 utterances annotated with tags signaling beginning and ending of the pedagogical strategies of feedback, worked-out example and analogy. We created a subcorpus of the utterances within all the analogical episodes plus a context of five utterances before the beginning and after the end of the episodes. Our subcorpus contains a total of 3,887 utterances– the size of our subcorpus relative to the size of the whole corpus is as in Table 1.

Utterances within analogical episodes	2,528	7.10%
+ context	1,359	3.81%
Utterances outside analogical episodes	31,722	89.09%
Total	35,609	100.00%

Table 1: Analogical episodes comprising the CSTD sub-corpus as percentage of whole corpus.

3.2 Base and Target Annotation Scheme Development

The development of our annotation scheme included the participation of six researchers, two of whom are the first two authors of this paper. Five of them participated in the annotation exercises. The following paragraphs explain the settings and results of each iteration.

3.2.1 Iteration 1

For the first iteration, two non-native English speakers annotated the BEEC subcorpus described in Section 3.1. The annotators were instructed to mark each utterance as including only the *base* (B), only the *target* (T), *both* (BT) or *none* of the analogy components. Before the annotation exercise, the annotators were provided with an annotated example of the career orientation dialogue from the BNC1994 subcorpus, which was 700 utterances long, of which 25 were annotated with B, T or BT. Each annotator then executed a practice run with an excerpt of a BEEC dialogue of 15 utterances with real-time feedback from the main author. During the interactive practice, the disagreements were discussed with a twofold purpose; identify any annotation rule which elicited disagreements and creating new rules if needed. Every modification and new rules were made explicit in an updated version of the manual. After the practice sessions, utilising the updated version of the annotation scheme manual and in individual sessions, they annotated the other BEEC dialogue comprised of 292 utterances.

The two-way Cohen's Kappa (Siegel and Castellan, 1988) inter-annotator agreement results are as in Table 2 where G is the Gold Standard we assume, which are the annotations by the first author, compared against annotators A1 and A2. The right-hand column also gives the Fleiss' Kappa multi-rater agreement of all three annotators reaching a moderate agreement level of 0.544.

	G&A1	**G&A2**	**G&A1&A2**
B, T, BT, N	0.644	0.304	0.544

Table 2: Kappa inter-annotator agreement on Iteration 1.

Table 3 shows a tutoring dialogue in which the three annotators agree about all the utterances containing the analogy component of type *base*. In this case, the tutor explains electrical potential energy (the target domain) by referring to a ball tossed in the air (the base domain).

Table 4 shows a dialogue excerpt in which the two annotators agree about all the utterances containing the analogy component of type *target*. In this case, the tutor explains conservation of energy.

Finally, Table 5 shows a tutoring dialogue in which the two annotators agree about all the utterances containing both analogy components of type *base* and *target* while the third annotator marks the last two of these as only being *base*.

In addition to the potential ambiguity between both (BT) and base (B), one of the most frequent sources of disagreement in this first iteration was due to not considering anaphoric references to base

P	Utterance	G	A1	A2
T	Think of a ball tossed into the air.	B	B	B
	At first the upward force caused by your hand throwing it causes it to move up.	B	B	B
	But eventually it stops - gravity causes it to slow down until it stops.	B	B	B
	Then it falls down.	B	B	B

Table 3: All three annotators in Iteration 1 agree on base annotation. P = participant.

P	Utterance	G	A1	A2
T	Again, energy would be conserved.	T	T	
	You just have to think what that energy was converted into.	T	T	B
	Some of it would be converted into heat because of the friction, etc.	T	T	B

Table 4: Two annotators agree on target.

P	Utterance	G	A1	A2
T	You're right that kinetic energy was zero, but at the maximum hight, when the ball stops, the height makes it possible for it to start moving again.	BT	BT	
	Now it's going to start moving in the opposite direction.	B	B	
	So that height, since it will make it possible for the ball to move, is a form of energy.	BT	BT	B
	It's the total energy that is conserved, not the kinetic energy, since the velocity of the ball is not constant.	BT	BT	B

Table 5: Two annotators agree on 4 utterances while the other annotator disagrees with them.

and target components and a lack of consistency about marking implicit references to components. As shown in the excerpt of the disagreement analysis of this first iteration on Table 6, the last utterance refers to the base, in this case "a ball tossed in the air", which is only marked as such by A1.

P	Utterance	G	A1	A2
T	Think of a ball tossed into the air.	B	B	B
	At first the upward force caused by your hand throwing it causes it to move up.	B	B	B
	But eventually it stops - gravity causes it to slow down until it stops.	B	B	B
	Then it falls down.	B	B	B
	But at every point, is not the energy the same?	BT	BT	
S	except for when it stops.		B	

Table 6: Excerpt of disagreement in Iteration 1

3.2.2 Iteration 2

For the second iteration, two monolingual native English speakers were recruited as annotators, with the purpose of increasing the inter-coder agreement. The set-up was adjusted such that annotators had to decide whether each utterance contained a base (B) or not as a binary decision, and also whether the utterance contained a target (T) or not. The annotators coded the same BEEC dialogue of 292 utterances used in iteration 1 and received the same coding rules, with the addition of the rule that considers anaphora. They were provided with the same annotated example which was provided as per the previous iteration and also executed the practice annotation with the main author giving live feedback on their decisions. This session allowed for discussion and clarification of the rules in the provided manual. As in iteration 1, all changes were registered in an updated version of the annotation manual. The results from iteration 2 on the two labels are as in Table 7 (again with Cohen's Kappa for the pair-wise agreement and Fleiss' Kappa for the three-way multi-rater agreement).

While very high agreement is reached on the base component, there was large disagreement on identifying target utterances, particularly the agree-

	G&A1	G&A2	G&A1&A2
B	0.878	0.880	0.807
T	0.615	0.211	0.140

Table 7: Kappa inter-annotator agreement from Iteration 2.

	G&A1	G&A2	G&A1&A2
B	0.886	0.731	0.779
T	0.735	0.775	0.772

Table 8: Final Kappa inter-annotator agreement results

ment between the gold standard annotation (first author) and annotator A2.

3.3 Final Annotation Scheme

For the third and final iteration, we used the CSTD corpus for both; the practice and the disagreement and language interpretation experiment, the main reason being the fact that the CSTD was the corpus we wanted to do the study of the semantic waves on. Another change in this iteration was the substitution of one of the two monolingual native English speakers.

The disagreements with A1 and A2 from iteration 2 were discussed and the manual updated accordingly. The annotators coded a new dialogue using the final version of the annotation manual based on these insights. The definitions and examples given to annotators for annotating base and target components is as in Fig 2. An expanded version of the instructions are as in the Appendix A.

The annotators were provided with an annotated example of 6 analogical episodes from the CSTD corpus consisting of 193 utterances. The annotators executed a practice annotation exercise with another selection of 3 analogical episodes and a total of 116 utterances of the same subcorpus and received feedback from the main author with the purpose of clarifying their questions when they judged the annotation rules did not fit in a particular case. Once this practice was executed, to test the agreement the annotators annotated a new selection of 5 analogical episodes for a total 188 utterances of the CSTD corpus. The final inter-annotator agreement Kappa results are as in Table 8, where it can be seen a strong overall agreement with all three annotators is reached for base and target at over 77%.

4 Corpus Study on Analogical Episodes in Tutorial Dialogues

With the appropriate level of agreement reached, the main author annotated the entire subcorpus of the CSTD analogical episodes of 3,887 utterances with the scheme described above (each utterance with the two binary decisions for *base* and *target* presence) which contains all the sections annotated as *analogical episodes* and, additionally, the *single analogies* (Di Eugenio et al., 2009) which consists of individual utterances, and in both cases the 5-utterance context window either side of the analogical utterances.

4.1 Descriptive statistics of all episodes

The histogram in Fig. 3 shows the distribution of lengths of the analogical episodes in number of utterances.

Table 9 and Fig. 4 show the distribution of utterances labelled as containing only base, only target, both or no analogy components, derived from the two binary labels B and T. Note that 33% of utterances contained no analogical components and mainly communicative grounding types of dialogue acts. For the analysis that follows, and consistently with the concept of *question under discussion* (Ginzburg, 2012), we assume those to have a B, T or BT component still under discussion based on the most recent label in the dialogue.

N	B	T	BT	Total
1269	1302	859	457	3887
32.65%	33.50%	22.10%	11.76%	100%

Table 9: Distribution of analogy components labeled in number of utterances and percentage of corpus.

4.2 Validation of Semantic Wave Structures in Analogical Explanations

As discussed in the introduction we aimed to test whether analogical episodes in tutoring dialogues adhere to the structure of semantic waves from Legitimation Code Theory (LCT) empirically. We used the base and target component annotations to validate whether analogical explanations do in fact adhere to this structure and if they do, what the distributions of different types of wave might be.

Analogy definition and examples:
An analogy is a linguistic device that uses a specific concept (a base*) to transfer information or meaning to another concept (a target*). Tutors use analogies to explain concepts. The following examples contain a section of a tutoring dialogue which includes an analogy. The base concept is formatted with underlined characters and the target concept is formatted with bold characters.

Utterance Examples	Base	Target	B	T
um the **stack** is a lot like a Lego set, okay?	lego set	stack	1	1
uh a **binary search tree** is +//		binary search tree		1
one way of looking at it is like a family tree.	family tree		1	
the comparison I like to make with **linked lists** is a movie line.	movie line	linked lists	1	1

Figure 2: Annotation definitions and examples for Base and Target for Annotators

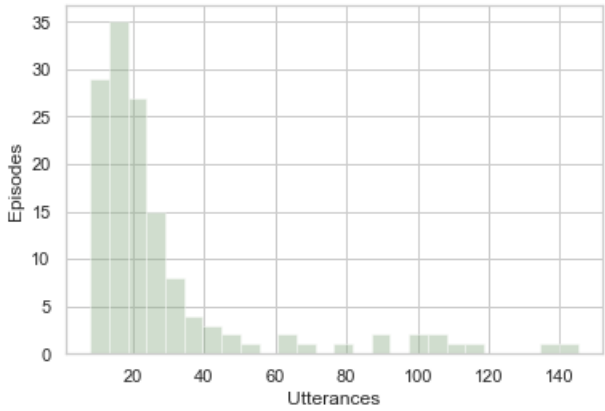


Figure 3: Distribution of analogical episodes length in number of utterances

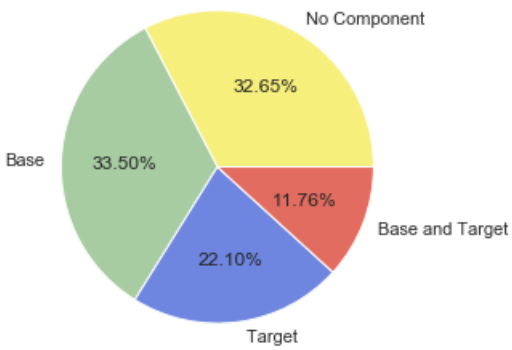


Figure 4: Distribution of analogy components

Vis-a-vis the complexity of meaning in semantic waves, we map the concepts of linked lists, binary search trees and stacks, our target analogical components, to the notion of low gravity (i.e. abstract concept), and the references to people queing at movie theaters, family members, restaurant trays, sheets of paper and other tangible examples, which are our base analogical components, to the notion of high gravity (i.e. concrete concept) in semantic waves.

We would expect analogical episodes to begin with the target component of an analogy, descend in terms of semantic gravity to the base component, and then ascend again to the target concept at the end of the episode– see Section 2 for details. Here we define a semantic wave as any descent in terms of semantic gravity between utterances and rise again, so, for instance beginning with a utterance with both base and target (BT), descending to base (B) and then back to BT is still counted as a wave. Dialogues excerpts exemplifying the semantic waves can be seen in Tables 10, 11 and 12.

To test this, we automatically searched for semantic waves in the 138 analogical episodes of our CSTD subcorpus and we found that 129 (93.47%) of them contained at least one, which supports the idea tutors use the semantic wave in analogical episodes. We also found that a mean of 2.1 waves are used in every analogical episode in a tutorial dialogue explanation.

64 of the 138 episodes had at least two consecutive semantic waves. The distribution over the number of waves per episode can be seen in Fig. 5. One episode contained 16 consecutive waves.

We additionally found that there were seven

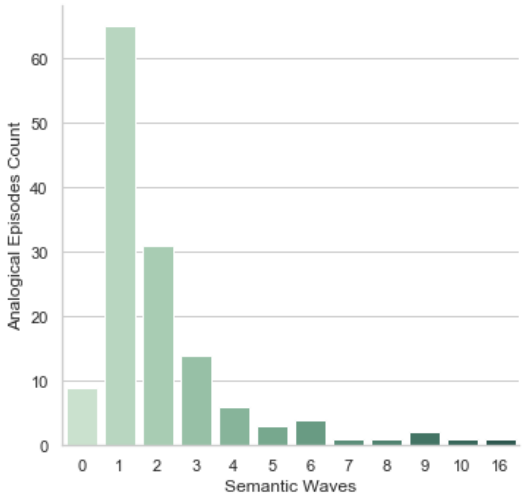


Figure 5: Distribution of analogical episodes density in number of semantic waves

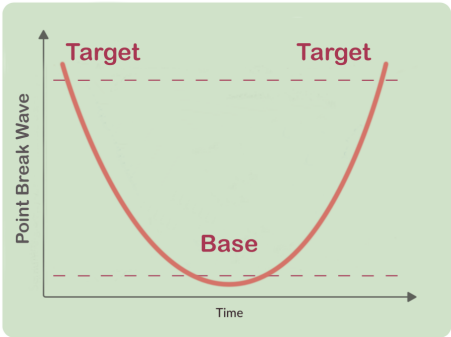


Figure 6: Point Break Wave (T-B-T)

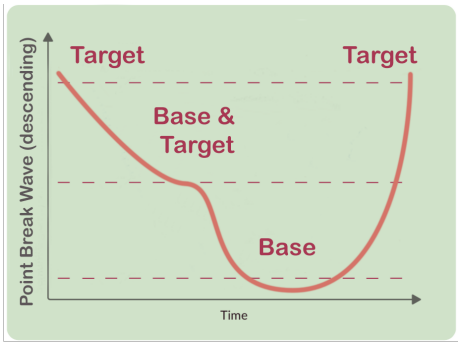


Figure 7: Point Break Wave (descending) (T-BT-B-T)

main types of waves which represent different patterns of base and target components. We take from the surfing domain the names of the types of waves, which vary from strong to weak. The strongest is the *point break wave* as shown in Fig. 6, and in our sequence model it represents starting the analogical episode with a target component, descending

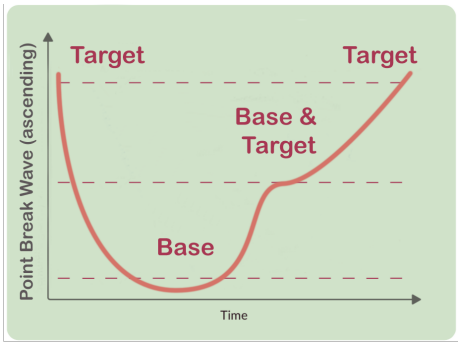


Figure 8: Point Break Wave (ascending) (T-B-BT-T)

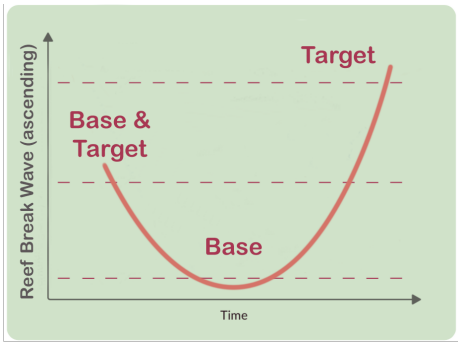


Figure 9: Reef Break Ascending Wave(BT-B-T)

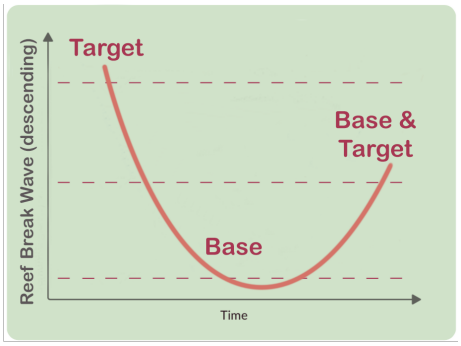


Figure 10: Reef Break Descending Wave (T-B-BT)

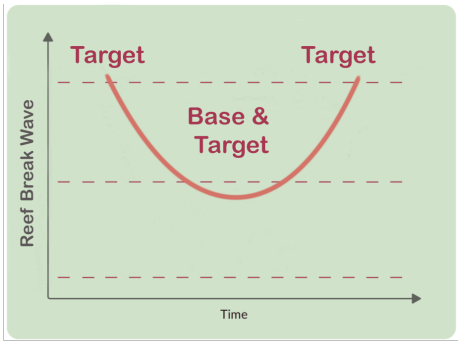


Figure 11: Reef Break Standard Wave (T-BT-T)

to the base component at some point during the

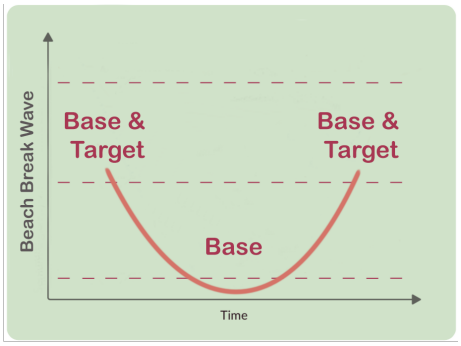


Figure 12: Beach Break Wave (BT-B-BT)

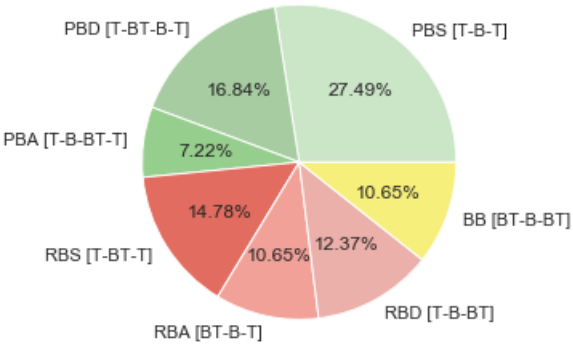


Figure 13: Distribution of semantic wave types

episode, and finishing with the target again. Two sub-types of *point break wave* were also observed, namely the *point break descending* and *point break ascending*– see Figs. 7 and 8. The next type of wave is moderate in intensity and is called a *reef break*, which can be *ascending* (Fig. 9), *descending* (Fig. 10) and *standard* (Fig. 11). Finally, the weakest type of wave was observed, the *beach break* wave (Fig. 12).

In total the 291 waves existing across the 138 episodes were distributed by type is as in Figure 13. Example dialogue excerpts showing a *point break*, *reef break (standard)* and *beach break wave* can be seen in Tables 10, 11 and 12.

5 Discussion and Conclusion

We have presented, to our knowledge, the first annotation scheme and corpus study which investigates how the base and target components of analogies are deployed by human tutors during their explanations. We used the annotation scheme to verify whether analogical explanations follow the structure of semantic waves, whereby they begin from the target component, descend to the base component and return to the target. 93.47% of the episodes contain the structure of a semantic wave and, 74% of the episodes used a series of semantic waves consecutively. We showed there are a variety of different wave types used and we define some shapes to understand these different structures. We claim this novel characterisation of analogical explanations in tutoring dialogues to be a formalisation that could be used as a tool for both; the design of human tutorial dialogue pedagogical strategies and intelligent tutoring systems.

Tutor	alright, stack is a very simple data structure.	T
Tutor	um, this is a shorter and shorter stack of paper.	B
Tutor	and it has a top sheet.	B
Tutor	you can pick up the top sheet or you can put another sheet on the top.	B
Tutor	so the stack +// lets make up one here that has x@1 in it, and d@1 and p@1 and q@1 how about that?	T
Tutor	so here is a stack of four elements.	T
Tutor	here are the operations you can apply to a stack.	T
Tutor	you can pop it.	T
Tutor	and when you pop the stack, that's a destructive function that returns the top element.	T

Table 10: Point Break Wave Dialogue Example

Tutor	it's destructive it takes the q@1 off and gives you the top element what I call n@1 xxx xxx.	T
Tutor	uh the insert is called a push and these come from the spring popping this thing up and the spring pushing down and you push on the stack some value n@1.	BT
Tutor	so this is a function that returns what popped off and that's a void function that takes the thing to make it come off.	T

Table 11: Reef Break (standard) Wave Dialogue Example

Tutor	*uh a binary tree is kind of like mother and father and xxx	BT
Student	a family tree.	B
Tutor	no that's not bad *uh that's bad.	N
Tutor	it's +// because families can have more than two kids.	B
Tutor	so here what it means is that binary is that each node can have two trees, two children.	BT

Table 12: Beach Break Wave Dialogue Example

In future, we intend to further investigate the semantic wave structure in analogies in a more fine-grained level, to analyse the mapping between the base and target domains which happens dynamically during the semantic wave explanation and to incorporate and test these explanatory models in a spoken dialogue system.

Acknowledgements

Del-Bosque-Trevino is partially supported by EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1) and CONACYT (National Council of Science and Technology of Mexico). Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.Thanks to Thomas Kaplan, Louise Bryce, Shamila Nasreen, Morteza Rohanian and Jack Ratcliffe for their participation in the development of the annotation scheme and annotations.

References

Mehrdad Alizadeh, Barbara Di Eugenio, Rachel Harsley, Nick Green, Davide Fossati, and Omar AlZoubi. 2015. A study of analogy in computer science tutorial dialogues. In *Proceedings of the 7th International Conference on Computer Supported Education*, volume 2, pages 232–237. SCITEPRESS-Science and Technology Publications, Lda.

David Michael Barbella and Kenneth D Forbus. 2011. Analogical dialogue acts: Supporting learning by reading analogies in instructional texts. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services http://www.natcorp.ox.ac.uk/docs/userManual/.

Michelene TH Chi, Stephanie A Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G Hausmann. 2001. Learning from human tutoring. *Cognitive science*, 25(4):471–533.

Paul Curzon, Peter McOwan, J Donohue, Seymour Wright, and William Marsh. 2018. Teaching computer science concepts. In Sue Sentance, Eric Barendsen, and Carsten Shulte, editors, *Computer Science Education: Perspectives on Teaching and Learning in School*, chapter 8, pages 91–108. Bloomsbury Publishing, London.

Barbara Di Eugenio, Lin Chen, Nick Green, Davide Fossati, and Omar AlZoubi. 2013. Worked out examples in computer science tutoring. In *International Conference on Artificial Intelligence in Education*, pages 852–855. Springer.

Barbara Di Eugenio, Davide Fossati, Stellan Ohlsson, and David Cosejo. 2009. Towards explaining effective tutorial dialogues. In *Annual Meeting of the Cognitive Science Society*, pages 1430–1435.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Mary L Gick and Keith J Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology*.

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

Douglas J Hacker, John Dunlosky, and Arthur C Graesser. 2009. *Handbook of metacognition in education*. Routledge.

Keith James Holyoak, Dedre Gentner, and Boicho N. Kokinov. 2001. Introduction: The place of analogy in cognition. In Dedre Gentner, Keith James Holyoak, and Boicho N. Kokinov, editors, *The Analogical Mind*, pages 1–19. MIT Press, Cambridge, MA.

John E. Hummel, John Licato, and Selmer Bringsjord. 2014. Analogy, explanation, and proof. *Frontiers in Human Neuroscience*, 8:867.

Frank C Keil. 2006. Explanation and Understanding. *Annual Review of Psychology*, 57(1):227–254.

Diane J. Litman, Johanna D. Moore, Myroslava Dzikovska, and Elaine Farrow. 2009. Using natural

language processing to analyze tutorial dialogue corpora across domains modalities. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pages 149–156, Netherlands. IOS Press.

Karl Maton. 2011. Theories and things: the semantics of discipinarity. In F Christie and Karl Maton, editors, *Disciplinarity: functional linguistic and sociological perspectives*, pages 62–84. Continuum, London.

Karl Maton. 2013. Making semantic waves: A key to cumulative knowledge-building. *Linguistics and education*, 24(1):8–22.

Matthew Purver. 2001. Score: A tool for searching the bnc.

S Siegel and Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Book Co.

Paul Thagard. 1989. Explanatory coherence. *Behavioral and brain sciences*, 12(3):435–467.

Jane Waite, Karl Maton, Paul Curzon, and Lucinda Tuttiett. 2019. Unplugged computing and semantic waves: Analysing crazy characters. In *Proceedings of the 1st UK & Ireland Computing Education Research Conference*, pages 1–7.

A Annotation Instructions

Purpose of the annotation excercise The main purpose of this annotation protocol is to identify interactive patterns of explanations analogies in human-human tutoring conversations. This study will be conducted on a dataset consisting of 54 one-to-one human basic computer science tutoring dialogues collected in the late 2000's. The dialogues topics are limited to three basic computer science (CS) data structures, which are: stacks, linked lists and binary search trees

Analogy definition and examples An analogy is a linguistic device that uses a specific concept (a base*) to transfer information or meaning to another concept (a target*). Tutors use analogies to explain concepts. The following examples contain a section of a tutoring dialogue which includes an analogy. The base concept is formatted with underlined characters and the target concept is formatted with bold characters.

Utterance Examples	Base	Target
um the **stack** is a lot like a Lego set, okay?	lego set	stack
uh a **binary search tree** is +// one way of looking at it is like a family tree.	family tree	binary search tree
the comparison I like to make with **linked lists** is a movie line.	movie line	linked lists

Analogies are easily tractable when they are directly observable, they can be denoted by the presence of particular words or combination of words or keywords which depict either the base or the target of the analogy. Occasionally, analogies are stated in an utterance by the use of coreference. Analogies which are alluded by the use of coreference should also be annotated.

CHILDES notation The research group which annotated the 54 dialogues used the CHILDES notation. The following table contains the symbols you might encounter during your analysis. Use this table as a reference when doing your annotations.

Anaphora, cataphora and refering noun phrases Some analogy bases and analogy targets are alluded indirectly by the use of coreference.

Example	Marker	Meaning
uh# you know what a linked list is?	#	pause between words
it's a +// it's a concept, not a language thing.	+//	self interruption
so all you're given is this header, that why h@l is here.	h@l	the letter h
<*uh, and they want us>[//] oh O_K_.	<>	angle brackets group words marked by the following symbol, in this case, retracing with correction
<*uh, and they want us>[//] oh O_K_.	[//]	retracing with correction
then you're losing all your xxx.	xxx	unintelligible speech
yeah, but yeah, then you know +...	+...	The trailing off or incompletion marker (plus sign followed by three periods) is the terminator for an incomplete,but not interrupted, utterance.
<the second one wants>[///] so that was an insertion, the second one is a deletion.	[///]	retracing with reformulation
(be)cause the xxx.	(be)	noncompletion of a word
you got to start here at the root, just like in <a binary> [//] in a linked list you have to start at the first node.	[//]	retracing with correction
++ right.	++	other completion
star+wars	+	compound or rote form marker

The three cases of coreference that we should be able to spot and annotate are called anaphora, cataphora and coreferring noun phrases.

Continuation and preambles Some utterances contain few words (4 or less) and are continuations of the previous utterance of the same speaker, or a preamble of the following utterance. Assign the same annotations that you gave to that speaker's previous or subsequent utterance.

Session management Session management utterances should not be considered or annotated.

Utterance Example (Metacognition)
let's start off
let me just grab a clean sheet of papper

Metacognition Utterances which are observations or reflections of the mental processes that occur during the tutoring session should not be considered or annotated.

Utterance Example (Session Management)
OK, now you are on to something
I'm so happy that you understand now
It's good that you recognize that your answer doesn' t look right.
That shows you 're thinking!
I like to see that

Additional considerations

- An utterance can be allocated as base and target at the same time.
- Rely only on the text.
- Within the linked lists analogies, tutors and students sometimes refer to letters as if they were people (e.g. b@l is looking to c@l). If this is the case, annotate as BASE. Referencing letters does not determine that speakers are talking about the TARGET
- Annotate as TARGET when there is an explicit reference to concepts within the TARGET domain.

Let's do that first! A Comparative Analysis of Instruction-Giving in Human-Human and Human-Robot Situated Dialogue

Matthew Marge[1], **Felix Gervits**[1*], **Gordon Briggs**[2], **Matthias Scheutz**[3], **Antonio Roque**[3]
[1]U.S. Army Research Laboratory, Adelphi, MD 20783
[2]U.S. Naval Research Laboratory, Washington, DC 20375
[3]Tufts University, Medford, MA 02155
matthew.r.marge.civ@mail.mil, gervif@gmail.com,
gordon.briggs@nrl.navy.mil, {matthias.scheutz, antonio.roque}@tufts.edu

Abstract

We present an annotation scheme that captures the structure and content of task intentions in situated dialogue where humans instruct robots to perform novel action sequences and sub-sequences. This representation identifies patterns and structural differences between human-human and human-robot communications. We find that humans engage in more dialogue about updating beliefs with other humans, while they are significantly more direct in their intentions with robots, incrementally instructing physical actions. Additionally, humans talk significantly less about plans with robots compared to other humans.

1 Introduction

Robots will inevitably be placed in open-world contexts and be expected to accommodate substantial changes in environment and task ability. Moreover, task-based communications between humans and robots will likely involve instructing robots to perform novel actions as it is impractical if not impossible for robots to have pre-defined knowledge that accommodates all situations. However, it is currently not clear whether and to what extent such task-based instructions differ dependent on whether the instructee is a robot or a human. To shed light on this question, we examine task-based instructions given by human instructors to either human or robotic instructees using a novel annotation scheme that captures the structure and content of task intentions in situated dialogue. Ultimately, this scheme will help improve the types of situated dialogues that robots are capable of handling, especially in terms of beliefs and plans.

To explore similarities and possible differences in instruction-giving, we analyze a task-based corpus where the task is for a human instructor to provide an instructee (human or robot) a series of instructions that require the movement of objects situated in the physical world. In advance of the interaction, the instructor prepares a diorama representing a miniature configuration of full-size objects in the space, and verbally tasks the instructee to move objects such that they match the diorama. Instructors participate in two trials, one with another human and one with an anthropomorphic robot as the instructee. The robot is controlled using a "Wizard of Oz" methodology, where wizard experimenters control the robot's behaviors without the instructor's awareness. This dataset, the *Diorama corpus* (originally collected in Bennett et al. (2017)) provides an opportunity to investigate whether human instructors teach robot instructees differently from human ones.

The annotation scheme we present in this paper aims to capture the content of **dialogue moves** in a human-robot interaction task. In doing so, we seek to better understand how humans will frame sequences of instructions to robots. Previous schemes for dialogue acts (e.g., the ISO standard (Bunt et al., 2017, 2020)) focus on observations in human-human dialogue and general-purpose dialogue structures. In contrast, our scheme is concerned with robot-directed language and describing dialogue acts in the context of a specific domain, e.g., performed as part of a conversational game (Carletta et al., 1996) or performed while updating an information state (Traum and Larsson, 2003). While the scheme focuses on the Diorama corpus task, it also builds on established dialogue move taxonomy work (Marge et al., 2017).

While the dialogue move taxonomy we present in this paper can be applied to human-human as well as human-robot dialogue, we consider our taxonomy to be *robot-centric* because at the highest level it is organized around the type of responses that might be expected from a robot. We are partic-

*Most contributions were performed as a graduate student at Tufts University.

ularly interested in *instructions*, defined as instructor utterances that include one or more dialogue moves meant to achieve a task-related physical, verbal communicative, or belief update task intention. A *task intention* is an instructor goal that the instructor wants the instructee to adopt (Cohen and Levesque, 1990); it can include a plan consisting of a sequence of one or more actions. Thus we classify dialogue moves based on the type of response that the instructor intends the instructee to perform, given the current context.

We divide the content of dialogue moves into (1) the dialogue move **type**, which indicates whether the move conveys a physical action intention, verbal communicative action intention, or belief update intention, and (2) the dialogue **move** itself, which identifies the specific actions or plans that the instructor intends that the instructee perform in the environment.

Along with this type and move analysis, instructions are further assessed in terms of their **structure**: whether they include an individual or more than one instruction, and if there is more than one instruction, whether or not the instructions are specified as needing to be in a certain sequence.

Our analysis is performed in this way to help identify notable differences between human-directed and robot-directed instructions in the Diorama corpus. Our contributions are the following:

- We introduce a novel dialogue move taxonomy that captures the structure and content of task intentions in situated dialogue.
- We present a comparative analysis of communications involving the instructing of novel action sequences in human-human and human-robot dialogues conducted in a physical environment, with results that indicate that humans convey more direct intentions to robots compared to other humans.

2 Related Work

2.1 Instruction-Giving in Human-Robot Dialogue

Extensive prior work has explored the core properties of instruction-giving in human-robot dialogue, as well as differences in how humans communicate with robots compared to other humans. Much of this work has focused on the task domain of *navigation* (Bugmann et al., 2004; Koulouri and Lauria, 2009; Marge and Rudnicky, 2011; Marge et al., 2017), where a desired capability for robots, especially in teaming with humans, is to move around the physical world (e.g., to support collaborative exploration, search-and-rescue, or delivery tasks). In a corpus of route instructions, Bugmann et al. (2004) found that robot instructees only succeeded in following the human-provided routes 63% of the time compared with 83% for human instructees. The difference lay in the ability of human instructees to detect and correct errors autonomously.

Other work has explored how error-handling can be achieved through dialogue interaction and how such strategies differ in human-human versus human-robot interactions. In a kitchen scenario, Gieselmann (2006) found that humans largely use "achievement" strategies for error-handling regardless of whether they are talking to a human or a robot. Examples of such strategies include paraphrases, repetition, and requesting missing information.

Other comparative analyses have identified differences in instruction-giving to humans versus robots. Koulouri and Lauria (2009) examined spatial language in the IBL corpus of route instructions in a Wizard-of-Oz task. They found that most instructions to a robot were in the form of simple actions, and that there were more of these kinds of instructions when the human could monitor the robot's actions versus when they could not. Without monitoring, people used more landmark references in their instructions to ensure grounding. Tenbrink et al. (2010) compared the ways in which route instructions were generated in human-human versus human-computer interaction. They found that human-provided utterances to a human partner were more complex than those provided to a computer partner. Specifically, this included more spatial language, more perspective shifting, more location references, and more complex syntax. On the other hand, when the partner was a robot, the instructions tended to be simplified, turn-by-turn commands. As a whole, these results highlight important differences in instruction-giving behavior for navigation tasks, and suggest that humans often simplify their language when instructing robots.

Other task domains have also been explored, including language-guided assembly and manipulation. In a dyadic collaborative task, Bennett et al. (2017) found that human instructors are polite regardless of whether their conversational part-

ner is a human or robot. In particular, they found that humans commonly used *indirect speech acts (ISAs)* (Searle, 1969) to give instructions, and that there was no difference in the frequency of ISAs for human versus robot instructees. In a similar task, Briggs et al. (2017) also found high rates of ISAs in human-generated instructions to robots. The tasks surveyed so far are novel ones without conventionalized social norms, but Williams et al. (2018) found that ISAs are increasingly used in scenarios involving conventional social norms, e.g., restaurant ordering. These results suggest that in non-navigation tasks, people speak to robots much like they do to humans, highlighting the need for robots to represent politeness norms and interpret ISAs in task-oriented dialogue.

2.2 Teacher-Learner Dialogue with Robots

Limited work has explored differences in human-human versus human-robot communication in tasks involving the teaching of novel action sequences. One study by Schreitter and Krenn (2014) investigated the language of instruction-giving in a teacher-learner assembly task. They found that people used a broader vocabulary when instructing humans compared to robots, suggesting that teachers may adapt to the perceived capabilities of a learner. Some work has also explored the role of a robot teaching a human. For example, Torrey et al. (2013) found that human and robot assistants that used more hedges (e.g., "kind of") or discourse markers (e.g., "basically") in a cooking domain were perceived as more likable, more considerate, and less controlling. Moreover, robots were perceived as less controlling than humans even when they used identical discourse markers. These findings were replicated in Strait et al. (2014) and shown to be influenced by additional factors including robot appearance and interaction distance.

While the literature has explored important differences in how humans communicate with robots compared to other humans, what is missing is an understanding of the ways in which humans structure and communicate novel concepts and action sequences. This is important for the development of mechanisms to enable robots to identify plans and actions from dialogue interaction (Laird et al., 2017; Appelgren and Lascarides, 2019). The present work fills this gap by introducing a scheme to capture human task intention and using it to annotate a corpus of human-robot dialogue to better understand the properties of human instruction-giving that robots will need to process.

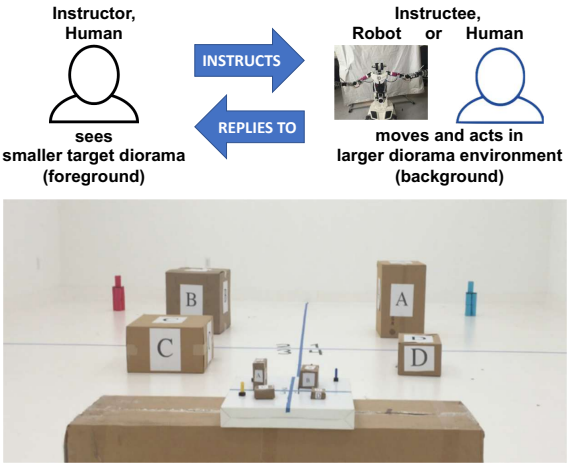


Figure 1: Instructors manipulate a miniature diorama (foreground) to a desired configuration, then command the instructee, who is either human or a robot, to move objects in the larger diorama environment (background) to match. During the task, the instructee cannot directly observe the miniature diorama.

3 Diorama Corpus Overview

The Diorama corpus (Bennett et al., 2017) consists of dialogues between human instructors and either a human or robot instructee. Instructors would teach the human and robot in separate trials a new "skill" (i.e., how to rearrange a set of objects *however they wanted*). As depicted in Figure 1, instructors reconfigured a miniature version of the room (i.e., a *diorama*) from an initial start state, then verbally conveyed a sequence of actions that would be required to move objects to match. The human instructee was a confederate experimenter, while the robot instructee was an anthropomorphic robot.

Instructors (naïve participants) told an instructee situated directly in the environment to perform this reconfiguration. In the segment of the corpus that we analyzed, an experimenter informed the instructor that the robot was fully autonomous. No specific capabilities of the robot were described. Two other experimenters used a Wizard-of-Oz interface to control the robot's movements and verbal responses, which were synthesized as speech.

3.1 Diorama Task

In all trials, the environment was divided into two parts, a *teaching* area where the instructor (seated) would have a diorama, and a larger *experiment* area. The instructor would reconfigure the diorama of

miniature objects from an initial start state, then tell an instructee (human or robot) to replicate their configuration with the full-sized objects in the experiment area.

Beyond minor restrictions, instructors were free to rearrange the objects however they desired. The diorama and experiment area were comprised of four quadrants with seven movable objects: four cardboard boxes labeled with letters and three color-coded towers (in the diorama, towers were composed of Lego blocks while in the experiment area they were metal cans). After configuring their diorama, an instructee would enter the experiment area and introduce themselves. They would engage in spoken dialogue so the instructee would move the objects in the experiment area.

3.2 Interaction Design

Here we briefly outline the interaction design of the Diorama experiments (for additional details, see Bennett et al. (2017)). Example dialogues from the same instructor are presented in Figure 2. During the trials, both instructees had a limited set of dialogue responses. Their responses were constrained to "Okay", "Yes", and "No", with the exception of a few clarification questions regarding status; for example, they could ask "Are there any more instructions?" to verify the trial was over.

The robot was controlled by an experimenter with an interface that permitted the robot to pick up and put down objects with its arms. At the end of the trial, the instructee would say "Goodbye" and leave the experiment area. The order in which the instructor interacted with either the robot or human was randomized.

3.3 Corpus Statistics

The entire corpus consists of thirty-three participants (21 female, 12 male; aged 18 to 25, M=20.85, SD=1.37), but of these, seventeen interacted with a robot they were told was fully autonomous. The focus of our analysis was on this subset because we are interested in the dynamic of perceived robot autonomy and instruction-giving. Not counting standard introductory and closure exchanges, participants said anywhere between 2 and 27 utterances to the instructee in a trial (M=8.7, SD=4.8). Overall, each dialogue between the participant and the instructee ranged from between 6 and 50 task-related utterances (M=17.1, SD=8.8).

Human Instructee Condition	
Instructor:	Um so, I'm going to ask you to move Box D forward...
Instructor:	I guess forward once you turn around, and to the left so that it's right on the number 4.
Instructee:	Okay.
Instructee:	Should I do that now?
Instructor:	Yes please.
Instructee:	Okay.

Robot Instructee Condition	
Instructor:	Please move Box D backwards and to the left so that it sits on top of the number 4.
Instructee:	Okay.
Instructor:	Next, please move Box C backwards and to the right so that it rests on top of the number 3.
Instructee:	Okay.

Figure 2: Dialogue excerpts for an Instructor in Human and Robot Instructee conditions.

4 Corpus Annotation Scheme

We annotated the corpus on three levels: dialogue move type, dialogue move, and instruction structure. To calculate inter-annotator agreement, dialogues from seven randomly-selected participants (totaling 299 utterances) were annotated by two annotators. We report agreement for each level below.

4.1 Dialogue Move Types

Our taxonomy is meant to apply to human-human as well as human-robot interactions, but we think of the taxonomy as being *robot-centric* because its dialogue moves are organized into *types* based on the nature of response that is expected from a robot, for the given dialogue move and the given context. Informally, a robot might be expected to perform a physical action (such as moving a block), or to perform a verbal communicative action (such as describing its capabilities), or to update a belief without further action (such as understanding a plan that the instructor is describing).

More formally, we classify the **dialogue move type** based on *the primary reaction*[1] *that the instructor intends*[2] *the instructee to perform given*

[1]When we say *primary* reaction, we recognize that, for example, when the instructee performs a physical or verbal action, the instructee will typically also update their beliefs (i.e., that the action has been committed, and that the world has changed because of it). However, for the purposes of classification, the *primary* intention of the instructor is that the physical or verbal action be performed.

[2]When we refer to the instructor's *intention*, we use the term in the sense of an action goal that the instructor wants the instructee to adopt (Cohen and Levesque, 1990): to perform a physical action, a verbal action, or a belief update action.

p-action	move-box	*move box D into square 1*
	knock-over-tower	*knock over tower number 2*
	rotate-box	*move box B so its facing that way*
	switch-boxes	*switch box B with box D*
	go-to	*walk over to the yellow tower*
	turn	*turn to your left*
	stop	*okay, stop*
	start-plan	*let's do that first*
	other-action	*yeah, put the yellow ones back up*
	non-task-related	*turn towards me (before we get started)*
v-action	check-task-ability	*see the tallest one over there to your left?*
	check-capability	*can you move those cans with your hands?*
	request-info	*what do I need to do?*
	non-task-related	*how's it going?*
b-update	plan	*then we'll do the same for the last box*
	past-action	*do the same thing with box A*
	update-plan	*we'll just forget about the towers*
	acknowledge	*that's good*
	share-info	*I didn't realize the towers are separate cans*
	non-task-related	*my name is Alex*

Figure 3: Taxonomy of Dialogue Move Types and Dialogue Moves, with examples.

atomic	*move to quadrant 2*	
non-atomic	compound	*knock down the blue and red towers*
	chain	*return to quadrant 1 and grasp the blue tower object*
	complex	*knock down the blue and red towers, then push the fallen blocks to quadrant 2*

Figure 4: Taxonomy of Instruction Structure Types, with examples.

the current context[3]. As shown in Figure 3 and as further described in Section 4.2, the three dialogue move types we identified are:

- **p-actions**: physical action intentions
- **v-actions**: verbal communicative action intentions
- **b-updates**: belief update intentions

There was substantial inter-annotator agreement for dialogue act types with 91.5% raw agreement (Cohen's unweighted $\kappa = .79$).

The dialogue move type is not determined by the full range of possible follow-up actions, responses, or updates that the instructee may perform; if the instructor says, "We'll just forget about the yellow tower," the instructee may legitimately move on to another tower (p-action), answer "okay" (v-action), or just update their beliefs (b-update), but clearly the utterance should not be classified as all three types. As defined above, the specific dialogue move type is defined by the speaker's intentions for the hearer, which is determined based on the context. This definition is motivated by our experience designing interactions for robots, during which one of the main concerns is determining what is expected of the robot at any given moment. Although there are numerous dialogue moves that may vary based on domain and scenario, we maintain that there are a far more limited number of *types* of expectations that those moves represent: expectations that the robot perform an action, that the robot say something, or that the robot do neither of those but instead just add something to its knowledge base of beliefs. From those types of expectations, we identified the three dialogue move types presented here.

[3]When we say the classification should consider the *current context*, this refers to the common ground (Clark and Schaefer, 1989), i.e., mutually understood material.

4.2 Dialogue Moves

We distinguish our *dialogue moves* from established schemes for describing *dialogue acts* (e.g., the ISO standard (Bunt et al., 2017, 2020)) by defining dialogue moves in the context of a specific domain, e.g., performed as part of a conversational game (Carletta et al., 1996) or performed while updating an information state (Traum and Larsson, 2003).

To define the range of dialogue moves that we would study, we began with an existing taxonomy of robot dialogue moves that had been developed in a collaborative navigation task (Marge et al., 2017). We used this as a basis for an initial dialogue move annotation of two representative interactions (four dialogues total) to develop the set of dialogue moves shown in Figure 3, which are divided into three types as described in Section 4.1. There was also substantial inter-annotator agreement for dialogue moves, with raw agreement of 81.6% (Cohen's unweighted $\kappa = .77$).

One type of dialogue move (*p-actions*) involves requests for physical actions such as manipulating a box in various ways, knocking over towers, and performing motions. A qualitatively different move, *start-plan*, indicates that a mutually understood plan (presumably established through belief update dialogue moves) should be initiated.

Another type of dialogue move (*v-actions*) involves requests that the instructee perform a verbal communicative action. For example, *check-task-ability* requests a verbal action that would provide information about the instructee's current ability (e.g., ability to see something, or reach something, or if there is a path somewhere). In contrast, *check-capability* requests a verbal action that would provide information about the instructee's general, potential capabilities (e.g., capability of bending over, or of grasping something in one hand).

The third type of dialogue move (*b-updates*) involves requests that the instructee update one or more of their beliefs without further physical or verbal action. This may involve, for example, utterances in which task-related plans are provided before the plans are actually enacted, utterances that update mutual beliefs while a plan is underway, and maintaining mutual beliefs. *Acknowledgment* dialogue moves provide a general expression of understanding or approval, usually intended as positive feedback. *Plan* dialogue moves refer to future or possible actions in more than a general way; *update-plan* dialogue moves revise previously-established plans, and *past-action* dialogue moves involve references to past actions in any other way – they are unusual in that they include p-action dialogue moves as a parameter, although they are not commands for action in themselves. *B-updates* highlight how the *primary* intended reaction is used to classify the dialogue move type because in a way, all dialogue moves involve updating beliefs. However, b-updates do not have any additional requirement for physical or verbal action.

All three dialogue move types include a catch-all dialogue move. For example, if an annotator encounters a physical action intention that is not explicitly defined by a dialogue move, that would be classified as an *other-action*. An undefined verbal action intention would be a *request-info*, and an undefined belief update would be a *share-info*.

All three dialogue move types also include a *non-task-related* move for utterances that are not part of the Diorama task. For the purpose of this analysis, we consider *preparing for the session* and *providing session setup instructions* to be different tasks than *performing the Diorama task*. As shown in Figure 3, a *p-action:non-task-related* dialogue move might be a request for the instructee to move towards them before beginning the Diorama task; only one example of this, "turn towards me," was seen in the corpus. Similarly, a *v-action:non-task-related* dialogue move might be a general off-topic question; the example in Figure 3 is notional, as no examples were seen in the corpus. Finally, a *b-update:non-task-related* dialogue move might be introductions or a farewell expression.

4.3 Instruction Structure

Along with the type and move annotations described above, instructions are further assessed in terms of their **structure**, as summarized in Figure 4.

An *atomic* instruction is an individual low-level instruction, while *non-atomic* instructions are made up of more than one atomic instruction. This is similar to the distinction between *minimal* and *extended* instructions made in Lukin et al. (2018) and Traum et al. (2018), but we expand upon the non-atomic/extended case in the following ways.

A *compound* non-atomic instruction has no constraints on the execution order of the atomic instructions that make it up. Consider Figure 4: in

		Instructee			
		Human		Robot	
Move Type	**Move**	%	N	%	N
p-action**		21 %	31	52.7 %	79
	move-box	13.6 %	20	27.3 %	41
	knock-over-tower	4.1 %	6	12 %	18
	rotate-box	0 %	0	1.3 %	2
	switch-boxes	1.4 %	2	2 %	3
	go-to	0 %	0	5.3 %	8
	turn	0 %	0	2 %	3
	stop	0 %	0	2.7 %	4
	start-plan	1.4 %	2	0 %	0
	other-action	4.8 %	7	4 %	6
v-action		5.4 %	8	5.3 %	8
	check-task-ability	2 %	3	4.7 %	7
	check-task-capability	0 %	0	0 %	0
	request-info	3.4 %	5	<1 %	1
b-update**		73.5 %	108	42 %	63
	plan*	32.7 %	48	11.3 %	17
	past-action	10.2 %	15	5.3 %	8
	update-plan	1.4 %	2	1.3 %	2
	acknowledge	9.5 %	14	6 %	9
	share-info	22.4 %	33	19.3 %	29
Utterance Count		100 %	147	100 %	150

* p < .05; ** p < .01

Table 1: Counts and percentages of dialogue move types and dialogue moves out of total number of instructor utterances. Significant differences between Instructee groups are marked by * $p < .05$, ** $p < .01$. Note that move type is only counted once per utterance, while each move may be counted more than once.

the example compound instruction, "knock down the blue and red towers," the towers in question can be knocked down in any order.

By contrast, a *chain* non-atomic instruction does have constraints on execution order; in the example chain instruction, "return to quadrant 1 and grasp the blue tower object," the instructee must return to quadrant 1 before grasping the blue tower object.

Finally, a *complex* utterance involves both compound and chain instructions. In the example complex instruction, "knock down the blue and red towers, then push the fallen blocks to quadrant 2," the blue and red towers can be knocked down in any order, but this knocking down must be performed before the fallen blocks can be pushed to quadrant 2.

We found a high rate of inter-annotator agreement for instruction structure, with a raw agreement of 97.1% (Cohen's unweighted $\kappa = 0.89$).

5 Results

The raw counts and percentages of dialogue move types and dialogue moves out of total instructor utterances are reported in Table 1. For statistical testing, we analyzed the proportion of move types, moves, and instruction structure to instructor utterances per trial using analyses of variance

		Instructee			
		Human		Robot	
Instruction Structure	**Structure Type**	%	N	%	N
atomic		78 %	60	88 %	88
non-atomic		22 %	17	12 %	12
	chain	9 %	7	7 %	7
	compound	13 %	10	5 %	5
	complex	0 %	0	0 %	0
Instruction Count		100 %	77	100 %	100

Table 2: Counts and percentages of instruction structure types out of total number of instructions.

(ANOVAs) where instructee type (human or robot) and the order (robot first or not) were considered as fixed effects. Proportions were preferable in the analysis to raw counts because they mitigate the possible variations in dialogue length per participant[4]. After confirming that there were no main or interaction order effects[5], we performed ANOVAs on the proportions by instructee type. All *non-task-related* moves were excluded from our analysis.

5.1 Dialogue Move Types

When analyzing the proportion of dialogue move types to the total number of utterances by the instructor, several significant differences were found. A significant main effect for instructee type was found for both physical action (*p-action*) and belief update (*b-update*) proportions. Instructors used significantly more direct physical action intentions in their dialogues with robots (M=.48, SD=.36) compared to humans (M=.22, SD=.22), $F[1,32] = 6.64$, $p < .01$. In contrast, instructors used significantly more belief updates in their dialogues with humans (M=.71, SD=.27) compared to robots (M=.46, SD=.31), $F[1,32] = 6.01$, $p < .01$.

Of the dialogue moves that participants used, the most meaningful contrast was the use of the *plan* move. There were significantly greater proportions of plans in instructor utterances with the human instructee (M=.33, SD=.30) than the robot instructee (M=.13, SD=.26), $F[1,32] = 4.40$, $p < .05$. Among the other moves, the p-action moves *move-box* and *knock-over-tower* were markedly higher, though not significant, in dialogues with robots over humans.

5.2 Instruction Structure

We report the raw counts and percentage of instruction structure types out of total number of

[4] A secondary ANOVA-based analysis of raw counts confirmed all significant differences found with proportions.

[5] No significant main effects for order were found.

instructions in Table 2. No significant main effects were found for instructee type when comparing proportions of *atomic* instruction structure to total instructions per trial. The proportion of atomic instructions was similar for both instruction-giving to robots (M=.80, SD=.28) and to humans (M=.72, SD=.35).

6 Discussion

The work presented in this paper aims to identify similarities and differences in how humans instruct new action sequences to robots compared to other humans. Our findings suggest several fundamental differences both in the intentions for and the content of instructions.

6.1 Dialogue Move Types and Dialogue Moves

We found that instructors formulate instructions with different intentions when the task is to have an instructee perform actions in the physical world: they tend to engage in more dialogue about updating beliefs with humans, while more directly instructing physical actions with robots. We observed significantly higher proportions of belief update moves in human-human dialogues, while there were significantly higher proportions of physical action moves in human-robot dialogues. This result suggests that the underlying human mental models of the instructee differ with respect to what humans (intuitively) believe needs to be addressed: with humans, instructors know that their instructees know what to do and how to carry out actions if their beliefs are aligned with the instructor, whereas with robots, instructors assume that robots will do what they tell them to do.

Our results also support the conclusions of other studies presented in Section 2.1 where humans formulate less complex instructions to robots compared to other humans. This was further supported by a much higher use of plans in utterances directed to humans compared to robots. Given the complex nature of teaching a set of novel actions to a dialogue partner, these results suggest that humans prefer to rely on belief updating all at once when instructing other humans which is both more succinct and less burdensome for human instructees (because human instructees are aware of all the required actions) than instructing each action directly and incrementally.

In the case of robots, human instructors do not typically know whether their instructees are capable of doing the right thing if given high-level belief states; in fact, human instructors might not believe that robots can understand those types of belief utterances or have those types of belief states. This may be in part due to the fact that instructors were not given any details on the robot's capabilities, forcing them to rely on their own prior notions of what such a robot could do and possibly inferences based on their perceptions of the robot.

6.2 Instruction Structure

Looking at how many actions instructors chose to include in an utterance, we found no significant difference in the proportion of *atomic* to *non-atomic* instructions to either the human or robot instructee. In general, there was an overall heavy use of atomic instructions (at least 70% for both conditions). We believe this is in part due to the shared gaze permitted by the study setup: while they were in the same space, without any limitations on how many instructions they could give, the principle of *least collaborative effort* (Clark and Schaefer, 1989) suggests the least costly way of doing the task is to teach one chunk of the larger task at a time. We suspect that if the instructor and instructee were in separate spaces, or if there were limitations imposed on how many instructions they could give to the instructee, there could be substantial differences.

6.3 Limitations

A complication we identified while performing the annotations was that we received only the utterance-speaker pairs of the corpus, without information about timing, prosody, or visually co-occurrent actions. In the few cases where the instruction-giver appeared to repeat themselves, timing information would have confirmed whether this was actually a self-correction or a re-issuing of an instruction that was not acknowledged by the instructee. Additionally, a few utterances (e.g., "Like this?") referred to a gesture that the instructee was performing. While it was clear a physical action was underway, identifying the intention could be made clearer with access to visual recordings. We recommend in future studies that investigate such dialogue phenomena that visual and timing information be collected in addition to the spoken dialogue.

We also note that some of the differences in instruction-giving to human and robot instructees are because humans generally perceive robots dif-

ferently from other humans (Bartneck et al., 2020). This fundamentally implies that there will be differences in the instructions given to robots; our results confirm this general concept.

7 Conclusions

In this paper, we presented a novel annotation scheme that enables comparisons between how humans instruct either a human or a robot instructee to perform previously unknown actions. The scheme centers on capturing the structure and content of task intentions in situated dialogue. We annotated a corpus of instruction-giving dialogues (Diorama corpus), and found substantial differences in how humans instruct robots compared to other humans. Humans tend to use dialogue to update beliefs of their instructees, as also indicated by a marked increase in the use of plans. In contrast, human instructors were more direct with robots, instructing physical actions in the space.

The results highlight the need to better understand people's implicit mental models of robots and what effects these mental models have on how humans formulate instructions for robots. This is particularly critical for the design of natural language understanding systems as well as cognitive architectures for future, more sophisticated robots. Such robots will likely be confronted with more humanlike instructions, and so will need to understand high-level beliefs and translate them into executable actions, which requires an understanding of the links between high-level goals, purposes, and how to realize them – all capabilities that current robotic systems lack. Future work should then explore how human instructors relate to perceived or known robot capabilities, and what level of sophisticated instructions robots would then be able to handle based on the changes to human mental models of robots.

Acknowledgments

This research was sponsored by the Basic Research Office of the U.S. Department of Defense with a Laboratory University Collaboration Initiative Fellowship awarded to the first and third authors and by a NASA Space Technology Research Fellowship (NSTRF) awarded to the second author. The authors would like to thank the anonymous reviewers for their helpful comments.

References

Mattias Appelgren and Alex Lascarides. 2019. Coherence, symbol grounding and interactive task learning. In *Proc. of SemDial*.

Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-Robot Interaction: An Introduction*. Cambridge University Press.

Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *Proc. of IROS*.

Gordon Briggs, Tom Williams, and Matthias Scheutz. 2017. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 6(1):64–94.

Guido Bugmann, Ewan Klein, Stanislao Lauria, and Theocharis Kyriacou. 2004. Corpus-based robotics: A route instruction example. In *Proc. of Intelligent Autonomous Systems*.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proc. of LREC*.

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*. Springer.

Jean Carletta, Amy Isard, Jacqueline Kowtko, Gwnyeth Doherty-Sneddon, and Anne Anderson. 1996. *HCRC Dialogue Structure Coding Manual*. Human Communication Research Centre.

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

Philip R. Cohen and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261.

Petra Gieselmann. 2006. Comparing error-handling strategies in human-human and human-robot dialogues. In *Proc. of KONVENS*.

Theodora Koulouri and Stanislao Lauria. 2009. A corpus-based analysis of route instructions in human-robot interaction. In *Proc. of Towards Autonomous Robotic Systems (TAROS)*.

John E. Laird, Kevin Gluck, John Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario Salvucci, Matthias Scheutz, Andrea Thomaz, Greg Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive task learning. *IEEE Intelligent Systems*, 32(4):6–21.

Stephanie Lukin, Kimberly Pollard, Claire Bonial, Matthew Marge, Cassidy Henry, Ron Artstein, David Traum, and Clare Voss. 2018. Consequences and factors of stylistic differences in human-robot dialogue. In *Proc. of SIGdial*.

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

Matthew Marge and Alexander Rudnicky. 2011. The TeamTalk corpus: Route instructions in open spaces. In *Proc. of the RSS Workshop on Grounding Human-Robot Dialog for Spatial Tasks*.

Stephanie Schreitter and Brigitte Krenn. 2014. Exploring inter- and intra-speaker variability in multimodal task descriptions. In *Proc. of RO-MAN*.

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

Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proc. of HRI*.

Thora Tenbrink, Robert J Ross, Kavita E Thomas, Nina Dethlefs, and Elena Andonova. 2010. Route instructions in map-based human–human and human–computer dialogue: A comparative analysis. *Journal of Visual Languages & Computing*, 21(5):292–309.

Cristen Torrey, Susan R Fussell, and Sara Kiesler. 2013. How a robot should give advice. In *Proc. of HRI*.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory J. Hayes, and Susan G. Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proc. of LREC*.

David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353.

Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. 2018. "Thank you for sharing that interesting fact!" Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proc. of HRI*.

Proactive Systems and Influenceable Users: Simulating Proactivity in Task-oriented Dialogues

Vevake Balaraman[1,2], **Bernardo Magnini**[1]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]ICT Doctoral School, University of Trento, Italy
{balaraman, magnini}@fbk.eu

Abstract

We investigate proactivity, the capacity of a dialogue system to provide relevant information even when not explicitly requested, in the context of task-oriented dialogues. We propose to extend the current task-oriented framework, and have investigated four aspects of proactivity: (i) the degree of proactivity provided by the system during a dialogue; (ii) the propensity of the user to be influenced by the system proactivity; (iii) the complexity of the domain ontology; (iv) the relation between user needs and application domain, in terms of expected failure situations. Under the hypothesis that proactivity helps to increase effectiveness and efficiency of dialogues, we set up a framework based on dialogue simulations, and experimented the four aspects mentioned above. Although the current implementation allows to simulate a limited amount of dialogue phenomena (e.g., system initiative only), we are able to show that proactivity might have strong effects on dialogues, reducing up to 60% of dialogue turns in an application domain of medium complexity.

1 Introduction

Proactivity is a fundamental characteristic of human-human collaborative dialogues, consisting in the attitude of the speakers to provide information that can be used to achieve the goal of the dialogue, even when such information is not explicitly requested. This attitude obeys to the so called *principles of cooperative dialogue*, which have been summarized in the popular Grices's maxims (Grice, 1975). Among the others, proactivity is very common in instruction-giving dialogues. For instance, assuming the following dialogue between speakers A and B,

(A) *What time is the next train to London?*
(B) *The next train is at 1015. It arrives at 1245.*

we say that *It arrives at 1245* is a proactive response, because information given by B is more than was asked for by A, and because B guesses that this is the sort of information A might also need, and so offers it unsolicited. The major effect of proactivity is that, by anticipating user needs (e.g., a question like *What time this train will arrive in London?* in our example), it avoids long dialogues, making communication more efficient.

Proactive behaviours share some aspects with other communicative devices. Persuasion (Fogg, 2002; Wang et al., 2019) provides not requested information too, although not necessarily in response to a user request (i.e., we can try to persuade even without a dialogue). Recommendation (Chen et al., 2019), has assumed an application connotation, i.e., suggesting items that match the user profile during a search process. We prefer to use *proactivity* as a general term to indicate any situation in a task-oriented dialogue where the system takes the initiative to provide a piece of non requested information with the goal of better achieving the user needs.

Finally, proactivity goes beyond verbal communication, and may involve multimodal modalities. A well known experience has been Microsoft Clippy, launched as an helper for the Microsoft operative system. The substantial difficulties encountered by Clippy to be well accepted by users, mostly because of its perceived invasiveness, should be considered as a signal that effective proactivity is actually a very complex phenomenon.

Although very common in human-human dialogues, proactivity is almost absent from the current research in task-oriented dialogue systems. To the best of our knowledge, proactivity is largely under represented in most of the datasets (e.g., ATIS (Price, 1990), WoZ (Mrkšić et al., 2017), MultiWoz (Budzianowski et al., 2018), Snips (Coucke et al., 2018)) that are used to train either natural language understanding or dialogue state tracker components

Non-proactive Dialog	Proactive Dialog
User: Are there any Eritrean restaurants in town? **Sys:** No, there are no Eritrean restaurants in town. Would you like a different restaurant?	**User:** Are there any eritrean restaurants in town? **Sys:** No, there are no eritrean restaurants in town. *Most of the restaurants are Chinese, particularly in the East area of the town*
User: How about Chinese food? **Sys:** There is a wide variety of Chinese restaurants. Do you have an area preference or a price preference to narrow it down?	**User:** Chinese food in the East area is fine. **Sys:** Yu Garden is a Chinese restaurant in the east area, *with mid level price.*
User: I would like the East part of town. **Sys:** Yu Garden is a Chinese restaurant in the east area.	
User: Could I get the address, phone number, and postcode of Yu Garden? **Sys:** Phone is 01223 248882, address and postcode are 529 Newmarket Road Fen Ditton C.B 5, 8 P.A	**User:** Could I get the address, phone number, and postcode of Yu Garden? **Sys:** Phone is 01223 248882, address and postcode are 529 Newmarket Road Fen Ditton C.B 5, 8 P.A
User: Thank you. That is all the information I needed. Bye bye!	**User:** Thank you. That is all the information I needed. Bye bye!

Figure 1: On the left side: a non proactive sample dialog from the WoZ dataset. On the right side: a proactive version of the same dialogue (arranged by us). Turns are separated by dotted-lines; proactive contributions of the system are in *italic*.

of conversational agents. A more detailed analysis on proactivity in existing task-oriented datasets is presented in Section 3).

Figure 1, left side, shows a typical task-oriented dialogue from the WoZ dataset (Wen et al., 2017), collected through Wizard of Oz and focused on booking restaurants. Here we notice the behaviour of the system in case of a failure situation (i.e., no match with the user needs), at turn 1: *No, there are no Eritrean restaurants in town. Would you like a different restaurant?*, and compare it with a proactive version of the same dialogue, arranged by us, showed in Figure 1, right side. In this proactive version the systems responds: *No, there are no Eritrean restaurants in town. Most of the restaurants are Chinese, particularly in the East area of the town.*, offering to the user potentially useful domain information to reorient the search. At the next turn the user may accept or not the proactive contribution of the system. In case the user accepts it (as in Figure 1, right side), there is no need of a further question about food type and the area of the restaurant, and the system can directly move to propose a specific restaurant, saving one turn in the dialogue. In case the user does not accept the proactive contribution, the dialogue will follow the non-proactive path, and the system can implicitly assume a low match with the user needs for the facts involved in the proactive contributions, which will not be proposed to the user anymore. In both of the cases, we see that a proactive behaviour has brought the system to a more precise definition of the user needs (i.e., the belief state of the system).

In the paper we aim at a first step toward extending the current task-oriented framework to accommodate proactive behaviours. We investigate the following research questions: (i) which are the main features that affect proactivity in task-oriented dialogues? and (ii) can we model such features within the architecture of current dialogues systems? and (iii) how can we show the impact of proactivity on task-oriented dialogues, e.g., in terms of effectiveness and efficiency of the dialogues?. The code used for the simulation is made publicly available.[1]

2 Related Work

In this section we report works that are related to proactive dialogues in both task-oriented and open domain (i.e., chit-chat) dialogue systems.

2.1 Proactivity in Task-oriented Dialogues

Task-oriented dialogues, usually distinguished from so called chit-chat dialogues, cover a broad range of applications (Allen et al., 2001; Bohus and Rudnicky, 2009), including giving instructions,

[1] https://github.com/vevake/Proactive_dialog_simulation

question answering, executing commands, conversational search, recommendation dialogues, and also explanatory dialogues. In this paper we focus on task-oriented dialog systems whose aim is to help the uses to retrieve items (i.e., entities) in a certain domain of interest, through *conversational search* (Radlinski and Craswell, 2017). Such systems (Young et al., 2013; Bordes et al., 2017; Lei et al., 2018; Wen et al., 2017) are modelled to achieve the task by actively questioning the user for slot values in a knowledge base, and then to retrieve entity instances satisfying the user needs. We report some work on recommendation, dialogue initiative, and persuasion, which we believe is related to proactivity.

Recommendation dialogues. (Thompson et al., 2004) propose ADAPTIVE PLACE ADVISOR, an adaptive conversational recommendation system designed to help people to select an item from a Knowledge Base. (Sun and Zhang, 2018) propose a conversational recommender system based on user's past ratings and the current interactions to make personalized recommendations to the user. The cold-start problem for recommendations in task-oriented dialogue systems was investigated by (Christakopoulou et al., 2016), proposing a preference elicitation framework to learn the preferences of a new user in fewer interactions and then use preferences to make recommendations. (Yoshino and Kawahara, 2015) propose a news navigation system that proactively presents users with information by tracking on the user focus.

Mixed initiative dialogues. Collaboration is particularly investigated in the context of mixed initiative dialogues, with the general goal of modelling dialogue policies for several situations (e.g., negotiation and argumentation dialogues). (Guinn, 1996) examined how to incorporate the initiative in a task-oriented dialogue system and proposed attaching a initiative level to each goal and as the goal of the dialogue changes, so does the dialogue initiative. (Yang and Heeman, 2007) investigated human-human conversations, proposing that initiative normally belongs to the speaker who initiates the task. They support the claim though simulations with two variations of mixed-initiative strategy. (Nouri and Traum, 2014) addressed mixed-initiative dialogue patterns in a negotiation dataset and the relationship between such patterns and the goal/outcome of the dialogue. They propose an annotation schema for dialogue initiative that allow to identify key features in a negotiation dialogues.

Persuasive dialogues. Unlike recommendation and dialogue initiative, where system and user cooperate to achieve a common goal, persuasive dialogues aims at convincing the user to achieve the systems goal (e.g., buying a target item or donating for charity). (Hiraoka et al., 2014) analyse human persuasion and the relation with user satisfaction. They propose an annotation scheme for persuasive dialogues and analyze a set of dialog acts by collecting a human-human dialog dataset for camera sales, showing that the main dialogue acts are information exchange and argumentation. (Hiraoka et al., 2013) propose a dialogue manager based on Bayesian network framework. They construct a knowledge-base that captures the relationship between various topics, which is then used to guide the user from topic to topic. (Wang et al., 2019) work on identifying correct persuasion strategies and analyse which strategies were effective based on user backgrounds.

While recommendation, user initiative and persuasion address collaborative phenomena in dialogue, they also tend to influence the user toward a specific target goal/item. Although the border between recommendation and proactivity is fuzzy, we intend proactivity as a general collaborative strategy aiming at improving the conversation quality and effectiveness, particularly in case of failure situations.

2.2 Proactivity in Open-Domain Dialogue Systems

Proactivity is also addressed for open domain dialogue systems due to high user expectations in open domain conversations. (Li et al., 2016) propose STALEMATEBREAKER, a system that can proactively introduce new content during a conversation, when a stalemate situation (e.g., "..." or "Errr", etc.) in the conversation is detected. The new content is dependent on the context of the conversation and is retrieved from a knowledge graph. (Yan and Zhao, 2018) address proactivity combining the response generation and a suggestion generation, and proposing a hybrid approach that provides a smart response (in addition to the system response), which the user can use as the input for the next turn. While this approach is feasible for open domain systems, for task-oriented dialogue systems such as smart responses, this is inefficient because of

the nature of the task. (Wu et al., 2019) propose a view on proactivity where the system is actively leading the conversation to achieve a specific goal. This requires the system to generate plans over a knowledge graph.

3 Proactivity in Existing Task-oriented Datasets

Modeling proactivity for task-oriented dialogues requires the system be aware of the content of domain and be able to provide responses that help the user to better achieve his/her goals. In general, a proactive system should be guided by two strategies: (i) try to avoid failure situations, providing the user with domain information before the point of failure; (ii) in case a failure is reached, use proactivity to recover the dialogue. In this Section we are interested to analyse whether the two strategies mentioned above are represented in current available datasets.

Existing datasets for task-oriented dialogue systems are widely collected using a Wizard-of-Oz (WoZ) approach. According to the WoZ methodology dialogues are collected following a script that is provided both to the user, who believes to interact with a system, and to the wizard, a person who simulates the behavior of a system. Typically, the wizard extracts information to be provided to the user searching in a domain Knowledge Base (e.g., restaurants in a town). However, since the dialogue follows a pre-defined script with a description of the actions that the wizard can take, if proactive behaviours are not included in the script, the responses of the Wizard will be quite standard, resulting in a general lack of proactivity. To support this intuition, we analyse two of the most widely used multiturn datasets for task-oriented dialogue systems, namely WoZ (Wen et al., 2017; Mrkšić et al., 2017) and MultiWoZ (Budzianowski et al., 2018), searching for system proactivity. In both the datasets proactivity is not marked in any way (i.e., there is no a specific dialogue act for it). In order to analyse proactivity we focused on failure situations, where there is no match between the constraints posed by the user and the content of the system domain (see example reported in Figure 1). These situations are both easy to be individuated in the dialogue corpus through simple patterns (e.g., *There are no...*), and, according to our expectations, are those where the Wizard should be mostly proactive to help the user. As a consequence, we assume that restricting our search to those situations can give a good approximation of proactivity in the corpus.

In the WoZ dataset we found over 200 failure situations, i.e., turns where the system was not able to find any matching with respect to the user constraints. In all such turns, except four of them, the response of the Wizard was a non proactive behaviour of the type "*There are no <SLOT-VALUE> restaurants in town. Would you like a different restaurant?*" where <SLOT-VALUE> corresponds to a combination of slot values requested by the user (e.g., Eritrean, as in our example in Figure 1). We can conclude that proactivity was not a relevant design criteria for the WoZ dataset.

As for the MultiWoZ dataset, we have individuated nearly 3000 turns where the user constraints do not yield any matching result. Out of them, we noticed around 10% of turns providing a proactive response to the user, such as "*There are no hotels that fit your criteria in the <AREA>, but there are two Guesthouses. Would you like to book one of those?*". Such responses in MultiWoZ were explicitly mentioned in the dialog task description of the Wizard, with the goal of modeling more realistic conversations (Budzianowski et al., 2018). We conclude that in MultiWoZ proactivity, at least in case of search failure, has been considered by design. However, we think that: (i) the amount of proactivity is still under represented, and (ii) maybe more important, being not annotated in any way, there is virtually no difference between proactive and non proactive behaviours.

4 Modeling Proactivity

In this section first we introduce the relevant background about modeling task-oriented dialogues, and then we attempt to extend such framework to integrate proactivity.

4.1 Task-oriented Dialogue Framework

According to most of the recent literature (Budzianowski et al., 2018; Bordes et al., 2017; Mrkšić et al., 2017), we assume a task-oriented dialogue between a system and an user, composed of a sequence of turns $\{t_1, t_2, ...t_n\}$. The goal of the dialogue is to retrieve a set of entities (possible empty) in a domain knowledge base (KB), which satisfy the user needs. A domain ontology O provides a schema for the KB, and typically represents entities (e.g., RESTAURANT, HOTEL, MOVIE) according to a pre-defined set of slots S

(e.g., FOOD, AREA, PRICE, for the RESTAURANT domain), and values that a certain slot can assume (e.g., EXPENSIVE, MODERATE and CHEAP, for the slot PRICE). On the base of the entities defined in the domain ontology, the application knowledge base, KB, is then populated with instances of such entities. As in most of the literature, we distinguish *informable slots*, which the user can use to constraint the search (e.g., AREA), and *requestable slots* (e.g., PHONENUMBER), whose values can be asked only when a certain entity has been retrieved. We say that the *entity space* of a domain ontology O is the combinatorial product of the slot values defined for each informable slot in O.

$$Entity\ Space_{O_d}(E) = \prod_{i=1}^{|S|} |V_i|$$

where $|V_i|$ is the number of possible values for an informable slot S_i in a given domain ontology O_d. At each turn in the dialogue, both the user and the system may refer to facts in the KB, the user with the goal of retrieving entities matching his/her needs, and the system to pro-actively propose entities that can help the user to achieve the dialogue goals.

Finally, in such task-oriented systems a dialogue state tracker (DST) maintains a distribution over the dialogue states based on the dialogue history. A dialogue state d_i for a turn t_i is typically represented as a set of slot-value pairs, such as {PRICE=MODERATE, FOOD=ITALIAN}, meaning that at t_i the system assumes that the user is looking for an Italian restaurant with a moderate price.

4.2 Proactive Units

We define a proactive behaviour as any information that: (i) is introduced by the system; (ii) was not previously introduced in the dialogue by the user; and (iii) it is assumed to be relevant to achieve the user needs. We call a piece of proactive content *proactive unit*. Intuitively, a proactive unit pu_i relative to a certain dialogue state d_i is a sort of guess of the user need given the information available at turn t_i. For instance, assuming
d_i= {AREA=CENTER}, then both:
{AREA=CENTER, FOOD=ITALIAN} and
{AREA=CENTER, FOOD=ITALIAN,
PRICE=MODERATE}
are possible proactive units for d_i.

Although in principle there might be different ways to identify a proactive unit during a dialogue, in this paper we focus on proactive units generated through *completion* of the dialogue state. Given a dialogue state $d_i = \{s_1 = v_1, ...s_n = v_n\}$, a completion for d_i is defined as an extension of d_i with slot-value pairs defined in the Ontology O, with the constraint that only one slot-value pair for each slot is admitted. Notice that, according to the task-oriented framework we adopted (Budzianowski et al., 2018), a dialogue state can only contain informable slots (see Section 4.1) and we apply the same constraint to its completions. This means that proactive units only contain informable slots, and that suggestions containing requestable slots (e.g., "Bella Napoli offers a large variety of pizzas") are not allowed. This is the way we capture the distinction between proactivity and recommendation mentioned in the Section 1.

Depending on the complexity of the domain ontology (i.e., number of slots and slot values), the number of proactive-units generated through completion for a dialogue state may be high, and it becomes useful the capacity to rank the units according to their relevancy, so that the more relevant are selected as proactive suggestions to the user. As a first approximation to a relevance function, we consider the probability that the entities described in the proactive units are actually present in the domain KB. Intuitively, the more the chances that a proactive unit have occurrences in the KB, the easier for the dialogue system will be to accomplish the user goals. To compute the probability of a proactive unit we consider the fraction of KB entities that matches the slot-value pairs out of the KB entities that match the dialogue state. More precisely:

$$P(pu_i|d_i) = \frac{\#matches(pu_i)}{\#matches(d_i)}$$

where the number of matches corresponds to the number of entities (i.e., instances) in the KB that satisfy all the slot-value pairs in pu_i or d_i.

For instance, if the user asks for Italian restaurants (d_i={FOOD=ITALIAN}), the following could be a partial rank of the proactive units that complete the information in the dialogue state:

([FOOD=ITALIAN], [AREA=CENTRE], $p = 0.8$)
([FOOD=ITALIAN], [AREA=SOUTH], $p = 0.2$)

assumed that 80% of the ITALIAN restaurants in the KB are in CENTRE, while 20% are in the SOUTH). According to this ranking, the first proactive unit

will have more chances to be selected in that dialogue turn.

Notice that the definition of completion provided above is constrained by the slots defined in the domain ontology, and does not consider other possible features, such as those related to the user identity (e.g., gender, age), nor situational features (e.g., time of the day).

4.3 System Proactivity

Having now clear our definition of proactive units, a first aspect that needs to be modelled is the attitude of the system to inject such proactive units during the dialogue. Based on the fact that proactive units are "completions" of a dialogue state, we consider those turns in the dialogue involved in the search through informable slots, and exclude turns where the choice is already made and the user asks additional information on requestable slots (e.g., telephone number of the selected restaurant).

Intuitively, at each (search) turn the system has to take two decisions: first, whether to be proactive or not, and, second, in case the system decides to be proactive, the amount of proactive units to be displayed to the user. We regulate the two decisions using two parameters. a is a value in the range [0,1] that defines the apriori attitude of the system to be proactive. It is interpreted as the proportion of turns in the dialogue that show proactivity. For instance, $a = 0.5$ means that 50% of the turns in the dialogue are expected to be proactive. A second parameter is k, indicating the average number of proactive units to be used by the system at each proactive turn. k ranges from 1 (i.e., an average of one proactive unit) to $|E|$, the cardinality of the entity space (see Section 4.1) defined over the domain ontology O. In practice, k is typically very small, as high numbers of proactive units will not be cognitively accepted by users.

The model we propose defines a and k as average behaviours of the system, and it is agnostic about how the average is actually achieved. However, in most situations it seem reasonable that both a and k distribute, respectively their proactive turns and proactive units, quite uniformly through the dialogue, avoiding dialogues where, e.g., all proactive turns are at the end, or all the proactive units are used in a single turn. In our experiments (Section 6) we adopt a similar strategy.

4.4 User Influenceability

A second aspect that needs to be considered while modeling proactive dialogues is the propensity of the user to accept the system suggestions. Intuitively, in order to produce the expected benefit, the user has to somehow "accept" the proactive content provided by the system. Remind that we defined proactivity as a neutral behaviour, i.e., general information about the domain of the task (e.g., "Most Italian restaurants are in the city center"), rather than specific recommendations (e.g., "Bella Napoli offers a large variety of pizzas"). While acceptance may depend on several reasons (e.g., quality of the proactive content), we approximate the propensity of the user to accept a proactive information, i.e., *user influenceability*, with a probability value $[0, 1]$, where 0 corresponds to a user that never accepts the system information (i.e., a non influenceable user), and 1 to a user that always accepts proactive content (i.e., a very influenceable user). For instance, a user influenceability set to 0.5 means that 50% of the proactive content provided by the system will be "accepted" by the user, while the remaining 50% will be ignored.

4.5 Complexity of the Application Domain

A third aspect that may affect proactivity is the complexity of the application domain. The intuition is that solving a conversational search task (e.g., retrieving a restaurant) in a complex domain might require more interactions than solving the same task in a less complex domain. According to this intuition, the more the domain complexity, the more proactivity is necessary to help the user to make efficient dialogues. We define domain complexity considering the number of entities defined in the domain ontology, their slots and their slot values. Practically, domain complexity is approximated with the domain entity space, defined in Section 4.1, and calculated as the combinatorial product of the slot value sets, for each slot of the domain.

4.6 Failure Situations

The last aspect that we aim to capture is that proactivity is particularly useful to recover failure situations, when the user needs do not match with any actual content of the domain KB, and some form of "adjustment" of the initial needs is typically attempted in order to successfully conclude the dialogue. This is the situation shown in Figure

1, where initially the user is looking for an Eritrean restaurant, and, after being informed that there are no such restaurants (i.e., a failure), he/she makes use of the system information and modifies the initial goals. Intuitively, the more failure situations happen in a dialogue, the more the role of system proactivity, and the more the expected benefits in term of task accomplishment.

We denote the probability of a failure situation as *failure expectation* and estimate it, given a certain application KB, considering the prior probability that a certain user need (e.g., a restaurant with certain characteristics) actually occurs in the domain KB, and average such probability over all the potential user needs (i.e., the entity space defined over the domain ontology O).

$$\begin{array}{c} Failure \\ expectation \end{array} = 1 - \frac{1}{|E|} \sum_{i=1}^{|E|} \begin{cases} 1 \text{ if } E_i \text{ in } KB \\ 0 \text{ else} \end{cases}$$

where E is the entity space. The more the failure expectation (e.g., "we do not have such restaurant"), the higher the need to help the user with proactivity. We practically estimate the probability of a user need to occur in the KB with the number of entity instances in the KB: the more the instances, the more the probability that the user need can be found in the KB.

Note that the size of the KB and the domain complexity described in section 4.5 are independent: a domain can be very complex (i.e., many slots and many slot values) although very poorly populated (i.e., few instances), and vice versa.

5 Simulating Proactivity

As there are no available datasets that can be used to test our model of proactive dialogues, we base our investigation on dialogues generated through simulation. We used an existing multi-domain dialog generator, *SimDial*[2] (Zhao and Eskenazi, 2018), and extended it by incorporating both the system proactive behaviour and user influenceability, as defined in Section 4. SimDial is based on a set of *domain specifications* describing the ontology of the dialogue task and a set of *complexity specification* that describe dialogue policies. As for domain we focus on restaurant booking defining its possible slots and values, while for dialogue policies we used the default specifications provided by SimDial.

[2]https://github.com/snakeztc/SimDial

5.1 Domain KB and Proactive Units

Our dialogue simulator is based on a default domain ontology for restaurants, consisting of three informable slots (i.e., AREA, PRICE, FOOD) with, respectively, 12, 10, and 3 pre-defined values, and two requestable slots (i.e., PARKING, OPEN). We also populated a default KB with 200 instances, randomly using the set of pre-defined slots.

Then, we extracted all the proactive units (see Section 4.2) from the default Ontology, considering both single and combinations of multiple slot-value and pairs, and then we computed their probability distribution in the KB. Such proactive units are stored in a repository called *KB-stats*, and made available as possible proactive information that the system can choose to convey to the user. The *KB-stats* is re-computed for each different configuration of the KB of our experiments.

5.2 Proactivity Module

The proactivity module that we incorporate to SimDial has access to *KB-stats*, which contains information about proactive units for the domain, as defined in Section 5.1. In order to keep the efficiency of the simulator under control, the content of the *KB-stats* is pre-computed. Specifically, we pre-compute the complete set of proactive units (see Section 4.2) considering the slots and the slot values in the ontology, as well as the probability of each proactive units on the base of the entities of the populated KB. We use the *KB-stats* to efficiently generate the pro-active-units for a certain dialogue state. The query to the *KB-stats* is the current dialog-state and the result is the list of proactive units that "complete" the dialogue state with their probability, according to relevancy definition provided in Section 4.2. For example, if the current dialog state is FOOD=ITALIAN, the proactive information could be (FOOD=ITALIAN, AREA=WEST, 0.8), which informs the system that 80% of the *Italian* restaurants are in the *west*.

5.3 Dialog Agent

The SimDial dialog-agent has a default policy which consists to request the user-agent for all the informable slots and, after that, to perform a domain query to retrieve entities from the KB. The order in which the informable slots are requested to the user-agent is randomized. This is the non proactive policy of SimDial.

The dialog-agent with the proactive policy re-

quests the informable slot to the user-agent, as in the default policy, and, in addition, provides the proactive information (i.e., the proactive units) that may help the user to achieve his goal. To convey the proactive information, we have defined proactive templates, which are used to transform the abstract information of the proactive units into a natural language utterance.

In particular, we define two kinds of templates that show proactive information: i) the first template indicates that all the proactive units for the current dialogue state lie in a single search space (e.g., All *Italian* restaurants are in *Palo Alto* and are *expensive*); ii) the second template indicates that most of the proactive units for the current dialogue state lie in a single search space (e.g., Most *Italian* restaurants are in *Palo Alto*). For example as seen in Figure 2, the system at turn 7 requests the user-agent for the PRICE slot and then provides proactive information that all restaurants are cheap. The selection of the template is carried on based on the proactive information retrieved from the *KB-stats*. If the selected proactive unit for the current dialog state has probability 1.0, then the first template is chosen otherwise second template is preferred.

We also added an additional policy to the dialog agent: if, for a particular dialog state, there is no proactive information retrieved, then it implies that there are no entities in the KB matching the dialogue state. In this case, instead of the dialog-agent continuing to query the user-agent for additional slots, we directly inform the user-agent to change his goal, as there exists no result for the current goal.

5.4 User agent

The default SimDial user-agent responds to the slot requested by the dialog-agent with a slot value based on his current goal. When the dialog-agent returns a restaurant match for the user-agent's goal, the user-agent may ask for any additional information on the available requestable slots (i.e., PARKING and OPEN).

In case of a proactive dialogue agent, the user-agent receives an additional proactive information at each turn, and has the option to either accept or ignore this suggestion. We simulate this user behaviour through the user influenceability probability value, as defined in Section 4.4.

An interesting situation happens when the selected proactive unit contains multiple slots (e.g., Most *Italian* restaurants are in *Palo Alto* and are *expensive*). In this case we designed a policy that allows the user-agent both to accept the entire proactive unit, or just a part of it. This policy, partial acceptance, in the current version is not explicit as a dialogue act, rather it is implemented by the user-agent repeating the slot value suggested by the dialog-agent. The reason for that is due to the limitations of the natural language expressions that can be handled by the simulator.

Finally, the maximum number of turns in a dialogue is set to 100 and, if the user-agent can not achieve its goal in 100 turns, the user-agent decides to quit the dialogue.

5.5 Evaluating Proactive Dialogues

The working hypothesis of this paper is that proactive behaviors help to improve both the *effectiveness* of the dialogue, i.e., the capacity of the user to accomplish his goals, and the dialogue *efficiency*, i.e., accomplishing the task with shorter dialogues. In our simulation framework we say that a task has been accomplished when a match with the user needs is found within a certain number of turns (fixed to 100), and we measure efficiency with the number of turns of the dialogue.

5.6 Limits of the Current Simulator

The current implementation of the proactive simulator has a number of limitations, which affect our capacity to carry on simulations. The most relevant is that SimDial allows for system initiative dialogues only. Secondly, it is not trivial to integrate a proactive response within the simulator dialog policy. As an example, it happens that while the proactive response for the current dialog state shows that there is no match in the KB, the system would still continue to ask for further informable slot values from user before it makes the KB query.

An example of a simulated proactive dialogue is shown in Figure 2. Although with the above limitations, the current simulator allowed to run a number of experiments, reported in the next Section.

6 Experiments and Results

In this section we describe the dialogue configurations that we have simulated, we report the obtained results, and finally we discuss our achievements.

6.1 Experimental Setting

We simulated task-oriented proactive dialogues for restaurant booking. The default configuration is

Sys: Welcome to restaurant recommendation system. How can I help?

User: Hi. I need a restaurant.
Sys: What type of restaurant?. Most of the restaurants are Korean

User: Chinese food.
Sys: Which city are you interested in?. Most of the Chinese restaurants are in Austin and are cheap.

User: At San Jose.
Sys: What's the average price? All Chinese restaurants in San Jose are in cheap pricerange.

User: cheap price.
Sys: Restaurant 62 is a good choice. Are you happy about my answer?

User: No more questions. Thank you.
Sys: See you next time..

Figure 2: A simulated dialog using the simulator for the restaurant domain.

as described in Section 5.1. In addition, we simulated four levels of domain complexity (see Section 3.5) varying the number of slot values for the three slots of the ontology. As for the size of the KB, we simulated six configurations (including the default), ranging from 50 to 1000 instances. For each configuration the set of proactive units included in the *KB-stats* has been fully recalculated.

As for system proactivity, we set the average amount of proactive units per turn k to 1 and vary the system proactive attitude a in the range $[0, 1]$. User influenceability has been simulated considering ten values in the probability range $[0, 1]$ with increments of 0.1.

The estimation of a failure expectation, see Section 4.6, given that the simulated KB has been populated randomly, the failure expectation has been simply calculated considering the KB size for a fixed ontology: the bigger the KB the less failure expectations are expected.

Dialogue effectiveness is calculated on the number of turns of the dialogue. A simulated dialogue stops either when a restaurant is found, or when a maximum of 100 turns is reached. Finally, for each experimented configuration, we simulate 1000 dialogues.

6.2 Results of the Simulations

The goal of the simulation is to highlight the dependencies between the proactivity aspects described in Section 4. We are interested to test how dialogue effectiveness changes as a function of system

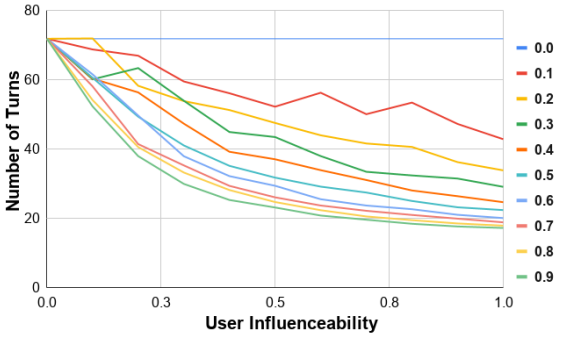


Figure 3: Dialogue effectiveness w.r.t the system proactivity attitude and the user influenceability. Numbers in legend indicate the system proactivity rate.

proactive attitude, user influenceability, domain complexity and failure expectations.

Dialogue effectiveness as a function of user influenceability and proactivity attitude. In this experiment we keep constant the ontology and the KB (using the default configurations), and we vary user influenceability and system proactivity attitude. Figure 3 shows for each proactivity attitude against the user influenceability values in the range $[0, 1]$. We notice that dialogue effectiveness increases (i.e number of turns decreases) with the increase of both proactive attitude and user influenceability. As expected, it can be noticed that a highly influenceable user (probability 1) results in the least number of turns in the conversation. While such highly influenceable users are quite unrealistic, and maybe undesired behaviour, we can infer that a moderately influenceable user (probability 0.5), would nevertheless result in a considerable reduction of unwanted interactions, a situation that helps the user to achieve the dialogue goals more effectively.

Dialogue effectiveness as a function of failure expectation and user influenceability. In this experiment we keep constant the ontology (using the default configuration) and the proactivity attitude (set to 1), and we vary the size of the KB and the user influenceability. Figure 4 shows the effectiveness of the proactive attitude over different failure expectation values (i.e., KB size). We can infer that, in case of a small KB, a user would struggle to find the desired results (i.e., 24 dialogue turns on average), as there is high probability that his/her constraints do not match with the KB instances. However, we notice that the number of turns considerably decreases even if the user con-

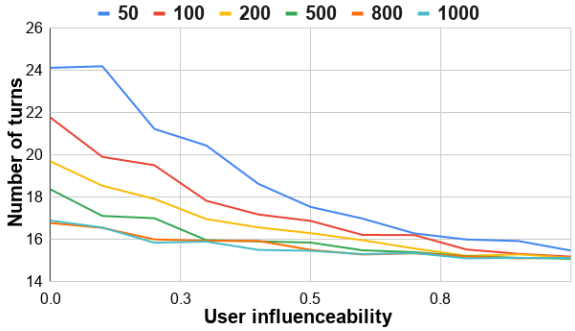


Figure 4: Dialogue effectiveness w.r.t the complexity of domain KB and the user influenceability. Values in the legend indicate the number of instances populated in the KB.

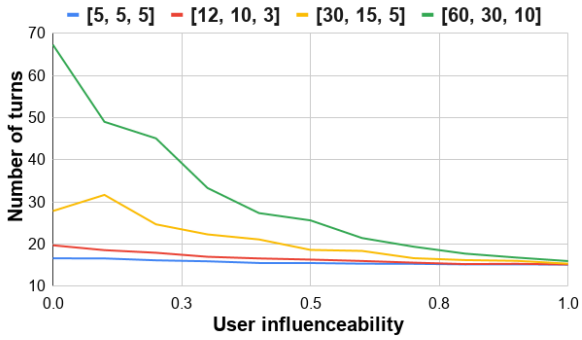


Figure 5: Dialogue effectiveness w.r.t the complexity of domain ontology and the user influenceability. Numbers in legend indicate the number of values for each of the three informable slots.

siders only half of the system proactive responses (i.e., users influenceability of 0.5). This is promising particularly for domains with a very small KB, although with a range of possible user constraints.

Dialogue effectiveness as a function of domain complexity and user influenceability. In this experiment we keep constant the KB size (using the default configuration of the KB with 200 instances) and the proactivity attitude (set to 1.0), and we vary the ontology complexity (varying the number of values for each slot), and the user influenceability.

We recall that for certain domains the number of values for a given slot could be unbounded (e.g., *time of departure*) and it could be exhausting for a user to find the KB result that matches his/her constraints without having any feedback or proactive behaviour from the system. Figure 5 shows that increasing the domain complexity, the number of turns necessary to achieve a goal increases, and this number can be drastically reduced by following a proactive behaviour.

Both proactive attitude and the user influenceability have the same effect on the dialogue effectiveness: the first generates proactive units, while the second filters them out, both with a certain probability. For this reason we do not present the corresponding experiments.

6.3 Discussion

In this section we discuss the two research questions at the core of the paper, in the light of the simulations that we carried on.

Which are the features that affect a proactive behaviour in task-oriented dialogues? The aspects proposed in Section 4 proactive unit, system proactivity rate, user influenceability and complexity of the application domain are all relevant for describing proactive dialogues. Particularly, the notion of proactive unit has been shown to be essential to make operative our simulation of proactivity.

How can we model such features within the architecture of current dialogue systems? While the current modeling of task-oriented dialogues provides a general framework, specific extensions were necessary to model our intuitions about the attitude of the user (i.e., influenceability) and the role of the system (i.e., proactivity). However, simulating proactivity revealed itself as more complex than initially expected, and, due to the issues described in Section 5.6, we had to adopt quite simplistic solutions.

How can we show the impact of proactivity on task-oriented dialogues, in terms of effectiveness and efficiency of the dialogues? Although we are aware that an effective evaluation should involve human judgements, the simulations that we conducted allow to show that the proactivity attitude of the system and the user influenceability have similar effect on the dialogue, and that they act as multipliers of the dialogue effectiveness. On the other side, failure expectation and domain complexity capture orthogonal properties of proactivity. Through our simulations we obtained quantitative evidences that small KBs and complex ontologies are the main motivations behind the need of proactivity. This conclusion, together with our predictive data, may have relevant impact on the practical construction of proactive dialogue systems.

7 Conclusion

We propose to extend the current task-oriented framework to manage proactive behaviours of the system. We analysed four aspects: (i) the degree of proactivity provided by the system during a dialogue; (ii) the propensity of the user to be influenced by the system proactivity; (iii) the complexity of the application domain, and (iv) the failure expectation of a dialogue goal. We conducted a series of simulation experiments, which allowed to investigate our intuitions about the role of the four proactivity aspects with respect to the effectiveness of the dialogue.

As for future work, we are currently working along three directions. First we intend to further extend SimDial, to be able to simulate more complex proactive situations: particularly mixed initiative dialogues and more complex domain ontologies. Second, we will conduct a comparison of the simulated dialogues based on human judgements, in order to have a qualitative evaluation. The third research direction is about modeling proactive dialogue policies, to be used within current neural dialogue architectures.

References

James F. Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the 6th International Conference on Intelligent User Interfaces, IUI 2001, Santa Fe, NM, USA, January 14-17, 2001*, pages 1–8. ACM.

Dan Bohus and Alexander I. Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. *Comput. Speech Lang.*, 23(3):332–361.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *ICLR*. OpenReview.net.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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, Brussels, Belgium. Association for Computational Linguistics.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. 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 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 815–824, New York, NY, USA. Association for Computing Machinery.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, pages 12–16.

BJ Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Curry I. Guinn. 1996. Mechanisms for mixed-initiative human-computer collaborative discourse. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 278–285, Santa Cruz, California, USA. Association for Computational Linguistics.

T. Hiraoka, Y. Yamauchi, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. 2013. Dialogue management for leading the conversation in persuasive dialogue systems. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 114–119.

Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Construction and analysis of a persuasive dialogue corpus. In *5th International Workshop on Spoken Dialog Systems (IWSDS)*, Napa, California, USA.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1437–1447.

Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2845–2851. IJCAI/AAAI Press.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788. Association for Computational Linguistics.

Elnaz Nouri and David Traum. 2014. Initiative taking in negotiation. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 186–193, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of spoken language systems: the ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, pages 117–126. ACM.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, page 235–244, New York, NY, USA. Association for Computing Machinery.

Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley. 2004. A personalized system for conversational recommendations. *J. Artif. Int. Res.*, 21(1):393–428.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Rui Yan and Dongyan Zhao. 2018. Smarter response with proactive suggestion: A new generative neural conversation paradigm. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4525–4531. AAAI Press.

Fan Yang and Peter A Heeman. 2007. Exploring initiative strategies using computer simulation. In *Eighth Annual Conference of the International Speech Communication Association*.

Koichiro Yoshino and Tatsuya Kawahara. 2015. *News Navigation System Based on Proactive Dialogue Strategy*, pages 15–25. Springer International Publishing, Cham.

S. Young, M. Gašić, B. Thomson, and J. D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

On Laughter and Forgetting and Reconversing: A neurologically-inspired model of conversational context

Jonathan Ginzburg
Université de Paris
Laboratoire de Linguistique Formelle
Laboratoire d'Excellence LabEx-EFL
Institut Universitaire de France
yonatan.ginzburg@u-paris.fr

Andy Lücking
Université de Paris
Laboratoire de Linguistique Formelle
Goethe-Universität Frankfurt
Text Technology Lab
luecking@em.uni-frankfurt.de

Abstract

We argue that phenomena from both dementive and neurotypical individuals motivate the need for a synthesis between existing formal semantic theories of dialogue context and theories of memory and emotion. We sketch here such a synthesis: we embed certain very transient aspects of dialogical semantic states as components of a Multicomponent Working Memory (Baddeley, 2012) style working memory and we also sketch how to embed dialogues globally within Long Term Memory. We exemplify how the emergent framework can account for the conversational phenomena.

1 Introduction

Consider the exchange in (1) between nurses and inhabitants of a care home suffering from dementia. The nurses are prompting the patients to complete well known Swedish proverbs. Among this cohort, studied by Lindholm (2008) laughter is commonly used to mark memory failure.

(1) a. NURSE: strain at a gnat and, (0.5 sec) PATIENT: (ah) (0.8 sec) *nothing else* [heh heh heh] NURSE: strain at a gnat and swallow a camel. (Lindholm, 2008, ex. (2))

b. NURSE1: no smoke, (1.9) PATIENT: (.h) without (1.7) heh heh heh NURSE2: fire ((whispers)) (Lindholm, 2008, ex. (3))

Note that cases such as (1) are not limited to people with dementia, as in the constructed (2)—here after a failure by the speaker to recollect the name or title of a particular individual, they reformulate and laugh about it:

(2) A: The ... whoever it is (laughs) decided against reopening the schools.

In order to develop an account of examples such as these, we need to synthesize at least three ingredients: a theory of dialogue meaning and coherence (to explain the coherence requirement of the nurse's and patient's responses, and the content of self-repairs and laughs), a theory of memory incorporating long-term and short-term/working memory (LTM, WM) distinctions, and a theory of emotion to explain certain of the effects brought about by laughter; indeed recent work has demonstrated the need for synthesizing emotion and dialogue given the complex dialogical meaning of signals such as laughter and frowning (Ginzburg et al., 2020). As we discuss in section 2, (Neuro)Psychological Memory theories help explain memory *failure* in terms of (i) trace deterioration caused, e.g., by lack of consolidation (Wixted, 2004), by time–dependent contextual drift (Sadeh and Pertzov, 2020), or by disease, or (ii) capacity limits—the episodic buffer, assumed to serve as a link to perception and to LTM and to hold multidimensional representations, is assumed to hold up to four chunks. However, since no semantics is assigned to traces and such theories lack notions of relevance, they cannot be used to explicate partial recollection and how patients react to their failure, as in the laughter in (1) and (2), or in the question posed in (3). For the latter, such theories can be used to explain the confusion evinced by the patient in terms of working memory capacity, as we explain below, but not the emergence of the clarification question:

(3) (Context: B (a dementia patient) is watching a concert on television featuring wind instruments while eating dessert consisting of cooked apples): How will the apples get through the pipes? (Citation-suppressed, 2019)

Dialogical theories of context (e.g., Ginzburg,

2012; Lascarides and Asher, 2009) have as their inspiration logical derivation architectures and they offer intricate notions of relevance and inference. However, their Achilles heel is the absence of forgetting and of explicit interface with LTM.[1] Forgetting has typically been viewed as a "performance" issue to be finessed at a future date. But as illustrated in (4a), it can play an important role in dialogues with dementive patients, and even with neurotypical participants.

(4) a. A: When will you come? B: Not for a few weeks. A: Yes. B: So good night. A: So will you come tomorrow for lunch? B: I'm not in town. A: Yes. B: So good night. (Citation-suppressed, 2019)

b. A: When will you come? B: Not for a few weeks. A: (absent mindedly) Yes. (10 minutes later) A: So, when are you coming? B: I told you! A: Sorry, I was multitasking.

A converse issue, showing the need for an explicit interface with LTM, is the phenomenon of resumed conversations, as in the constructed example (5):

(5) A: How can we solve the equation? B: I'll have to think about it, but now I have to run.
(3 days later) A: So? B: Right, yes,um I'd say just integrate three times and ...

Indeed, continued resumption is important as far as inner dialogue goes (on which see Kracht, 2010), which is particularly relevant for therapeutic genres, where apparently resolved issues can rearise indefinitely (as has been reported, for instance, for depression (Curry, 2014) and schizophrenia (Kennedy and Xyrichis, 2017)).

The rest of the paper is structured as follows: in section 2, we review literature on memory, emotion, and dialogue. We effect a synthesis in section 3— implement basic linguistic and non-verbal interactions including signs in a memory–oriented and felt way. We return to the above examples in section 4.

[1] Absence of forgetting is not to be conflated with lack of structure. Such theories have means to limit accessibility to antecedents, including structural (e.g., the right frontier constraint), stacking/partial ordering of contextual repositories such as QUD. But once a proposition becomes part of the common ground, however modelled, then it remains as such, unless eliminated by correction.

2 Background

2.1 Memory

Recent (neuro)psychological work Recent work on memory has offered evidence for various distinct kinds of long-term memory (LTM) systems and the brain areas they activate—much of it based on disassociation phenomena (Bastin et al., 2019). (6a,b) exemplify two distinct types of memory failure: partial recollection (details of an event) and loss of familiarity (with an entity):

(6) a. Inv: Who did you go with? Man: Went with you (wife). Wife: yeh me and who else? Man: I don't know. Wife: Trevor (son) Man: Aye Trevor. (From Shakespeare, 2013 extract 2)

b. Son: Who am I? Mother: I don't know. (Shakespeare, 2013 extract 23)

The LTM systems include: the relational episodic (hippocampus), entity (perirhinal and parahippocampal cortices), and the procedural (striatum) subsystems. With respect to working memory (WM) Baddeley (2012) summarizes a model, Multicomponent Working Memory (M-WM), that has been highly influential in the last 40 years: on this view, M-WM has four components and informational flow, as summarized in Fig. 1 (slightly simplified from Baddeley, 2012, Fig. 5). Initially, as described in (Baddeley, 1997), the framework had three components, the C(entral) E(xecutive), the phonological loop (phon-loop), and the visuo-spatial sketchpad (VSSP). The basic idea here is to split attentional control from temporary storage—the phonological loop and the visuo-spatial sketchpad correspond to separate verbal and visuo-spatial short-term systems, which are limited in capacity. The phonological loop has been argued to be necessary for new long-term phonological learning and to play a significant role in the initial stages of vocabulary acquisition. The CE serves four main functions (Baddeley, 2012, 14): 1. focus attention, 2. divide attention, 3. switching between tasks, and 4. interface with the LTM. 25 years after the initial 3 component model, an additional component was postulated, namely the *episodic buffer*. This is a buffer that can maintain information from several modalities that has been bound together by the central executive. (Baddeley, 2012, 17).

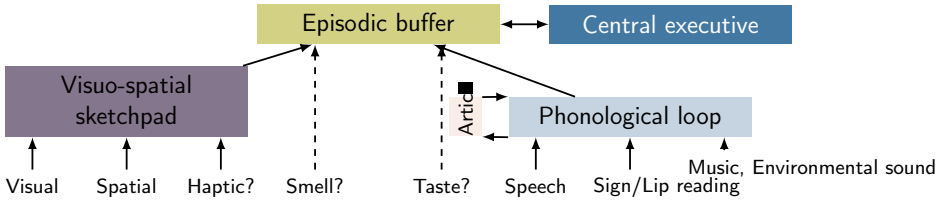


Figure 1: M-WM (Baddeley, 2012)

There are still a variety of open issues in this area:

- Baddeley's phon-loop is a (speech-based) short-term storage, while the episodic buffer is a chunk based mechanism. Cowan (e.g. Cowan, 2008) claims that there is not enough empirical evidence to differentiate two processing limiting systems; he rather argues for a single 'attentional system used both for processing and for storage' (*op. cit.*, p. 13) that operates on activated areas of the LTM. However, as Baddeley (2012, 20) remarks, Cowan's concerns pertain mainly to the CE and the episodic buffer (in M-MW's terms) so that there is a at least a terminological 'translation' between both accounts.

- Both Baddeley's episodic buffer and Cowan's focus of attention are chunk limited buffer stores (Cowan's focus of attention also acts as a temporal storage – see the previous item), and both models by and large agree on a capacity limit of four chunks (Baddeley, 2012, 15). In fact, Cowan (2001) argues in great detail for a limit of four chunks of short-term memory processes (though there still is some discussion, as the comments to Cowan's article show). However, a chunk may itself be a compound structure, for instance, in case of multimodal representations (i.e., bindings of various VSSP and possibly also phon-loop stimuli), so that there is some flexibility concerning the complexity of the chunks themselves. In any case, a short-term memory limit of four chunks seems to be a first capacity constraint for dynamic semantic theories, as desired in Sec. 1.

- Alternative proposals suggest that, in fact, the correct generalisation is that one can maintain only *one* temporally extended *event* or *epoch* in focal attention (McElree and Dosher, 2001).

2.2 Emotion

There are a variety of theories of emotion. Although important early work on facial gestures came from researchers espousing theories postulating a small number of *basic emotions* see e.g., Ekman et al. (1987), there has been little evidence to support the existence of physiological characteristics that instances of a single emotion share but that other emotions do not (Scherer and Ellgring, 2007; Barrett, 2017). A variety of approaches have emerged that avoid such an assumption. We mention here two influential approaches, *emotion constructivism* and approaches based on *appraisal*; both provide useful means for classifying laughables, smileables, sighables (the events triggering such signals) etc and the emotional episodes they give rise to.[2] Emotional episodes are viewed as arising from a categorisation process of the triggering event in terms of previous event 'exemplars' on the basis of a resemblance in terms of certain dimensions (Russell, 2003; Barrett, 2017). In the account of Russell (2003) these include *core affect* (a two dimensional matrix of (un)pleasantness and arousal), *affective quality*, and the object causally involved in the event. With respect to neural implementation, one might mention latency results from a recent extensive review of intracranial recordings of the human amygdala, one of the key junctures in affective processing (Murray et al., 2014; see Fig. 2 for an illustration). This suggests that the amygdala neuronal activity occurs in three latency windows: an early window subsumes effects respective to exogenous stimulus-driven affective processing of faces and emotion; an intermediate window consists of effects related to explicit attention to novel task-relevant stimuli, irrespective of sensory modality; whereas, the late window subsumes effects from tasks soliciting working memory, semantic pro-

[2]For discussion of the similarities and differences between these approaches, see Brosch (2013); Barrett (2017); Sander et al. (2018).

cessing, attentional focus and memorization during affective processing.[3]

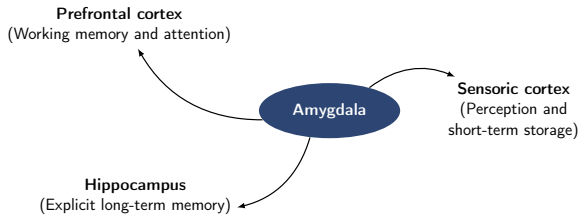


Figure 2: (Reproduced from LeDoux, 2001, 308) Regions of the amygdala project into various cortical regions. This includes projection into all stadiums of cortical sensoric processing, into the prefrontal cortex an to the hippocampus and its neighbouring cortical areas. Based on these projections, the amygdala influences current perception, imagination, attention, short-term memory, working memory, long-term memory, and various higher-order thought processes resting on the former.

2.3 Dialogue

We use KoS (Ginzburg, 2012) for modelling dialogue for two reasons: (1) as we explain below, it is already partially structured in memory terms (2) it is underpinned logically by Type Theory with Records (TTR) (Cooper, 2020), whose structured nature enables writing both memory trace construction rules and de-activation principles. In TTR propositions are identified with records of the type $\begin{bmatrix} \text{sit:Rec} \\ \text{sit-type:RecType} \end{bmatrix}$, involving a situation *sit* and a situation type *sit-type* and true iff sit:sit-type.

KoS as a theory of dialogue offers a theory of how utterances and other events change an individual's perspective of the public context, by developing an account of a highly structured class of states—*dialogue gameboards*. Dialogue gameboards track various aspects of the emerging context, including turn ownership, shared assumptions (FACTS), questions under discussion (QUD), the visual field, moves that are in the process of being or have been grounded (Pending, Moves) and MOOD—a weighted sum of appraisals. Here MOOD represents the publicly accessible emotional aspect of an agent that arises by publicly visible actions (such as non-verbal social signals), which can but need not diverge from the private emotional state.

$$DGBType := \begin{bmatrix} \text{spkr} & : Ind \\ \text{addr} & : Ind \\ \text{utt-time} & : Time \\ \text{c-utt} & : \text{addressing(spkr,addr,utt-time)} \\ \text{facts} & : Set(Prop) \\ \text{vis-sit} & = \begin{bmatrix} \text{fova} : Ind \vee Sit \end{bmatrix} : RecType \\ \text{pending} & : List(LocProp) \\ \text{moves} & : List(IllocProp) \\ \text{qud} & : poset(Question) \\ \text{Mood} & : Appraisal \end{bmatrix}$$

Such cognitive states can represent the results of locutionary, (7a,b), illocutionary updates, as in (7c,d), and emotion-based updates, such as (7e):

(7) 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, exemplified in section 3; 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.

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

c. 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.

d. 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).

e. Positive affect incrementation of Mood: given the LatestMove being an incongruity proposition by the speaker, the speaker

[3] An anonymous reviewer for SemDial asks whether we think that these windows correspond to the three dimensions that in RussellâĂŹs account underlie the categorization process, namely core affect(a two dimensional matrix of (un)pleasantness and arousal), affective quality, and the object causally involvedin the event. We think this is an interesting hypothesis, but do not have evidence to support/refute this currently.

increments the (positive) pleasantness recorded in Mood to an extent determined by the laughter's arousal value.

KoS provides a theory of meaning for highly context dependent elements such as non-sentential utterances (8a,b) and non-verbal social signals such as laughter (8c), which figure further below.

(8) a. yes $\mapsto$ p (p? is MaxQUD);

b. Dunno ('I don't know') $\mapsto$ $\neg$Know(spkr,MaxQUD)

c. Right $\mapsto$ Understand(A,u) (u is MaxPending, A current speaker);

(all Ginzburg, 2012)

d. Um $\mapsto$ Makes λxMeanNextUtt(spkr,Pending,x) MaxQUD (Ginzburg et al., 2014)

e. laugh: Given A as speaker, l as laughable event, τ as topos $\mapsto$ Assert(A, Incongruous(l, τ)) (Ginzburg et al., 2020)

3 Synthesizing Memory, Emotion, and Dialogue

3.1 Distributing the DGB between WM and LTM

In synthesizing KoS and M–WM, there are two main issues:

1. how to ensure that while the mechanisms we use for dialogue states are individual memory states, they represent records of *interaction*?

2. how to partition the various components of the DGB across WM and LTM?

With respect to the first issue, we will make a rather obvious move, namely view each conversation as an *episode*, one which gets initialized by the first move—typically a greeting—and concluded by the final move—often a (counter)-parting. Our conversations, then, constitute a particular class of episode, which DGBs provide structure for, so are subsumed within episodic memory:

(9) a. Conversation initiation: add the conversation as the next episode in LTM, with its initial state characterized by Init-State.

b. Conversation termination: if the speakers have *disengaged*—for instance by means of a parting exchange, a new episode starts.

With respect to the second issue, some aspects seem fairly straightforward: we can think of the Central Executive as mapping into an Agenda (as specified in work following Larsson, 2002), specifying the next action the inputs recorded in *Pending* (M-WM: the phonological loop) and *VisSit* (M-WM: the visuo-spatial sketchpad). The guiding intuition we will employ is that the episodic buffer needs to be capacity limited, which we will construe as pertaining in a broad sense to a single eventuality. How and whether one needs to capture the four chunk constraint remains an intriguing question we cannot resolve here. A potential starting point is the model of (Takac and Knott, 2016b) concerning how events and their participants (in monological discourse) are represented in WM. A key assumption of this model is that events are experienced through sequentially structured sensorimotor routines, and similarly for the event's participants. This yields a novel mechanism for binding representations of individuals to semantic roles such as AGENT and PATIENT and directly captures capacity constraints in terms of number of event argument roles; for more on the relationship between syntax and WM see (Knott, 2012). We do address capacity constraints in one respect. We separate the speaker and addressee information, assuming current speaker is within the episodic buffer, motivated in part by data from (Alberoni et al., 1992) that suggests that Alzheimer Disease patients have difficulty tracking who is speaking, with a limit being attained at four participants. We assume that the addressee is information that arrives from the VSSP.[4]

(10) EpisodicBuffer $=_{def}$

$$\begin{bmatrix} \text{spkr : Ind} \\ \text{u-t : Time} \\ \text{c-u : speaking (spkr,u-t)} \\ \text{MaxQUD : Question} \\ \text{LatestMove : LocProp} \\ \text{TopicalFact : Prop} \end{bmatrix}$$

We will restrict attention to three fields of LTM: the first is Episodic, which we assume has a field that tracks conversational episodes; we allow that

[4]The set up here is 2 person dialogue, nonetheless, for reasons of simplicity.

it might track other types of episodes, but do not consider how here. For a conversation in progress, it will record successive values of FACTS, QUD, MOVES, and Mood. In addition, LTM has two additional fields mentioned above: (i) Entities, which keeps track of individuals with whom a given agent has familiarity, and (ii) Procedures, which stores procedural knowledge:[5]

(11) a. LTM $=_{def}$

$$\begin{bmatrix} \text{Episodic} : \begin{bmatrix} \text{Conversational} : \text{list(LDGBType)} \end{bmatrix} \\ \text{Entities} : \text{set(RecType)} \\ \text{Procedures} : \text{set(topos)} \end{bmatrix}$$

b. LDGBType $=_{def}$

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

The partitioning of the DGB means that some conversational rules need to become more complex. We give two examples: incrementing QUD as a consequence of a query involves both updating the Episodic Buffer's MaxQUD, but also the LTM QUD. Similarly, FactUpdate/QUDDowndate requires updating TopicalFact and MaxQUD in the Episodic Buffer and the LTM FACTS and QUD:

(12) a. **Ask QUD-incrementation**:

$$\begin{bmatrix} \text{pre} : \begin{bmatrix} \text{q} : \text{Question} \\ \text{LatestMove} = \\ \text{Ask(spkr,addr,q)} : \text{IllocProp} \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{LTM.QUD} = \left\langle \text{q,pre.QUD} \right\rangle : \\ \text{poset(Question)} \\ \text{EB.MaxQUD} = \text{q} : \text{Question} \end{bmatrix} \end{bmatrix}$$

b. **Fact Update/ QUD Downdate:**

$$\begin{bmatrix} \text{pre} : \begin{bmatrix} \text{p} : \text{Prop} \\ \text{LatestMove} = \\ \text{Accept(spkr,addr,p)} : \text{IllocProp} \\ \text{LTM.QUD} = \left\langle \text{p?,pre.QUD} \right\rangle : \\ \text{poset(Question)} \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{FACTS} = \text{pre.FACTS} \cup \left\{ \text{p} \right\} : \\ \text{Set(Prop)} \\ \text{EB.TOPICALFACT} = \text{p} : \text{Prop} \\ \text{QUD} = \\ \text{NonResolve(pre.QUD,FACTS)} : \\ \quad \text{poset(Question)} \\ \text{MaxQUD} = \text{Max(QUD)} \end{bmatrix} \end{bmatrix}$$

A summary of the emerging picture is provided in Fig. 3.

We start with conversation initialization, focussing solely on LTM. If A and B have just started a conversation, their LTM will have the form in (13)

(13)

$$\begin{bmatrix} \text{Episodic.Conversational} = \\ \left\langle \begin{bmatrix} \text{x} = \text{A} \\ \text{y} = \text{B} \\ \text{participants} = \left\{ \text{A,B} \right\} \\ \text{Moves} = \left\langle \text{Greet(A,B)} \right\rangle \\ \text{qud} = \left\langle \right\rangle \\ \text{facts} = \text{cg1} \end{bmatrix}, \ldots \right\rangle \\ \text{Entities} = \text{entset} \\ \text{Procedures} = \text{procset} \end{bmatrix}$$

We now illustrate the interaction between the EB and LTM in this set up with the examples in (14):

(14) a. A: I see Jo. B: Right.

b. (A and B watching Vladimir Putin on tv walking in a hospital wearing a yellow Hazmat suit)[6] A: laughs. B: laughs.

Consider how B might process 'I see Jo': upon perception the locutionary proposition (15) updates the phonological loop—we abstract away for now from issues of lexical access. This leaves four contextual parameters to be resolved. The value for 'I' and for the utterance time are to be found in the episodic buffer, while we assume the VSSP to be the described situation; 'Jo' is found in LTM.Entities. This updated locutionary proposition then updates EB.LatestMove. Subsequently,

[5]For structure within episodic memory, see Takac and Knott (2016a) which describes a neural network model of episodic memory and its interfaces to entity representation in WM and LTM.

[6]From `https://abcnews.go.com/International/putin-dons-hazmat-suit-russia-admits-virus-numbers/story?id=69768452`, last accessed at 10th May 2020.

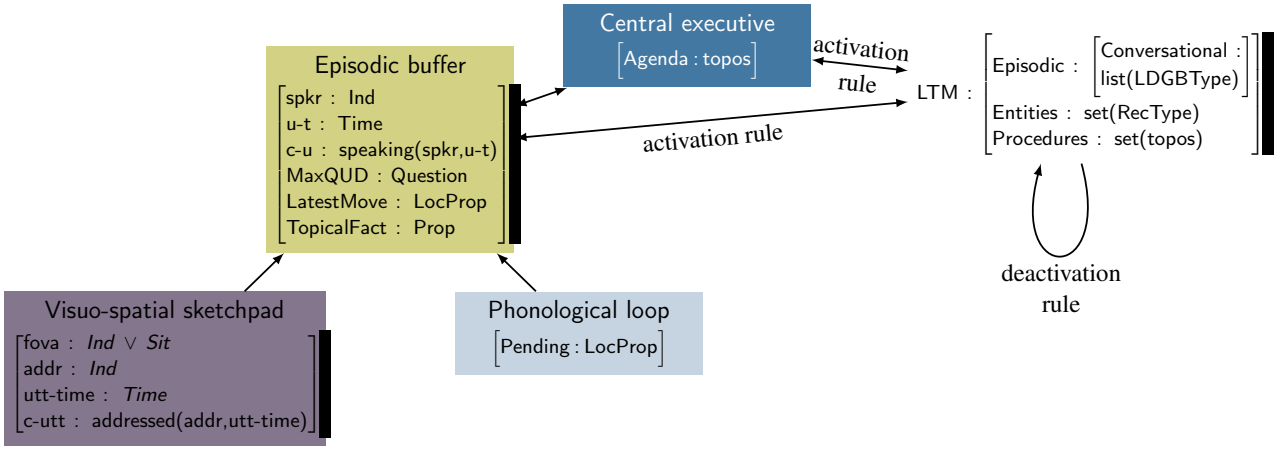


Figure 3: Fusing M-WM and DGB, and adding LTM.

CE—fed by LTM.Procedures—updates Agenda with the action Ground(LatestMove). This, in turn, gives rise to B's utterance whose contextual parameters are Pending.sit and B found in the VSSP:

(15)

$$\begin{bmatrix} \text{sit} = \begin{bmatrix} \text{phon} = \texttt{aysijo} \\ \text{cat} = \text{V[+fin,+root]} \\ \text{constits} = \left\{\text{ay,si,jo}\right\} \\ \text{dgb-params} = [] \\ \text{cont} = \begin{bmatrix} \text{sit} = \text{s0} \\ \text{sit-type} = \text{See(j,t0)} \end{bmatrix} \end{bmatrix} \\ \text{sit-type} = \begin{bmatrix} \text{PHON} : \texttt{I see Jo} \\ \text{CAT} = \text{V[+fin,+root]} : \text{syncat} \\ \text{constits} = \left\{\text{I, see, Jo}\right\} : \text{set(sign)} \\ \text{DGB-PARAMS} : \begin{bmatrix} \text{s0: SIT} \\ \text{spkr : IND} \\ \text{addr : IND t0: TIME} \\ \text{c2 : addressing(spkr,addr.t0)} \\ \text{j : IND} \\ \text{c3: Named(j,jo)} \end{bmatrix} \\ \text{cont} = \begin{bmatrix} \text{sit} = \text{s0} \\ \text{sit-type} = \text{See(spkr,j,t0)} \end{bmatrix} : \text{Prop} \end{bmatrix} \end{bmatrix}$$

Turning now to A and B sitting on couch watching Putin on tv in a Hazmat suit. A laughs. The laugh updates B's phonological loop with a locutionary proposition akin to (15) *mutatis mutandis*—the content is as given in (16). In this case the contextual parameters are A, the laughable l, and the topos τ. A and l can be instantiated from the EB and the VSSP, whereas τ involves a call on LTM.FACTS (e.g., the topos *Presidents wear formal suits*). Given this resolution, the VSSP can be appraised as pleasant, thereby updating her private Mood. Subsequently, CE—fed by LTM.procedures—updates Agenda with the action Accept(LatestMove). This, given B's Mood update, licenses B's laugh, which in turn triggers an update of her public Mood.

(16) Assert(A, Incongruous(l,τ))

We now consider a case that emphasizes one significant innovation of this set up, namely postulating that QUD is part of LTM and solely MaxQUD is part of the episodic buffer. Consider (17). Here the question (1) becomes EB.MaxQUD, but when (2) is asserted (2)? becomes EB.MaxQUD, whereas QUD now consists of both (1) and (2); after A's utterance in (3), (3a)? and (3b)? successively become EB.MaxQUD, whereas LTM.QUD consists of all four questions. Once B accepts (3), this updates both FACTS and TOPICALFACT with (3b) and makes (1) EB.MaxQUD as well as the sole member of LTM.QUD:

(17) a. A(1): Who's a good candidate?
B(2): Petra.
A(3): (3a) No, (3b) Pauline is.
B(4): OK.

3.2 Episodic and Entity activation and deterioration

We hypothesize that relational episodic traces, which we identify with TTR propositions (and potentially neurally implemented as explicated in Cooper, 2019) arise as a consequence of activation occurring during e.g., propositional acceptance. These, in turn, via projection operations on fields in the record type, construct entity traces.

Given any proposition in FACTS with x : IND fields, increment entities with record types for that individual. For instance (18a) will give rise to the entity in (18c), constructed from the type in (18b):

(18) a. A: What happened yesterday? B: A Mexican guy named Ramon killed Trotsky.

b. RamonM = $\begin{bmatrix} \text{x : Ind} \\ \text{c1 : mexican(x)} \\ \text{c2 : named('Ramon',x)} \end{bmatrix}$

c. $\begin{bmatrix} \text{z = RamonM : IND} \end{bmatrix}$

We can equally model memory deterioration operations: for neurotypicals occurring as result of underactivation, for dementives punctually and randomly as result of lesion or other mal-events. For relational episodic traces these involve elimination of fields from a record type constituting an event trace, for entity traces their downdating from Entities.

4 Laughter, memory, emotion: sample analyses

We now return to consider some of our initial examples.

4.1 Memory failure laughter

Consider (1a) repeated here as (19):

(19) NURSE: strain at a gnat and, (0.5 sec) PATIENT: (ah) (0.8 sec) *nothing else* [heh heh heh] NURSE: strain at a gnat and swallow a camel. (Lindholm, 2008, ex. (2))

The patient has in the phonological loop the locutionary proposition associated with the Nurse's utterance u_{N1}. Lexical (or collocational) access problems give rise to the issue *What word follows* u_{N1}*?* By articulating the filled pause the patient causes this question to become MaxQUD (see (8d) above). Given his continued collocational access failure, he responds to this question negatively. This failure makes the patient feel uncomfortable—laughter will help smooth the situation. Concretely, then, the laughable in this case is the collocational access failure which clashes with the topos in Agenda (When A asks a question B answers it), hence the incongruity.[7] This is an instance of laughter whose generation/resolution makes no appeal to episodic memory, and is therefore quite available to a dementia sufferer. We do not have here data on *failed* laughter triggers in dementia; see (Baumgartner and Renner, 2019) for some discussion of humour failure in dementia.

4.1.1 Stimuli merge

The CE is the 'attentional homunculus' of the M-WM. Among others, it is responsible for focusing or dividing attention (cf. Sec. 1). According to neuropsychological evidence, items in the object of attention (of mammals) are represented in the prefrontal cortex (Güntürkün, 2005). Initiating and maintaining items within attention is driven by dopamin release (Rose et al., 2010; Lohani et al., 2019). Sustaining activations in the prefrontal cortex with dopamin excitation keeps the focus on the items associated with the activation patterns. Dopamin innervation, triggered, e.g., by concurrent representation from the VSSP via the EB or the LTM, can lead to a shift in the focus of attention. If this shifts happens during other short-term processes such as topos rehearsal, an interference of focused items can be a consequence (for a model, see (Manohar et al., 2019)). Such a dopamin-driven interference can give rise to a stimuli merge like the pipe/apple blend of an eating and a music playing situations (both staged in the VSSP) reported in example (3) in Sec. 1, since 'results from work in schizophrenia, PD, and ADHD patients point to an abnormal DA transmission as being responsible for behavioral deficiencies in some learning and memory tasks that depend heavily on PFC function'[8] (Puig et al., 2014, 4).

4.1.2 Resumed conversations

Consider again example (5), repeated here as (20):

(20) A: How can we solve the equation? B: I'll have to think about it, but now I have to run.
(3 days later) A: So? /Where were we? B: Right, yes,um I'd say just integrate three times and ...

The initial conversation ends with a non-empty

[7] Another alternative, suggested to us by Alistair Knott, is that the incongruity is a direct consequence of the inability to complete the self-repair. This is a plausible alternative in that laughter often occurs in self-repair situations and would equally not make any appeal to episodic memory for resolution.

[8] Abbreviations: PD = *Parkinson's disease*, PFC = *prefrontal cortex*, ADHD = *attention deficit hyperactivity disorder*.

QUD. A's initial utterance in the subsequent conversation makes reference to the previous conversation, which is stored as an episode in the interlocutors' LTM, with its final state consisting of *inter alia* the non-empty QUD, which happens to be also MaxQUD. Hence, she can address this issue directly, using MaxQUD to resolve the non-sentential utterance.

5 Conclusions and Future Work

In this paper we have moved towards embedding dialogical semantics within neurologically well established memory systems. We have embedded certain very transient aspects as components of an M-WM style working memory. We have also sketched how to embed dialogues globally within LTM.

Both moves constitute a first step towards enabling semantics to model issues pertaining to memory problems, which occur both to dementia sufferers and to neurotypical speakers. The latter has the important consequence of enabling reference to entire conversations and to capture phenomena related to resumption and recurrence across conversations. In both areas there are a variety of hypotheses to be tested: for WM these include the question whether patients suffering from dementia can deal with stacking in QUD, a phenomenon making intrinsic use of binding into LTM. For LTM it raises the issue of testing the reality of long distance ellipsis in resumed conversations.

A much longer term challenge is to try to map the posited structures to brain structures, as is done, for instance, in models like MUC (Hagoort, 2016). Baddeley (2012) cautions against this, however, given current imaging *know how* and given the fact that a given memory structure is often apparently distributed across multiple brain structures.

Acknowledgments

We would like to thank two anonymous reviewers for SemDial and Alistair Knott for their detailed and very helpful comments. We would also like to thank Robin Cooper and Chiara Mazzocconi for discussion and encouragement. This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the program *Investissements d'Avenir* (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001. We also acknowledge a senior fellowship from the Institut Universitaire de France to the first author.

References

Margherita Alberoni, Alan Baddeley, Sergio Della Sala, Robert Logie, and Hans Spinnler. 1992. Keeping track of a conversation: Impairments in alzheimer's disease. *International journal of geriatric psychiatry*, 7(9):639–646.

Alan Baddeley. 1997. *Human Memory*. Psychology Press, Hove.

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

Lisa Feldman Barrett. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23.

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*, pages 1–66.

Gudrun Baumgartner and Karl-Heinz Renner. 2019. Humor in the elderly with dementia: Development and initial validation of a behavioral observation system. *Current Psychology*, pages 1–14.

Tobias Brosch. 2013. Comment: On the role of appraisal processes in the construction of emotion. *Emotion Review*, 5(4):369–373.

Citation-suppressed. 2019. Patient b corpus.

Robin Cooper. 2019. Representing types as neural events. *Journal of Logic, Language and Information*, 28:131–155.

Robin Cooper. 2020. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). Book Draft.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24:87–114.

Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, 169:323–38.

John F. Curry. 2014. Future directions in research on psychotherapy for adolescent depression. *Journal of Clinical Child & Adolescent Psychology*, 43(3):510–526.

Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

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

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

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa*.

Onur Güntürkün. 2005. Avian and mammalian "prefrontal cortices": Limited degrees of freedom in the evolution of the neural mechanisms of goal-state maintenance. *Brain Research Bulletin*, 66(4–6):311–316.

Peter Hagoort. 2016. Muc (memory, unification, control): A model on the neurobiology of language beyond single word processing. In *Neurobiology of language*, pages 339–347. Elsevier.

Laura Kennedy and Andreas Xyrichis. 2017. Cognitive behavioral therapy compared with non-specialized therapy for alleviating the effect of auditory hallucinations in people with reoccurring schizophrenia: A systematic review and meta-analysis. *Community Mental Health Journal*, 53:127–133.

Alistair Knott. 2012. *Sensorimotor cognition and natural language syntax*. MIT press.

Marcus Kracht. 2010. The inner dialogue: Pragmatics for one person. In Sebastian Bab and Klaus Robering, editors, *Judgements and propositions: Logical, linguistic, and cognitive issues*, number 21 in Logische Philosophie. Logos-Verlag, Berlin.

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

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2):109–158.

Joseph LeDoux. 2001. *Das Netz der Gefühle: Wie Emotionen entstehen*. DTV, München.

Camilla Lindholm. 2008. Laughter, communication problems and dementia. *Communication & medicine*, 5(1):3–14.

Sweyta Lohani, Adria K. Martig, Karl Deisseroth, Ilana B. Witten, and Bita Moghaddam. 2019. Dopamine modulation of prefrontal cortex activity is manifold and operates at multiple temporal and spatial scales. *Cell Reports*, 27(1):99–114.

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 & Biobehavioral Reviews*, 101:1–12.

Brian McElree and Barbara Anne Dosher. 2001. The focus of attention across space and across time. *Behavioral and Brain Sciences*, 24(1):129–130.

Ryan J Murray, Tobias Brosch, and David Sander. 2014. The functional profile of the human amygdala in affective processing: insights from intracranial recordings. *Cortex*, 60:10–33.

Victoria M. Puig, Robert Rose, Jonas Schmidt, and Nadja Freund. 2014. Dopamine modulation of learning and memory in the prefrontal cortex: Insights from studies in primates, rodents, and birds. *Frontiers in Neural Circuits*, 8(93).

Jonas Rose, A.M. Schiffer, Lars Dittrich, and Onur Güntürkün. 2010. The role of dopamine in maintenance and distractability of attention in the "prefrontal cortex" of pigeons. *Neuroscience*, 167(2):232–237.

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.

David Sander, Didier Grandjean, and Klaus R Scherer. 2018. An appraisal-driven componential approach to the emotional brain. *Emotion Review*, 10(3):219–231.

Klaus R Scherer and Heiner Ellgring. 2007. Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7(1):158.

Pamela Shakespeare. 2013. *Aspects of confused speech: A study of verbal interaction between confused and normal speakers*. Routledge.

Martin Takac and Alistair Knott. 2016a. Mechanisms for storing and accessing event representations in episodic memory, and their expression in language: A neural network model. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, CogSci, pages 532–537.

Martin Takac and Alistair Knott. 2016b. Working memory encoding of events and their participants: A neural network model with applications in sensorimotor processing and sentence generation. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, CogSci, pages 2345–2350.

John T Wixted. 2004. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.*, 55:235–269.

Different Games in Dialogue: Combining character and conversational types in strategic choice

Alafate Abulmiti
Université de Paris, CNRS
Laboratoire de Linguistique Formelle alafate.abulimiti@gmail.com

Abstract

In this paper, we show that investigating the interaction of *conversational type* (often known as language game or speech genre) with the character types of the interlocutors is worthwhile. We present a method of calculating the decision making process for selecting dialogue moves that combines character type and conversational type. We also present a mathematical model that illustrate these factors' interactions in a quantitative way.

1 Introduction

Wittgenstein (1953); Bakhtin (2010) introduced language games/speech genres as notions tying diversity of linguistic behaviors to activity. Building on this, and insights of pragmaticists such as Hymes (1974); Allwood (2000), and earlier work in AI by (Allen and Perrault, 1980; Cohen and Perrault, 1979) Larsson (2002); Ginzburg (2012); Wong and Ginzburg (2018) showed how to associate global structure with conversations in a way that captures the range of possible topics and idiosyncratic moves. (Larsson, 2002) is also the basis for an approach to building spoken dialogue systems (SDS) which is essentially domain general, offers a fine-grained treatment of grounding interaction, and which was extended to clarification interaction in (Purver, 2004).

This body of work does not address, however the issue of strategic choice in conversation, which is the core issue underlying work in Game Theory. Asher et al. (2017) used game theoretic tools to develop a theory of strategic choice for dialogue. Although there are a variety of interesting insights captured in this approach, it is based on two assumptions that apply only to a restricted class of language games—games continue indefinitely and there exists a jury that assigns winning conditions to participants.

We seek to develop an approach to strategic choice applicable to the general case of dialogical interaction, where termination is an important consideration and where assessment is internal to the participants. Strategic choice is modelled by combining structure from conversational types with psychological and cognitive notions associated with the players.

Character type as a relatively independent factor abstracted out of the conversational type is important for dialogue. Although there is some analysis concerning both character effects and conversational types in the dialogue, combining them and analyzing their interactions in a quantitative way has not, to our knowledge, been carried out before. The purposes of this paper is, hence, to combine character type effect and conversational type analysis to yield a method that could help to analyse strategic choice in dialogue.

2 Background

The starting point of (Asher et al., 2017) is the framework of Banach-Mazurckiewicz games. They modify this framework to make it more amenable to analyzing certain kinds of NL dialogue, the emergent framework being *BM messaging games*. Asher et al. (2017) argued that each dialogue potentially continues indefinitely and has a winner adjudged by a third party jury. This is useful for modelling political discourse between rival groups or individual contenders in the public domain. But clearly this sort of conception is not appropriate for a variety, arguably most types of dialogue.[1] These have beginnings ($InitState$) and a variety of distinct terminations ($FinState$) (Wong, 2018), and there

[1]Strictly speaking, Asher et al. (2017) allowed also for finite games, by allowing for games to consist of a special move from a certain point onwards, but this would seem to either defeat the purpose of assuming potential extendibility or to be purely formal.

is no 'external jury' in most cases.

Burnett (2019) developed a formal model called *social meaning games* which explains how social variants affect the linguistic style of dialogue participants. And conversely, how the speaker's intrinsic linguistic style affects dialogue moves. Pennebaker and King (1999) shows that linguistic style is an independent and meaningful way of exploring personality. There is evidence that people's personality traits influence their use of language. For instance, extroverted individuals are able to deceive others more easily, while neurotic individuals are less likely to be deceived (Riggio et al., 1988). The charisma of a speaker has been shown to be closely related to an extroverted character (Bono and Judge, 2004). There is also a strong relation between extroversion and conscientiousness and positive influences, as well as between neuroticism and dissent and various negative influences(Watson and Clark, 1992). Thus, an individual's personality does affect decision-making process in dialogue.

Cooperation in dialogue is a widespread phenomenon, and Allwood et al. (2000) identified four features of cooperation: cognitive consideration, joint purpose, ethical consideration and trust. When engaging in a collaborative dialogue, where the interlocutor decides his next move based on the intentions of the other and a variety of information deriving from the context of the dialogue, it seems that character has a broad influence on the course of the dialogue. Thus, it seems natural that a dialogue participant (DP) should also take into account the other's character traits in order to choose the appropriate move.

In the next section, we will explain the method we propose of combining character type effects and conversational type.

3 Methodology

In this section, we wish to explore the interaction between character type and conversational type, by considering how given a possible range of moves, a quantitative analysis can be provided for move selection.

3.1 Character Type

Researchers have developed a relatively unanimous consensus on personality description patterns, proposing the Big Five model of personality(Goldberg, 1992). Within this model there are five traits that can be used to capture many aspects of human personality. The Big Five Personality (OCEAN) can be assessed by the NEO-PI-R(Costa Jr and McCrae, 2008). The traits can be described as follows:

- Openness: the ability to be imaginative, aesthetic, emotional, creative, intelligent, etc.
- Conscientiousness: displays characteristics such as competence, fairness, organization, due diligence, achievement, self-discipline, caution, restraint, etc.
- Extroversion: exhibits qualities such as enthusiasm, sociability, decisiveness, activity, adventure, optimism, etc.
- Agreeableness: the qualities of trust, altruism, straightforwardness, compliance, humility, empathy, etc.
- Neuroticism: difficulty balancing emotional traits such as anxiety, hostility, repression, self-awareness, impulsivity, vulnerability, etc.

Goldberg (1992) gave a pertinent method to quantify character types in terms of 5 dimensional vector $[o, c, e, a, n]$. We define χ_s for the self character type scale vector and χ_o represents other character type scale vector.

In addition, with the development of machine learning and deep learning methods within NLP, a variety of approaches have been implemented for automatic recognition of personality in conversation and text (Mairesse et al., 2007). Jiang et al. (2019) used attentive network and contextual embeddings to detect personality traits from monologues and multiparty dialogues. Given a text(or utterance), one can calculate the character type scale vector of this sentence with a robust prediction model. We define c_i as ith dialogue move vector prediction.

We note that by calculating the similarity between the χ and c_i, we obtain the extent to which a given dialogue move can fit either the self character type or the other character type. Note that χ_s is a dialogue interlocutor's intrinsic property which does not show great change during conversation, but considering one's imperfect information situation, χ_o will change once new evidence arises and can be modified by applying Bayes' rule.

3.2 Conversational Type

Pennebaker and King (1999) also indicated that linguistic style gets influenced by the situation in which the interlocutor find themselves. Wong and Ginzburg (2018) provided a topological perspective on the space of conversational types based on the distribution of Non-Sentential Utterances (NSU) within each type. Wong (2018) developed a model of a conversational type couched in TTR (Cooper, 2005). On this view, a conversational type is a 4-tuple $\{ConvRules, InitState, FinState, G\}$, here $ConvRules$ represents a set of conversational rules, transition rules between different dialogue states (*dialogue gameboards*) of type $DGBType$ (Ginzburg, 2012), $InitState$ and $FinState$ are the initial and final DGB states.

$$DGBType \mapsto \begin{bmatrix} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{utt-time: Time} \\ \text{c-utt: adressing(spkr, addr, utt-time)} \\ \text{FACTS: set(Proposition)} \\ \text{Pending: list(Locutionary Proposition)} \\ \text{Moves: list(Locutionary Proposition)} \\ \text{QUD: poset(Question)} \end{bmatrix}$$

G is the grammar which serves for the different conversational language use.

An Example Considering a commercial transaction conversational type involving shopping in a bakery, so that we have:

$$Bakery = \begin{bmatrix} \text{participant: InteractionGroup} \boxed{\wedge} \begin{bmatrix} \text{c1:customer(A)} \\ \text{c2:baker(B)} \end{bmatrix} \\ \text{qnud-set = QS : poset(question)} \\ \text{c1} : \left\{ \begin{array}{l} \lambda \text{ x.InShopBuy(A,x)}, \lambda \text{ P.P(A)}, \\ \lambda \text{ P.P(B)}, \lambda \text{ x.Pay(A,x)} \end{array} \right\} \in \text{QS} \\ \text{moves : list(IllocProp)} \end{bmatrix}$$

this involves:

- participants: Baker and customer.
- qnud-set: questions to be discussed during the interaction
- moves: Dialogue moves made.

Wong (2018) argued that clarification interactions provide data showing that interlocutors can be uncertain about the conversational type that classifies the interaction they are in, as for instance in (1):

(1) (Context: A is being interviewed by B) A: Hi B: Hi . . . B (1): have you seen Blackklansman yet? A (2): Wait— is this an informal chat or a formal interview? B: A bit of both. (Example (54) from (Wong, 2018).

Thus, the DP has an opaque or uncertain "guess" as to the conversational type and this also influences the decision-making process. Following probabilistic TTR theory (Cooper et al., 2014), we assume that the DP has a probabilistic conversational type in the private part of information state and this can be updated during dialogue by Bayesian inference. We model the information state as follows:

$$InformationState = \begin{bmatrix} \text{private} = \begin{bmatrix} \text{CharacterType:} \begin{bmatrix} \text{Self: Vector} \\ \text{Other: Vector} \end{bmatrix} \\ \text{Goals: Set(Prop)} \\ \text{Tmp: private} \\ \text{ConvType : ConvType} \\ \text{Conv-prob : [0, 1]} \end{bmatrix} \\ \text{dgb :DGBType} \end{bmatrix}$$

In the private part, we have the *CharacterType* stores for self and other's character type vector; *Goals* tracks a set of (futurate) propositions that the DP wants realized; *Tmp* is a backup of the previous private state. Its use is for reasoning about the current *CharacterType* and updating the *Conv-prob*; *ConvType* is the conversational type and here we introduce *Conv-prob*, which indicates the probabilistic conjecture concerning the current conversational type.

3.3 Move Space

Individuals choose the moves from different kind of possible options. Ginzburg et al. (2019) provided a taxonomy of the response space of queries. 14 possible categories are given by the study of English and Polish corpus. The field of Nature Language Generation has some useful probabilistic methods to generate the range of potential moves, such as the Sequence-to-Sequence model(Sutskever et al., 2014). See et al. (2019) proposed a new conditional weight control technique to make the response space more human-like. The popular pre-training model like BERT(Devlin et al., 2018) exemplifies

NLG capability with fine-tuning. We define here A for a dialogue move space vector composed by a_i representing the ith move.

In the current work we will not discuss how to specify the possible moves, but will leave this for future work.

3.4 Decision Making

3.4.1 Global View

Levelt (1993) proposed that speakers monitors themselves while speaking. In other words, individuals have a self-criticism mechanism enabling them to reflect on their behaviors, emotions and speaking. We dub this the *SelfMonitor*.

Given our analysis in the first two subsections, a DP's move choice in the dialogue is influenced by her character type, the other interlocutor's character type, and the conversational type. When a DP tries to respond to the other party's dialogue move, she first constructs a dialogue move space, which yields a set of possible utterances that the DP can use. The DP typically makes a conjecture about the other's character type in terms of individual personality traits, based on her a priori knowledge of that individual and the current state of the dialogue. In addition, the DP has a probabilistic assumption about the present conversational type given her cognitive state. Subsequently, the DP's *SelfMonitor* determines which factor is more valuable in this context. Eventually a move is selected considering each possible move's value by comparing the affinity of each move for each factor with the skewness of each factor as determined by the *SelfMonitor*.

3.4.2 Mathematical Modeling

After the above analysis, we offer a mathematical model to explicate this process. In evaluating possible moves, we have three important factors: self character type, other character type, and the conversational type. We want to provide a real valued function ρ to evaluate each move in the dialogue move space.

- a_i: ith move.
- $\overrightarrow{\chi_s}$: Self character type vector.
- $\overrightarrow{c_i}$: character type vector for ith move.
- α Weight for the self character type effect.
- $\overrightarrow{\chi_o}$: Current other character type vector estimation.
- β: Weight for the other character effect.
- p: Probabilistic conjecture of the conversational type.
- d_i: ith move conformity with the conversational type from -1 to 1.
- γ Weight for probabilistic conversational type.
- $W = [\alpha, \beta, \gamma]$.

Conformity represents the degree to which a dialogue move conforms to the current conversational type. In other words, it can be modeled as the evaluation score generated by this dialogue move in dialogue context.

In order to calculate the "affinity" among character type vector $\{\chi_s, \chi_o\}$ and move vector (c_i), we use *cosine* similarity, defined as follows:

$$simi(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

Then we define the function $\rho(a_i)$:

$$\rho(a_i) = \alpha \cdot simi(\overrightarrow{c_i}, \overrightarrow{\chi_s}) + \beta \cdot simi(\overrightarrow{c^i}, \overrightarrow{\chi_o}) + \gamma d_i \cdot p \quad (2)$$

and let $X = [simi(\overrightarrow{c_i}, \overrightarrow{\chi_s}),\ simi(\overrightarrow{c^i}, \overrightarrow{\chi_o}),\ d_i \cdot p]^T$ as a decision factor matrix, obtaining:

$$\rho(A) = W \cdot X \quad (3)$$

where $\alpha + \beta + \gamma = 1$

α, β, γ are in fact estimated by the *SelfMonitor* based on information deriving from the information state. We believe those weights are mainly fixed in the beginning of the conversation, because great changes in strategic choice lead to the suspicion of deception in some cases (Riggio et al., 1988), This estimation process can be back-engineered by observing DP's selection of moves in the dialogue.

After the calculation of ρ, we get the score for each move. This score alone cannot determine the final decision—we need to take into account the features that we have not yet discussed and observed, so here we will probabilize them, i.e., convert it into a probabilistic distribution space using the $softmax$ function.

$$softmax(a_i) = \frac{exp(\rho(a_i))}{\sum_j exp(\rho(a_j)))} \quad (4)$$

We then obtain a probability estimate for each move, which indicates that the greater the probability the DP is more inclined to choose this move.

3.5 Example

Here we illustrate our proposed approach with an example.

3.5.1 Scenario

In a bakery, we observe a customer and baker's buying and selling process during the pandemic of COVID-19.

Goals For simplicity we fix the final goals as follows:

(2) a. Customer goal: buy two croissants.

b. Baker goal: sell the two croissants and obtain the desired price.

It is worth noting that both players may have more complex goals or a series of goals. For example, the customer might want to have appropriate desserts for dinner. In order to achieve those kind of goals, we often divide into several specific and simple goals. In such a case, the sub-goal might be:

1. what are the best desserts in this bakery?
2. From among those best desserts, which one will fit the dishes I prepare tonight?

For our current purposes, however, we will avoid such complications, despite their importance in a variety of cases.

3.5.2 Move Space

We assess a baker's possible responses to the customer's initial utterance: "2 croissants". We assume the following possible responses of the baker:

(3) CLIENT: 2 croissants.

BAKER: (i) 1.90.
(ii) Get out of the bakery, you're not wearing a mask.
(iii) Please would be nice.
(iv) 1.90 and please would be nice.

(i) would lead to a quick end that meet their needs. This "style" is used in most cases in our life disregarding other conversational factors and attending only to the baker's final goals. If the baker considers particularly the conversational type's impact, he would choose (i) to advance the conversation, thereby looking forward to finalizing the conversation early and achieving his goal.

(ii) would lead to an "unpleasant" conversation. The baker thinks that the customer's behavior is disrespectful (a short demand lacking politeness). He therefore uses the lack of the mask as a pretext or has in mind a competing goal for fighting against this disrespect. Finally, neither participants goals is achieved. (iii) This choice shows that the baker wants to have a pleasant and respectful conversation above all. It is clear that baker does not assign much weight to his final goals or the conversational type. Instead he prioritizes his psychological needs. The question under discussion would shift to another topic and the conversation might evolve into a dispute, though with lower probability than in (ii).

(iv) indicate that the baker wants to persist with the trade without disturbing his final goals and his psychological needs. This seems to be very much of a compromise move, but still the final state would depend on the customer's interpretation of the dialogue pragmatics.

3.5.3 A worked example

For this example, we have a character type vector of [o, c, e, a, n] we assume that baker character vector is χ_s = [0.0, 0.3, 0.0, 0.0, 0.5] showing conscientiousness and relatively high neuroticism and *Conv-prob* is p = 0.98 assuming the baker's conformity with the bakery conversational type is high. Following the initial state S_0, the baker hears "2 croissants" from client, after which the baker updates the information state then the *SelfMonitor* chooses values for α, β, γ. For example, we assume α = 0.1, β = 0.1 and γ = 0.8, we assume, in this circumstance, that the baker thinks more about the conversational type rule than self character type and other character type, then:

$$CharacterType.Other = \chi_o$$

$$\chi_o = \mu(\Uparrow Tmp.CharacterType.Other, \Uparrow dgb)$$

μ is the other's character type updating function with parameters: the previous state's character type vector and the current DGB. We assume that after the other's character type update we have χ_o = [0.0, 0.0, -0.1, -0.4, 0.2] assuming the customer is not very agreeable (-0.4) and a little neurotic (0.2).

(i): We assume that "1.90" shows disagreeableness (-0.4) and slight neuroticism (0.2) so c_1 = [0.0, 0.0, -0.1, -0.4, 0.2] and d_1 = 0.8 shows "1.90"

shows high conformity with the bakery conversational type. Then according to function (2) we have: $\rho(a_1)$ = 0.7646

(ii): We assume that "Get out of the bakery, you're not wearing a mask" shows high disagreeableness (-0.7), low conscientiousness (-0.5) and high neuroticism (0.8) so c_2 = [0.3, -0.5, 0.0, -0.7, 0.8] and d_2 = -1.0 shows incongruity with the bakery type. As a result, we obtain $\rho(a_2)$ = -0.7080

(iii): We assume that "Please would be nice" shows high agreeableness (0.7) and slight extroversion (0.3) c_3 = [0.2, 0.0, 0.3, 0.7, -0.2] and d_3 = 0.3 shows low conformity with the conversational type. Consequently, we obtain $\rho(a_3)$ = 0.1201

(iv): We assume that "1.90 and please would be nice" shows relatively high openness (0.5), conscientiousness (0.6) and extroversion (0.4), high agreeableness (0.7) and low neuroticism (-0.4) so c_4 = [0.5, 0.6, 0.4, 0.7, -0.4] and d_4 = 0.7 shows high conformity with the bakery type. So we obtain, $\rho(a_4)$ = 0.4727

We then apply those preference score with the $softmax$ function: $softmax(\overrightarrow{\rho(a_i)})$ = [0.3998, 0.0917, 0.2099, 0.2986]

This indicates that the probability that the baker selects (i) is 39.98% (iv) is the next ranked score. All this assuming that the baker's character type is relative high neuroticism and the commitment degree to the conversational type is high. Given his character type, it is reasonable to conclude that he would not be extremely polite (option iv) but not rude because of the constraint of conformity with the conversational type.

Now we want to illustrate that if we change the distribution of certain of the parameters, things can flip significantly. We change the distribution of different variables (α, β, γ) to [0.3, 0.1, 0.6] showing that the baker is concerned a little more about his own character type (0.3 rather than 0.1) but slightly less than before on conversational type (0.6). And we assume the baker's character type vector is [0.5, 0.7, 0.3, 0.8, -0.5] which shows high conscientiousness (0.7), low neuroticism (-0.5) and high agreeableness. Then we obtain:$\rho(\overrightarrow{A})$ = [0.3457, -0.7277, 0.3217, 0.6946]. the resulting following with probabilistic distribution is $softmax(\overrightarrow{\rho(a_i)})$ = [0.2677, 0.0915, 0.2613, 0.3795]

This indicates that the baker would choose option (iv) for the next move with 38% probability, an option which balances the baker's final goal and personal psychological needs (high conscientiousness).

Finally, we modify (α, β, γ) to [0.8, 0.1, 0.1] and assume that the baker's character type vector is [0.2, -0.3, 0, -0.5, 0.8] which shows high neuroticism (0.8), low agreeableness (-0.5). What we obtain is $\rho(\overrightarrow{A})$ = [0.8007, 0.7652, -0.5229, -0.5032]. From this what results is the following: $softmax(\overrightarrow{\rho(a_i)})$ = [0.3996, 0.3856, 0.1064, 0.1085]

This indicates that the probability of choosing (i) and (ii) is now about the same. In this scenario the baker focuses more on his own character type. (ii) shows complete incongruity with the bakery type, which indicates a complete violation of the baker's original final goal. Hence, at this point, the baker is in a dilemma. It is worth noting that this state will lead to a "breakthrough point", and if the baker chooses (ii), it means that baker chooses to change his final goal or conversational type. The consequence of this is that the $Goals$, $ConvType$ and $Conv - prob$ should be changed in the private state of information state. How to effect this in a formal way, we leave to future work.

4 Conclusion

Character is a person's stable attitude towards reality and it also affects one's performance in dialogue. Conversational type is, in one way or another, since the early days of AI work on dialogue, one of the principal notions of dialogue research. It reflects domain specific interaction, in particular the move space given to the conversational participants. We have tried to show in this paper that investigating the interaction of these two factors is worthwhile. In particular, we present a method of the decision making process for moves that combines character type and conversational type. We present a mathematical model that combines these factors after assigning numerical values to the various parameters that are involved and demonstrate by means of an example how this works.

5 Future Work

In this paper, we have made a preliminary proposal concerning the modelling of character types and their combination with conversational types. We aim to refine this in future work in ways that include the following:

- Wong (2018) gave a formal method to classify conversations into types. We believe that

under realistic conditions, there are often multiple conversational types involved in a single conversation, which may involve sequential transformations or overlapping phenomena.

- In the approach sketched here, for classifying the conversational type we use a probabilistic TTR. However, in practice this assessment can change as the dialogue unfolds. We hope to develops methods that incorporate such dynamism.

- Personality analysis based on the Big Five theory is robust, but inference about each other's character types is also in flux during conversations, and examining the impact of the change process on our approach is worth looking forward to future work.

- Ginzburg et al. (2019) provided a categorical approach to the response to the question. We hypothesize that the conversational type has the role of delimiting the range of possible moves. We aim to characterize this variability.

- This article introduces the concept of move conformity, which is required for automatic detection of research based on different types of dialogue in the future work. This could be achieved by modifying an existing NLG evaluation model (e.g. BLEU (Papineni et al., 2002), ADEM (Lowe et al., 2017)).

- This paper discussed several cases in detail, but they are all constructed examples, so fitting this model on actual conversations (e.g. the British National Corpus) or scripted dialogues annotated for character types (e.g. FriendQA (Yang and Choi, 2019)) with experimental predictions is desirable.

Acknowledgments

This work was supported by an internship funded by the Institut Universitaire de France, within Jonathan Ginzburg's project *Unifying Verbal and Non-verbal Interaction*. We also acknowledge the support of the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083). Thanks to Prof. Ginzburg's patient guidance during the internship. Thanks also to three anonymous reviewers for SemDial for their detailed and perceptive comments.

References

James Allen and Ray Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178.

Jens Allwood. 2000. An activity based approach of pragmatics. In Harry Bunt, editor, *Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics*. Amsterdam, John Benjamins.

Jens Allwood, David Traum, and Kristiina Jokinen. 2000. Cooperation, dialogue and ethics. *International Journal of Human-Computer Studies*, 53(6):871–914.

Nicholas Asher, Soumya Paul, and Antoine Venant. 2017. Message exchange games in strategic contexts. *Journal of Philosophical Logic*, 46(4):355–404.

Mikhail Mikhalovich Bakhtin. 2010. *Speech genres and other late essays*. University of Texas Press.

Joyce E Bono and Timothy A Judge. 2004. Personality and transformational and transactional leadership: a meta-analysis. *Journal of applied psychology*, 89(5):901.

Heather Burnett. 2019. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 42(5):419–450.

Philip Cohen and Ray Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212.

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

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 72–79.

Paul T Costa Jr and Robert R McCrae. 2008. *The Revised NEO Personality Inventory (NEO-PI-R).* Sage Publications, Inc.

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*.

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

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.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Dell Hymes. 1974. Ways of speaking. *Explorations in the ethnography of speaking*, 1:433–451.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2019. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. *arXiv preprint arXiv:1911.09304*.

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

Willem JM Levelt. 1993. *Speaking: From intention to articulation*, volume 1. MIT press.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Matthew Richard John Purver. 2004. *The theory and use of clarification requests in dialogue*. Ph.D. thesis, University of London.

Ronald E Riggio, Charles Salinas, and Joan Tucker. 1988. Personality and deception ability. *Personality and Individual Differences*, 9(1):189–191.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

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.

David Watson and Lee Anna Clark. 1992. On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. *Journal of personality*, 60(2):441–476.

Ludwig Wittgenstein. 1953. Philosophical investigations. basil & blackwell. *OxfordWittgensteinPhilosophical investigations Basil1953*.

Kwong-Cheong Wong. 2018. *Classifying Conversations*. Ph.D. thesis, Paris Diderot University.

Kwong-Cheong Wong and Jonathan Ginzburg. 2018. Conversational types: a topological perspective.

Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.

An Outline of a Model of Situated Discourse Representation and Processing

David Schlangen
Computational Linguistics // Department of Linguistics
University of Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

Situated discourse provides opportunities for both exophoric and endophoric reference, hence requiring mechanisms linking both language and world, and language and (representation of) language. It is also, arguably, the primary site for learning about these mechanisms, which I take to be the meaning of words. I outline a model that brings together a treatment of discourse reference and a learning-based treatment of exophoric reference, based on the metaphor of "mental files". It combines a (rudimentary) model of visual attention with a model of exophoric reference resolution, built on a pluralistic model of word meaning that accounts for learning from ostension, learning from overhearing and learning from definition. Lastly, it offers several pathways to making inferences, accounting for fast judgements and slow justifications. At this stage, the model mostly provides a way of structuring the task of processing situated discourse, meant to be extensible both in the direction of stricter formalisation as well as in the direction of further implementation, the beginnings of which are reported here.

1 Introduction

Let's start with a story.

> Ann and Bert are taking a stroll in the park. *"Look at the dog, over at the waterfountain"*$_{(1)}$ Ann says, looking at a man carrying a dog in his arms. *"I know, isn't it cute!"*$_{(2)}$, Bert replies.
>
> A short while later they run into Chris and her young son, Dale. *"We just saw a man carrying a dog,"*$_{(3)}$ Bert exclaims. *"The cutest poodle ever!"*$_{(4)}$, adds Ann.
>
> *"Actually, I don't think that was a poodle."*$_{(5)}$, says Bert, who can't stop himself. *"It was too tall. I think it was a labradoodle."*$_{(6)}$ Ann knows that Bert is a serious hobby cynologist, so she says *"Oh. I guess you're right. You're the expert here."*$_{(7)}$
>
> After exchanging more pleasantries, they go their separate ways again, Chris and Dale walking into the direction from which Ann and Bert just came. And sure enough, shortly thereafter they see a man who is holding a beautiful white-ish dog on a leash. Chris happens to know the man, who had previously told her about his new dog, and so she thinks to herself: *"They were talking about Fredo!"*$_{(8)}$ Dale, however, thinks *"So that's what a labradoodle looks like."*$_{(9)}$

This is admittedly not a particularly exciting story. But mundane as it is, no existing computer system can model all the behaviours exhibited here, and neither can any formal model of situated discourse. So, what is it that Ann, Bert, Chris, and Dale need to be able to do to play their respective parts in this little story, and how could that be modelled in a machine?

• For exchange (1)/(2) to work (referring to individual utterances from the story via their subscripted numbers), both Ann and Bert need to be able to parse their environment into objects, and to recognise one of these as being appropriately described by *dog*. Furthermore, after (1), that dog can also be referred to by *it*, which wasn't the case before.

• To understand (3)/(4), Chris and Dale need to be able to connect *poodle* and *dog* in such a way as to let the phrases in which they occur refer to the same entity, or at least recognise that to do so is Ann's intention with (4). Unlike for Ann and Bert in (1)/(2), for Chris and Dale that entity is not also perceptually available. For Ann and Bert, it is not perceptually available *anymore*, but they will have a visual memory of it.

• (5)/(6)/(7) shows that categorisation decisions are *negotiable*: Bert denies applicability of *poodle* to the entity from discourses (3)/(4) and (1)/(2), provides reasons for doing so, and suggests a better category. (7) indicates that expertship is a reason

for accepting a proposed revision. We can assume that an effect of this interaction is that Ann changes her understanding of *poodle*, and possibly that of *labradoodle* as well, and perhaps Chris and Dale do so as well with their respective ones.

• (8) and (9) finally show some later consequences of the interaction. (8) shows that Chris can now make a connection she couldn't make earlier (the dog that Ann and Bert saw is Fredo, the new dog of her acquaintance). (9) shows Dale attaching visual information (from perceptual acquaintance) to a concept that previously, we assume, had only been introduced verbally to him.

We can derive from this analysis some desiderata for a model of agents that can participate in such situated discourse:

A) It needs to provide visually present real world objects as possible referents, as well as objects only introduced via discourse, as well as entities that previously were introduced either way.

B) It needs to provide word meanings that guide the mechanisms by which utterances are linked to either real world objects ("dog" in (1)), or to discourse referents ("dog" and "poodle", in (3)/(4)); which moreover are accessible for discussion and revision (5)/(6)/(7).

C) It needs to provide a way in which word meanings are composed into utterance meanings, and a way in which those are composed into discourse meanings.

In this paper, I attempt to outline a model that meets these desiderata. A particular focus in doing so is on the potential for recruiting current computational work for specifying the *mechanisms* that would be needed in a realisation of a dynamic model of situated discourse, and on arguing for a particular way of factorising the problem. This comes at the cost of a lack of formal rigour (for example, a lack of a model theory that would precisely specify the *conditions* under which an agent would evaluate a statement as true). This is something that I hope can be delivered later, where appropriate.[1]

2 Overview

Following closely the structure of the desiderata from the previous section, the proposal consists of several interlocking parts. The job of the **context representation component** is to provide the entities that the interpretation process links together. To perform this linking, it is not enough to just have access to (mental representations of) entities, there also needs to be further information connected to them. To describe this structured keeping of information, I will make use of the *file metaphor*, according to which each entity is represented by one file, on which additional information can be "written".[2] To give an illustration, Figure 1 shows what Bert's representation of the situation after utterance (1) looks like. A *file card* (along the lines of Heim's (1983) file change semantics) has been created for the discourse entity that (1) introduced; as part of the situated interpretation, Bert has been able to link that to one of the *object files* that his perceptual system created to represent the visual scene. Had he recognised the object as a previously know entity, he could also have made a link to what we call an *entity / event (e/e) file*; but here this is not the case.

What enabled Bert to make the connection are relevant *concept files*, which are part of the **conceptual component** of the model. Our proposal here is to account for the relevant desiderata rather directly, by representing the conceptual content that enables categorisation of perceptual objects (in the example here, identifying one object as a dog, and as white) and that which enables relating concepts (in the story, *dog* and *poodle*), separately. Moreover, we assume that at least parts of this content are accessible to their owner, and can be discussed and voluntarily revised (as in (5)-(7) of the story). The concept files come categorised into domains of knowledge, and this can factor in the decision to defer to a judgement (and revise one's own concept), or not.

Of course, it's not single words that trigger the creation of links in the context representation, it's the composition of words into phrases and utterances. The final part of the model then is the **composition component**, which composes utterance meanings out of the recognised structure and the activated concept files.

The remaining sections describe these components further and speculate on the computational mechanisms that might be used to create and maintain these representations, and discuss where cur-

[1] Maybe what I'm proposing here is not so much an actual model than more a research program, or a proposal for how to put together some existing smaller parts into one larger whole.

[2] This metaphor has a long, and wildly branching, history in linguistics (e.g., (Karttunen, 1969; Heim, 1983)) and philosophy (see the recent overview in (Recanati, 2012)), to the extent that not any single precursor can be declared here.

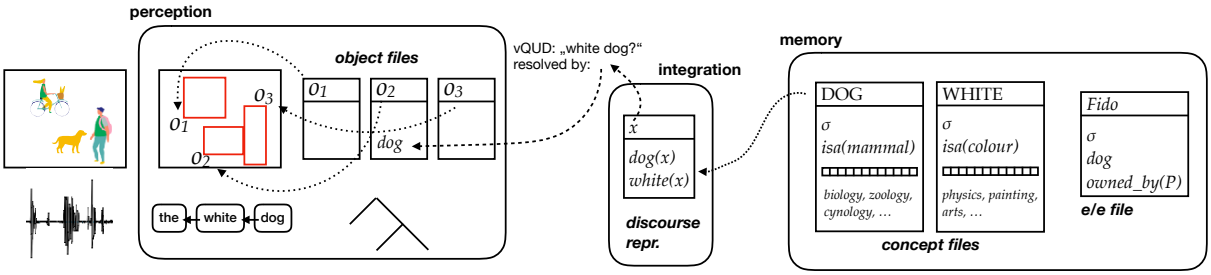


Figure 1: Overview of the Representational Components of the Model

rent implementations fall short.

3 Context Representation

As mentioned above, this is the component that is responsible for providing the mental representations that are linked together in the process of discourse interpretation. At the level of elaboration achieved here, I am assuming that any standard model of discourse representation can handle the part of representing discourse referents. The file metaphor I am using throughout here matches more closely with Heim's (1983) model, but the graphic representation in Figure 1 is more clearly inspired by DRT (Kamp and Reyle, 1993). (And a more sophisticated model like SDRT (Asher and Lascarides, 2003) would be needed to account for all the discourse coherence phenomena that I gloss over here.)

What is less often spelled out is how the visual scene can be made accessible in discourse representation (but see the related work discussed below). In the proposed model, the visual scene is represented via what I call *object files*. How do these object files connect to the actual objects in the world? I assume here that the indices with which the object files are indexed are created automatically by the perceptual system, and that these indices are pre-conceptual in that when created they do not represent more than raw objecthood.[3] I take this from Pylyshyn (1989), where a similar concept was introduced to account for the experimentally tested ability of human subjects to track a (limited, but not small) number of objects moving across a screen. Interestingly, this tracking worked even if some features of the objects (such as colour, shape, and texture) changed continuously; in fact, these changes were often not even noticed. Pylyshyn (1989) calls these tracking representations FINSTs ("Fingers of Instantiation"), with the idea that they function as indices for visually presented objects, individuating them across time and providing spatial information about them, without anything more needing to be recognised and represented about the objects. In later work, Fodor and Pylyshyn (2015) recognised the utility of this conception for providing a basis for conceptual reference, and it is this use that I assume here as well.

The indices or FINSTs only represent the presence of presence of *some* kind of object. Clearly, in many situations, this it is not enough for an agent, and it is of interested to know *which kind* exactly it is. However, any given object will be a member of many kinds. I make the assumption that the kinds that currently matter are determined by the current interests of the agent. As a simple device for specifying these interests, I introduce into the model the notion of a "visual Question under Discussion" (vQUD), in analogy to the discourse structuring device "QUD" proposed by Ginzburg (1995, 2012). This device is to act as an interface between bottom-up object *recognition* and top-down object *classification*.

In many situations, this question may default to something like "what's this?", which can perhaps be resolved by classification into what (Rosch, 1978) called *basic level categories*. But the mechanism should also account for how differently the world is parsed if there are pressing current concerns: If approached by a hungry hyena, you need a fast answer to "what can I throw?"; when hungry yourself and in the absence of any wild beasts, the salient question is a quite different one. To go back to our example from the introduction, and also to bridge to the next section, the vQUD mechanism shall also serve as the interface to reference resolution, where the question is directly provided by the

[3] The notion of "objecthood" that I am appealing to here is compatible with that investigated by Spelke (1994); Spelke and Kinzler (2007); Carey (2009) as likely to be innate, i.e. roughly "discrete, bounded entity that moves coherently, if force is exerted or produced by it".

task of utterance context integration.[4]

In any case, resolving the question on vQUD requires the application of relevant concepts to the objects. The objects are accessible via their indices / FINSTs, and the results of the classification are collected in the aforementioned *object files* that are connected to them.[5] These are meant as a representation of the current conceptual information about the object, active for as long as the current scene is attended to. (Let's assume that information about the perception event can be transfered into long-term memory in the form of an "event / entity file", as mentioned above. Also, re-recognition of an individual can be achieved by linking such an e/e file to an active object file; but more on that later.)

Lastly, let us also assume that an agent typically optimistically supposes that objects they recognise are also recognised by co-present agents.

Implementation With the dramatic improvements in quality that computer vision models have seen in recent years, it might seem that an implementation of this part of the model could be taken directly off the shelf. And it is true that current models like Faster R-CNN (Ren et al., 2017) and YOLO (Redmon et al., 2016) achieve high accuracy in segmenting and labelling still images. However, these models work by overgenerating proposed object regions and subsequent filtering through categorisation from a fixed set of categories. This misses two properties of the model as described above: a) it neglects that are always many valid classification possiblities, and which one is to be selected depends on current purposes (the reason why we've introduced the vQUD mechanism), and b) it restricts the set of classes to a prespecified finite set, whereas we want it to be possible to incrementally learn new classes. Hence, some adaptations to the basics of these algorithms would be required in an implementation of the model.

In the computational experiments we have run so far, we have skipped over this step and start with the segmentations provided by typical vision corpora (e.g., COCO, (Lin et al., 2014)), where objects are marked by bounding boxes. We take these, and the identifiers they come with, as simulating the output of the initial perception stage, providing the indices to which the object files are linked.

4 Word Meanings and Concept Files

As mentioned in the introduction, the concepts files collect the information that is required to create the links between representational objects; that is, between discourse referents and object files, discourse referents and antecedents, and discourse referents and e/e files. I first discuss some of the desiderata of what mental concepts should address, before presenting the proposal.

The Challenges I follow Bloom (2000) in assuming that learning the meaning of a word entails associating a form with a concept. If we understand concepts to be mental representations, as I am doing here, there is a problem. Mental representations are private, but words are public entities belonging to a public language, and so we will need a way to make the *content* of concepts publicly negotiable. Secondly, we need these concepts to do different things—or at least I will analyse these as being different—namely to support both exophoric and endophoric reference, of the kinds illustrated with our introduction story. (More generally, we will want them to support the computation of denotations and of material inferences.) Connected to this is that we want to capture different ways of learning—or at least we will analyse these as different—namely learning from observation of successful reference, learning from explicit definition (through linguistic explanation), and learning from implicit definition (through linguistic context). Lastly, we need to set them up in such a way that they are individuated in the right way, and the challenge to answer here is what Perry (2012) calls the 'co-reference and no-reference' problems (which are, of course, Frege's (1892) puzzles): How can an agent who possesses the words/concepts *Hesperus* and *Phosphorus* potentially learn something new when told that those name the same object? And what makes the concepts of Santa Claus and of unicorns different, given that they both refer to

[4] Below, as in this paragraph, I represent the questions on vQUD using natural language, without intending to commit to assuming a language of thought. But the relation between the kinds of categorial decisions one can make privately and those that one can make publicly via language, however, is of course a fascinating subject, left here to be explored in future work.

[5] Borrowing this label from Kahneman et al. (1992), who give a vivid example of why separating object *individuation* and object *classification* seems to fit the phenomenology of perception: "Imagine watching a strange man approaching down the street. As he reaches you and stops to greet you he suddenly becomes recognizable as a familiar friend whom you had not expected to meet in this context." It is the classification that changes here, not the identity of the object (as object) that you had tracked all along.

nothing?[6]

The Proposal The proposal I make here is to reify the observed (or claimed) differences, as it were. Namely, I assume that these different facets are indeed accounted for by different representations, brought together in what I call *concept files*.[7] Figure 1 gave the example of Bert's concept file for DOG (repeated below as Figure 2). The σ stands for the first component, the knowledge that allows an agent to recognise an object as falling under the concept or not. I follow the 'words as classifiers' model here (Schlangen et al., 2016; Kennington and Schlangen, 2015) (see also (Larsson, 2015, 2011)) and realise this technically through statistical classifiers that generalise via supervised learning and error-driven adaptations. (See below for technical details.) Such knowledge, I assume, can be picked up whenever successful reference is observed. This can be, but doesn't alway have to be, in episodes of ostensive teaching; I only assume that the learning agent must be capable of understanding what the intended referent was (so Bloom's (2000) arguments against associationism do not bite here).

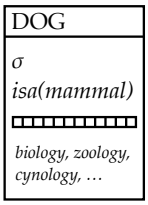


Figure 2: A Concept File for DOG

The second component is what accounts for the discussion in (5) and (6) of our introductory example, where Bert provides factual information about the concepts POODLE and LABRADOODLE. I assume that such factual information is indeed encoded as such, and can be recalled when needed, for example to provide justifications for classification decisions. (As I argued before in (Schlangen, 2016).) This kind of knowledge weaves concepts into a network,[8] and provides the kind of information that can support resolving discourse anaphoric phenomena like the intended co-reference between "the poodle" and its antecedent "a dog" in (4).[9]

I also assume that connections between concepts are learned and represented in less explicit ways, namely from distributional information that gets reduced into vector representations of word meanings. (This is what the row of boxes in the concept files in Figure 1 is meant to represent.) I am swayed here not only by the psycholinguistic evidence that such distributional information figures in meaning representations (see, e.g. (Andrews et al., 2009; McDonald and Ramscar, 2001)), but also use this to account for the phenomenological difference between trying to produce a definition of a concept and making a fast judgement of similarity between concepts. (More on that below in Section 5.)

Finally, and connected to the second component described above, I assume that it is part of conceptual knowledge to have a rough idea about the domain or field of knowledge that a concept belongs to. This can then serve as a guide for disputes about word meanings (or "litigations" of word meaning, (Ludlow, 2014)) on how to weigh other people's opinions, with the limiting case being to defer to the expert (Keil and Kominsky, 2015).

Challenges Addressed? First, the desideratum of accounting for exophoric and endophoric reference. Here, in the basic cases, the responsibilities are divided: Exophoric reference is resolved via the classifiers that work on perceptual input, and the resolution of endophoric reference, insofar as it needs inference, is supported by the semantic or distributional knowledge in the concept file. I do however assume that cross-overs are possible: For example, in fine-grained visual classification (Dubey et al., 2018), recognition can happen via explicitly recalling distinguishing features (e.g., between bird species) and recognising those, deriving from this the visual classification decision.

This carries over to the learning of the components of the concepts files. While each component primarily seems to be connected with one type of learning situation described above (observing a reference intention with learning a classifier; receiving semantic information with learning semantic information; observing linguistic contexts with learning a continuous representation), initial experiments (see below) make us hopeful that cross-modal /

[6] This list is only the beginnings of the trouble with concepts, as (Murphy, 2002; Margolis and Laurence, 2015) helpfully review, but it shall be enough for this first start.

[7] Marconi (1997) provides a book-length argument for a similar separation which I took as an inspiration, but develop in a different technical direction here.

[8] Whereupon the file metaphor starts to falter a bit; but think of relational databases and their links between records.

[9] I leave possible connections to the "theory theory" model of concepts (Morton, 1980) on the one hand, and Donald Davidson's (1986) coherentism unexplored for now.

zero-shot learning can be realised. For example, visual recognition on first encounters of instances of otherwise familiar concepts ("book knowledge") can back off to recognition via properties stored in semantic knowledge; semantic knowledge can be induced from perceptual similarities; distributional knowledge can support visual classification, and be discretised into semantic knowledge.

Allowing these types of knowledge to be incrementally adapted, through interactions such as in the example from the introduction, opens up the private concept to public inspection and debate. This then connects the individual's concept with the public (but virtual) intension. Of this, at least in certain domains, experts can be the custodians. So, while concepts are in the head(s), *meanings* (understood as the public norms governing use), in this model "ain't" (Putnam, 1975).

Then, there are Hesperus and Phosphorus.[10] What happens when an agent, who previously didn't know it, learns that these are names for the same entitiy? In this model, suprisingly little: the agent just has to bring together the respective files. That two files have to be merged ("Hesperus is Phosphorus"$_{(8)}$) is valuable information, whereas touting the identity of a file with itself ("Hesperus is Hesperus"$_{(9)}$) is not. Similarly, accounting for attitude reports is straightforward: They index the file that the agent calls on, and the label they use to access it, and so substitution of a co-extensive term (e.g., turning "A believes (9)" into "A believes (8)") describes a different mental state.

Finally, *no-reference*. In this model, we can assume that concepts of this type (like UNICORN) have as part the knowledge that no instances of them will ever be encountered. Nevertheless, *pretend* semantic information, and real distributional information, can still be encoded, so that the concept can be used meaningfully.

But with all this, aren't concepts too fine-grainedly individuated? It looks like there now isn't just one type of sense, but many? Can two individuals ever have the same conceptualisation connected to the same word? (They can't have the same *concept*, as concepts are mental entities.) Here, I am inclined towards a pragmatist reply: Whether two individuals connect sufficiently similar concepts with the same word shows in the consequences of their having these concepts. And this is again where dialogue comes in, and learning. If the agents encounter a situation where they appear to be disagreeing on the applicability of a term, they can discuss, and, hopefully, reach an agreement that makes one, or both, of them adapt their concept. Where Fodor and Pylyshyn (2015) seem to assume that any single experiential difference between two agents must make their concepts incommensurable (if the inferential roles of the concepts are what individuates them), it seems to me unlikely that the learning methods sketched here, run on what is fundamentally very similar data (if both agents live in the same language community), even if on the instance level fully disjunct, yield dramatically different results that cannot be made commensurable in clarification dialogues of the type illustrated in the introduction. But this is an empirical question that a large-scale implementation could begin to test.

Implementation The components of the concept files individually are well-established in computational work. I have already mentioned some work on the words-as-classifiers model (Schlangen et al., 2016; Kennington and Schlangen, 2015) (see also Matuszek et al. (2012) for a related approach). Representing conceptual information in so-called *knowledge graphs* is a thriving field (see (Hogan et al., 2020) for a recent review); computing continuous representations of words from distributional information, following Landauer and Dumais (1997); Mikolov et al. (2013), perhaps even more so. (These last two fields, however, do not seem to be concerned much with incremental learning, which I take to be an indispensable part of any realistic model of situated interaction.)

Zero-shot learning, as it is called in the machine learning community, where information from one modality (e.g., text) is utilised for a task in another (e.g., visual classificaiton) is also a field that has relevance for this model. An approach like that of Lampert et al. (2014), where a visual classifier is constructed out of semantic knowledge directly offers a connection between these two types of knowledge in our concept files. Zarrieß and Schlangen (2017b) have explored the use of distributional information in the words-as-classifiers model, whereas Zarrieß and Schlangen (2017a) have shown that referentially structured contexts can yield improved continuous representations. A

[10] These, of course, are *names*, which in the model sketched in Figure 1 would be the labels of what we called *event / entity files*. For the purposes of this discussion, however, we can view them as concepts that mandatorily have a singleton extension.

thorough analysis of the structure of meaning negotiation interactions is still needed, but some pioneering (formal) work is being done (Larsson and Myrendal, 2017). Previously, I presented a proof-of-concept implementation that makes use of visual classifiers and semantic knowledge in simple justification interactions (Schlangen, 2016).

To sum up, parts, and partial connections, between the proposed components, have been explored in computational work. What is still missing is an attempt to fully combine these efforts along the lines sketched here.

5 Composition into Utterance Meanings

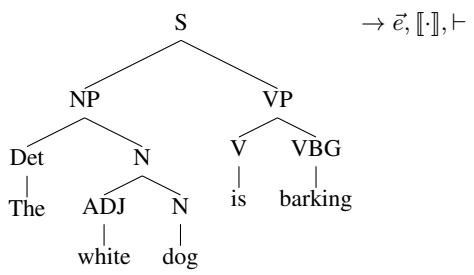


Figure 3: A single parse supporting three representations / uses

In keeping with the representational pluralism (or promiscuity?) of the previous section, I assume that there are several ways in which representations for larger expressions are computed, and used. The common basis for all of them is a syntactic analysis of the expression,[11] as in Figure 3.

First, I assume that a logical form is computed (e.g., (Sag and Wasow, 1999)) and integrated into the discourse representation (e.g., (Kamp and Reyle, 1993)), in the usual way. Besides other things, this representation forms the input to a within-context verification mechanism, which, for the visual context, is expressed via the vQUD as described above. For this, the interpretation functions for the non-logical constants are provided by the classifiers in the concept files. Following Schlangen et al. (2016), I assume that the contributions of the individual concepts are composed into a response for the whole phrase, on the basis of which (probabilistic estimates of) the denotations can be computed. For object-denoting expressions, this amounts to finding a link (or, ranking all possible links) to an object file. Finally, I also assume that a continuous representation for the expression is computed recursively (for example, through recursive neural networks, (Socher et al., 2014)), out of the continuous representations for the words / concepts.

This setup mirrors the tripartition in the concept files. What are the uses of these representations? I have already mentioned the role of the logical form in guiding the computation of (visually referential) denotations. Continuous representations, on the other hand, can provide a fast route to the computation of inferential relations between expressions (along the lines of the recently popular computational task of "natural language inference", (Bowman et al., 2015)). This could, for example, support the fast computation of coherence relations when integrating new content into the discourse (as is assumed by SDRT (Asher and Lascarides, 2003); and, as question-relevance, by KoS (Ginzburg, 2012)). In slower, explicit justifications of such judgements, the logical form can be used in sound, formal reasoning.[12]

Implementation In some preliminary experiments (to be reported in full elsewhere), we have parsed 50,000 captions from the COCO corpus (Lin et al., 2014), using the English Resource Grammar (ERG) (Flickinger, 2011) executed by the Answer Constraint Engine (ACE) (Packard, 2013) and accessed via pyDelphin.[13] In a task that uses captions from the test set as binary questions ("is this CAPTION?") paired with correct and incorrect images, we found that a reference computation approach based on the words-as-classifiers model can work satisfactorily. For a task that asks a model to predict whether two captions label the same image or not (used as a very rough proxy for entailment pairs), we computed representations for the captions using a TreeLSTM (Tai et al., 2015), and again found this to work satisfactorily.[14]

Extensions Before we move on, we note two possible further extensions of the model. First, it might be useful to allow the creation of object files

[11] We abstract here from the fact that a full model of situated discourse processing would need to work incrementally and hence this structure building would have to happen incrementally. However, we do think that the proposal here is broadly compatible with an incremental processing model such as for example that of Schlangen and Skantze (2009).

[12] It is tempting to see an approximation of Kahneman's (2011) System 1 / System 2 distinction here. How far this carries must be explored in future work.

[13] https://github.com/delph-in/pydelphin

[14] There is not enough space here to properly report these experiments, and so I won't attempt to report numbers here (which would not be interpretable anyway without the further detail). I only offer this information as anecdotal evidence that the approach could potentially be made to work.

also based on the discourse representation and not driven by perception. This could be used to model the state Chris is in after (3, "we saw a man carrying a dog") in our example in the introduction, where she can form a mental representation of the situation that is described. Schlangen (2019) describes a simple retrieval-based mechanism for constructing such imaginations of situations and their use in computing bridging relations (e.g., imagining "leash" when "dog" is mentioned). This, at least prima facie, seems to relate to the idea of understanding as simulation (Barsalou, 2012); to what extend needs to be explored in further work.[15] Interestingly, recent computational work on image synthesis from text also makes this distinction between predicting layouts (our object files) and predicting, based on that, the actual images (e.g., (Hong et al., 2018)).

The second extensions builds this out into the other direction (looking at Figure 1). We note that a composed expression representation (e.g., "the first president of the United States") could also be used to guide the recall of *entity files* from memory, via their stored semantic attributes. This might potentially address the definite description variant of Frege's puzzle (which we discussed above with names only), by making it possible for an agent to know of George Washington and still not retrieve the appropriate file via the description ("first president ..."), hence blocking substitution in a believe attitude report.

6 Back to The Story

We can now briefly return to the story from the introduction, and see how agents modelled according to the above would work here. Processing the various utterances involves the creation and linking of different kinds of mental representations, some of which stand in direct contact to the outside world (via visual indices), some of which are connected to memories of previous experiences. A proper test of this model would be an implementation (perhaps in simulation), showing that similar behaviour can indeed be created.

[15]This might even go some ways towards explaining the vast differences in how some people report phenomenal experience when imagining a scene (which in this model would mean "project back" the object files and their classifications into near-experiential impressions), while others lack this completely, but still seem to be able to simulate spatial relations. See Richardson (1999) for a recent overview of research on mental imagery.

7 Related Work

I have already mentioned in passing that there is a large amount of work that is related to the individual components which are brought together here. As a whole, the model probably has the most similarities with models from robotics on the one hand (e.g., (Kelleher et al., 2005; Kruijff et al., 2012), see (Tellex et al., 2020) for a recent overview), where multilayer (discourse and context) models are a salient choice. The literature is vast here, and the exact connections still need to be explored (see also older work on multimodality, e.g. (Luperfoy, 1992)). While the model likely re-invented parts already present in older work, some of the technical components that we aim to bring together are newer than most of this work, and together with the focus on learning, we think, bring a somewhat fresh perspective. On the formal side, the linking approach seems related to the idea of *anchoring* that has been discussed with respect to discourse representation (e.g., (Zeevat, 1999)).[16]

We also think that our proposal is broadly compatible with (and complementary to) approaches that more closely look into the role of coherence relations in situated discourse (Stone et al., 2013; Hunter et al., 2018), but this needs to be worked out. Lastly, the proposal also seems broadly compatible with the long-term efforts of the Gothenburg school (Larsson, 2015; Cooper, forth) to formally describe situated meaning, as well as with the dialogue model of Ginzburg (2012), but puts the emphasis more on the computational side.

8 Conclusions

I have outlined a model of how situated agents can deal with objects they encounter, in the environment, and in their discourses. It brings together elements from more formally oriented work (discourse representation, update), with methods from more processing oriented work (classifiers, distributed representations), in what hopefully at least roughly resembles a coherent whole. Many of the details are still missing, but a framework for where they can be filled in is provided.

Acknowledgements I have presented an even less fully developed version of this as part of an invited lecture series at the LabEx EFL Paris in 2019 (albeit with more details on

[16]This idea has already been present in the early DRT work (Kamp and Reyle, 1993); it is prominent in current unpublished work by Hans Kamp (2018) (as pointed out to me by one of the reviewers).

experiments), and I thank the audiences there for their patience and feedback.

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review*, 116(3):463–498.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Lawrence W Barsalou. 2012. Situated Conceptualization: Theory and Application. In Y Coello and Martin H. Fischer, editors, *Foundations of embodied cognition*, pages 1–17. Psychology Press.

Paul Bloom. 2000. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA, USA.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Susan Carey. 2009. *The Origin of Concepts*. Oxford University Press, Oxford, UK.

Robin Cooper. forth. *From Perception to Communication: An analysis of meaning and action using a theory of types with records*. na.

Donald Davidson. 1986. A coherence theory of truth and knowledge. In Ernest LePore, editor, *Truth and Interpretation, Perspectives on the Philosophy of Donald Davidson*, pages 307–319. Basil Blackwell.

Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ryan Farrell, Ramesh Raskar, and Nikhil Naik. 2018. Improving Fine-Grained Visual Classification using Pairwise Confusion. In *European Conference on Computer Vision (ECCV 2018)*, pages 1–17.

Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. In Jennifer E. Arnold and Emily Bender, editors, *Language from a Cognitive Perspective: Grammar, Usage, and Processing*. CSLI Publications, Stanford, CA, USA.

Jerry A. Fodor and Zenon W. Pylyshyn. 2015. *Minds without Meanings*. MIT Press, Cambridge, Massachusetts, USA.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Ztschr. f. Philos. u. philos. Kritik*, NF 100:25–50.

Jonathan Ginzburg. 1995. Resolving questions I. *Linguistics and Philosophy*, 18:459–527.

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

Irene Heim. 1983. File Change Semantics and the Familiarity Theory of Definiteness. In R. Bäuerle, Ch. Schwarze, and Arnim von Stechow, editors, *Meaning, Use and Interpretation of Language*, pages 164–189. De Gruyter, Berlin, Germany.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard de Melo, Claudio Gutiérrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2020. Knowledge graphs. *ArXiv*, abs/2003.02320.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. In *CVPR 2018*.

Julie Hunter, Nicholas Asher, and Alex Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11(10).

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. 1992. The Reviewing of Object Files: Object-Specific Integration of Information. *Cognitive Psychology*, 24:175–219.

Hans Kamp. 2018. Entity Representations and Articulated Contexts: An Exploration of the Semantics and Pragmatics of Definite Noun Phrases. Retrieved 2020-06-15.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Lauri Karttunen. 1969. Discourse referents. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*, Sånga Säby, Sweden.

Frank C. Keil and Jonathan F. Kominsky. 2015. Grounding concepts. In (Margolis and Laurence, 2015), chapter 24, pages 677–692.

J. Kelleher, F. Costello, and J. Van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167:62–102.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, Beijing, China. Association for Computational Linguistics.

Geert Jan Kruijff, Miroslav Janíček, and Hendrik Zender. 2012. Situated communication for joint activity in human-robot teams. *IEEE Intelligent Systems*, 27(2):27–35.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 36(3):453–465.

Thomas K Landauer and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211—-240.

Staffan Larsson. 2011. The TTR perceptron : Dynamic perceptual meanings and semantic coordination. In *semdial 2011*, pages 140–148.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of logic and computation*, 25(2):335–369.

Staffan Larsson and Jenny Myrendal. 2017. Towards Dialogue Acts and Updates for Semantic Coordination. In *Formal Approaches to the Dynamics of Linguistic Interaction*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing.

Peter Ludlow. 2014. *Living Words*. Oxford University Press, Oxford, UK.

Susann Luperfoy. 1992. The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of ACL 1992*, pages 22–31.

Diego Marconi. 1997. *Lexical Competence*. MIT Press, Cambridge, Mass., USA.

Eric Margolis and Stephen Laurence, editors. 2015. *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press, Cambridge, Massachusetts, USA.

Cynthia Matuszek, Nicholas Fitzgerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *ICML 2012*.

Scott McDonald and Michael Ramscar. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013)*, pages 3111–3119, Lake Tahoe, Nevada, USA.

Adam Morton. 1980. *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Clarendon Press, Oxford, UK.

Gregory L. Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA, USA.

Woodley Packard. 2013. The answer constraint engine (ace). Software package.

John Perry. 2012. *Reference and Reflexivity*, 2nd edition. CLSI Press, Stanford, CA, USA.

Hilary Putnam. 1975. The meaning of meaning. In *Mind, Language and Reality; Philosophical Papers Volume 2*. Cambridge University Press, Cambridge, UK.

Zenon Pylyshyn. 1989. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, 32:65–97.

Franois Recanati. 2012. *Mental Files*. Oxford University Press, Oxford, UK.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 779–788.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

John T. E. Richardson. 1999. *Imagery*. Psychology Press, Hove, UK.

Eleanor Rosch. 1978. Principles of Categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum, Hillsdale, N.J., USA.

Ivan Sag and Thomas Wasow. 1999. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford.

David Schlangen. 2016. Grounding, Justification, Adaptation: Towards Machines That Mean What They Say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.

David Schlangen. 2019. Natural language semantics with pictures: Some language & vision datasets and potential uses for computational semantics. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Gothenburg.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 710–718, Athens, Greece.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*, Berlin, Germany.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*.

Elizabeth Spelke. 1994. Initial knowledge: six suggestions. *Cognition*, 50(1-3):431–445.

Elizabeth S. Spelke and Katherine D. Kinzler. 2007. Core Knowledge. *Developmental Science*, 10(1):89–96.

Matthew Stone, Una Stojnic, and Ernest Lepore. 2013. Situated utterances and discourse relations. In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013 - Long Papers*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55.

Sina Zarrieß and David Schlangen. 2017a. Deriving continous grounded meaning representations from referentially structured multimodal contexts. In *Proceedings of EMNLP 2017 – Short Papers*.

Sina Zarrieß and David Schlangen. 2017b. Is this a Child, a Girl, or a Car? Exploring the Contribution of Distributional Similarity to Learning Referential Word Meanings. In *Short Papers – Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.

Henk Zeevat. 1999. Demonstratives in discourse. *Journal of Semantics*, 16(4):279–313.

Semantics with Feeling: Emotions for Abstract Embedding, Affect for Concrete Grounding

Daniele Moro
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
danielemoro@u.boisestate.edu

Gerardo Caracas
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
gerardocaracasur@u.boisestate.edu

David McNeill
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
davidmcneill@u.boisestate.edu

Casey Kennington
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
caseykennington@boisestate.edu

Abstract

An important yet underexplored aspect of meaning in both distributional and grounded models of semantics is emotion. In this paper, we explore how emotion can be predicted from descriptions of robot behaviors represented with embeddings. We then compare this approach with a grounded model that maps corresponding robot behaviors represented as internal states to the same emotion labels and discover comparable results. We then take the predictions from the second model and use them as a proxy for concrete affect (as opposed to abstract emotion) and use this derived affect to ground a semantic classifier in a retrieval task and see improvements on the retrieval task when affect is used as a grounded modality. This demonstrates that semantics can benefit from using a proxy of affect derived from human perceptions, given those perceptions are mapped to clear proxies, such the behaviors of an embodied robot.

1 Introduction

Semantic representation is a crucial part of language understanding for spoken dialogue systems, and the semantic meanings of many words have emotion as part of their connotative meaning (Lane and Nadel, 2002). Yet, including emotion in semantic models is complicated by the divide between distributional and grounded semantic representations, which rely upon separate assumptions for the source of meaning: distributional models only consider abstract meaning and grounded models only consider concrete meaning. Indeed, Bender and Koller (2020) and Bisk et al. (2020) observed that models of distributional semantics (i.e., embeddings) operate only on text and are missing key apsects of meaning. We therefore infer that distributional approaches make an *abstractness* assumption because the mode of acquisition for abstract language is other linguistic information (Della Rosa et al., 2010). Conversely, grounded semantic models make a *concreteness* assumption in that all semantic information can be acquired concretely by physical world denotations. Following Barrett (2017), emotions can similarly be dichotomized as abstract and concrete; abstract according their lexical categories (e.g., *happiness*, *fear*, *anger*) distributed with text, and concretely through *affect* which is a biological system and a fundamental part of embodiment (Vigliocco et al., 2014). In contrast to abstract emotion concepts, affect is a more basic underpinning for emotion, ranging from unpleasant to pleasant (valence) and from agitated to calm (arousal).

In this paper, we explore how emotion can be approached abstractly from embeddings derived from written descriptions describing a robot's behavior, and how affect can be approached concretely by deriving affect from the physical internal state data of a robot. We tie this to a semantic model by addressing the question of how a *grounded* semantic model could use a representation of affect and how this compares to existing work that links emotions to abstract semantic representations (like embeddings). In Experiment 1, we model abstract emotions from descriptions of robot behaviors in accordance with the abstractness assumption. In Experiment 2, we model concrete emotions from the robot's internal state data in accordance with the concreteness assumption. We hypothesize that both text descriptions and internal robot states can be used to classify emotion labels and conclude that these approaches are comparable. In Experiment 3, we take the model from Experiment 2 and use the distribution over emotion labels derived from internal robot states as a feature vector of affect for a grounded semantic model. The results show that the semantic model can effectively ground into our representa-

tion of affect.

2 Related Work

Our work relates to other recent work in connecting abstract emotions with embeddings. Xu et al. (2018) introduced Emo2Vec, a model of representing emotion as an embedding using multi-task learning. Agrawal et al. (2018) enriched word representations with emotional information, taking into account the fact that emotion words are distributed in similar ways in text, but have vastly different underlying emotional affect (e.g., *sad* is distributionally similar to *happy*). Similarly, Saravia et al. (2018) introduced CARER, a semi-supervised approach for representing lexically contextualized affect, enriched with embeddings. Finally, Alhuzali et al. (2018) extracted emotional information in text using Arabic dialectic first-person seed words. Ongoing work in emotion representation continues to follow the distribution hypothesis (i.e., assuming the semantics of all words are abstract), comparable to our findings in Experiment 1, but we move beyond this work by exploring how words can ground into an extra-linguistic affective representation.

Also related to our work is research on grounding semantic meaning of words into perceptual modalities beyond vision, such as auditory (Kiela and Clark, 2015) and olfactory perception (Grabski et al., 2012), haptics (Alomari et al., 2017; Thomason et al., 2016), and simulated hand muscle activations (Moro and Kennington, 2018). More directly related to our work is Song and Yamada (2018), which reported a multimodal approach in predicting a human's label of robot affect from seven basic emotions, as well as overall valence and arousal. Our work adds to a growing literature around a notion of embodied semantics (Johnson, 2008; Goertzel et al., 2010)–our experiments show that abstract words can ground more readily into representations of affect and emotion.

Our work builds directly off of McNeill and Kennington (2019), which explored how humans interpret affective display of robot behaviors, and Novikova et al. (2015) which explored how dynamic behaviors can be represented for mapping to emotions. We extend their work by linking dynamic robot modalities to emotion labels, and also to text descriptions of those robot behaviors.

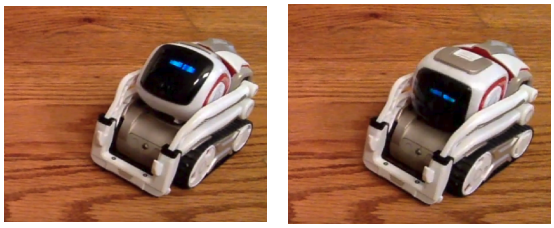


Figure 1: Two example frames of a recorded behavior.

3 Models

To model language abstractly for Experiment 1, we use the BERT model introduced in Devlin et al. (2018), which has been leveraged in many language tasks, in many cases resulting in state-of-the-art performance.

To model grounded semantics for Experiment 3, we use the *words-as-classifier* (WAC) model for grounded lexical semantics (Kennington and Schlangen, 2015) due to its simplicity and interpretability. The WAC model uses a task-independent approach to predicting the semantic appropriateness of a word given a physical context. It pairs each word w in its vocabulary V with a classifier, and this classifier maps the real-valued features x of an experience exp (e.g., information recorded by a robot's sensors) to a semantic appropriateness score (i.e., if the information belongs to the class given by a word):

$$[\![w]\!]_{exp} = \lambda \mathbf{x}.p_w(\mathbf{x}) \quad (1)$$

For example, to learn the grounded meaning of the word *turn*, the low-level features (e.g., wheel motor speed) of a robot behavior described by the word *turn* are given as positive instances to a supervised learning classifier. Negative instances are randomly sampled from the complementary set of descriptions (i.e., behaviors not described by the word *turn*). This results in a trained $\lambda \mathbf{x}.p_{turn}(\mathbf{x})$, where x is the feature-set of a behavior that can be described by $turn$.

4 Data

For all of our experiments, we use data introduced in McNeill and Kennington (2019) that used the Anki Cozmo robot (see Figure 1).[1] Cozmo was marketed as a toy robot for children, is small in size, and the SDK allows developers to access low-level information about the robot state and to

[1] https://github.com/bsu-slim/cozmo-affect-data

control several degrees of freedom including animated eyes, a head and lift that can move up and down, wheels that can be used to turn and move the robot forward or backward, and a speech synthesizer. Cozmo has a built-in camera and simple built-in object and facial detection software. Following McNeill and Kennington (2019), we use the data they collected by asking Amazon Mechanical Turk workers to observe and write English descriptions and label of Cozmo's 941 prescripted behaviors for emotion. These behaviors include movements, sounds, and facial animations that are easily observable by a person.

Their data collection resulted in 1,870 descriptions and 16 **emotion labels** (*interest, alarm, confusion, understanding, frustration, relief, sorrow, joy, anger, gratitude, fear, hope, boredom, surprise, disgust, desire*), at least two for each of the behaviors (their work focused on the emotion labels; we use both the workers' emotion labels and their written descriptions). We normalized the descriptions by lower-casing all text, replacing punctuation with spaces, and removing any non-alphanumeric characters, resulting in a vocabulary size of 1230 words. The average phrase length was 7.9 words. The most common reported emotion label was `interest` (22% of phrases). The original dataset has three modalities: audio, facial, and internal states which represent Cozmo's animation modalities for each behavior. In this paper, we focus only on internal states (due to the feature transform we apply below). Each state update recorded a vector of 47 continuous feature values. The number of state updates varied for each behavior; i.e., a behavior could have as few as one state update vector, or as many as over a thousand state update vectors.

We augment their dataset with features explained in Novikova et al. (2015) (see Table 1 in that paper) which we term *Novikova* features, that we use in Experiments 2 & 3. These features are a functional transformation from the internal states, (which could range from a few state updates to as many as two thousand state updates for a particular behavior) to a set of 9 features:[2]

- Approach 1 - Transfer weight forward (head bent or movement forward)
- Approach 3 - Move its body forward (track wheel movement forward)
- Approach 5 - Extend or expand its body (lift movement up)
- Avoidance 6 - Transfer weight backward (head bent or movement backward)
- Avoidance 8 - Move its body backward (track wheel movement backward)
- Avoidance 9 - Attract limbs close to body (lift movement down)
- Energy 11 - High strength (high wheel speed)
- Energy 12 - Low strength (low wheel speed)
- Flow 18 - High change in tempo (change in motor speed)

Each feature yields a value that is a percentage of the time that feature is true. For example, Approach 1 is the percentage of robot state changes with forward movement and Avoidance 9 is the percentage of the state updates where the lift was in the lower half. Taken together, these transformations result in a vector of 9 values, each value between 1 and 0. Using these features has the added benefit of being generalizable to other robots; one only needs to map internal state representations of their chosen robotic platform to the Novikova features for modeling emotion.

Example Figure 1 shows two frames of a recorded behavior corresponding to Example (1) below, which includes (a) one worker's typed description of the behavior, (b) that worker's chosen affects to label the behavior, (c) a table with a sample of 4 internal features at one state update, and (d) Novikova features.

(1) a. The robot comes closer to the camera, showing a desire to know what is going on. Then it slams down its arms, like it wants to know.

b. sad, frustrated, curious

c.

feature	value
left wheel speed	0.0
right wheel speed	0.0
lift height	4.0
head position	3.0
...	

d. [0.11, 0.27, 0.97, 0.8, ..., 0.34]

Analysis of Novikova Features Figure 2 shows an application of TSNE to our data, which are represented as the Novikova features, annotated by color for each emotion. There are some clear clusters for some of the emotions, such as frustration, surprise, alarm, boredom, interest, and sorrow, indicating that many of the robot behaviors are cleanly separated, but there are many instances where a behavior does not belong to a particular cluster, indicating noise from human judgements of emotion of robot behaviors can be subjective,

[2] We excluded the other 14 features because they did not change the results for a subset of Experiment 2; i.e., for our data, they were interpolations of the 9 remaining features.

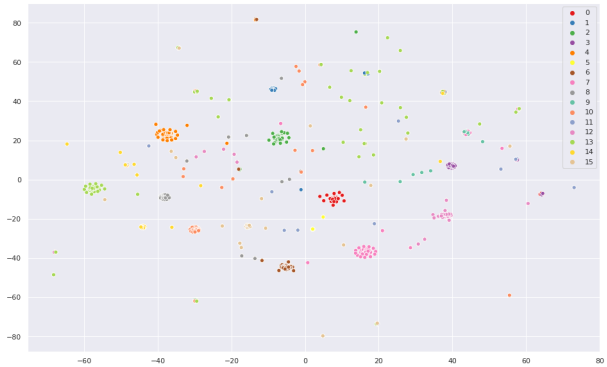


Figure 2: TSNE applied to all robot behaviors represented by the Novikova features colored for each of the 16 emotion labels. 0=interest, 1=alarm, 2=confusion, 3=understanding, 4=frustration, 5=relief, 6=sorrow, 7=joy, 8=anger, 9=gratitude, 10=fear, 11=hope, 12=boredom, 13=surprise, 14=disgust, 15=desire

but potentially useful if a classifier can pick up on differences, which we explore in Experiment 2.

5 Experiments

5.1 Experiment 1: Classifying Emotion Labels from Descriptions

In this experiment, we explore how language maps to abstract emotions; that is, by using the BERT model we are making the abstractness assumption on the descriptions, and we are assuming that emotions are labels of abstract concepts. As has been argued by Barrett (2017), emotions are indeed abstract categorical labels that are derived from affect and are tied into the abstract linguistic system. By mapping embeddings to a distribution over 16 emotion labels, which themselves are a representation of the human judgment of emotion, we can ascertain if emotion is a suitable modality that abstract language can map to.

Task & Procedure For this experiment, we evaluate a model that takes as input a written description of the robot behaviors, and outputs a distribution over the 16 emotion labels. The data was split into training and testing groups using a random 90:10 split.

Model and Training We used the BERT model introduced by Devlin et al. (2018), as BERT has been shown to improve over traditional recurrent models. Instead of fine-tuning the pre-trained BERT model, we use a feature-based approach that has been shown to approximate fine-tuning, where "fixed features are extracted from the pre-trained model" (Devlin et al., 2018). The embeddings extracted from the pre-trained BERT transformer encoder (pre-trained on BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words)) are padded to a length of 80, then fed into a two-layer feedforward neural network consisting of 128 nodes followed by 16 nodes (the small network is a result of our small amount of data). We trained using a binary crossentropy loss function and the adam optimizer, for a period of over 800 epochs (hyperparemeters were chosen empirically using a subset of the training data).

Accuracy	Precision	Recall	F1 Score
90.4%	69.6%	31.4%	43.3%

Table 1: Results from Experiment 1 model when predicting 16 binary emotion labels on the testing set using a threshold of 0.5.

Metrics We report accuracy, precision, recall, and F1 score of the trained model, assigning a probability above 0.5 to each label (i.e., more than one label could be correct) using a micro averaging strategy.

Results The results for this experiment are displayed in Table 1. These results show that our model can effectively map a description of a robot action to a distribution over emotion labels with relatively little data. Analysis of a confusion matrix shows that most of the errors are presented in the form of *false positives*. This could be caused by multiple users labeling the same robot action with conflicting labels.

5.1.1 Analyses

Here we report further analyses to determine what the model in this experiment is learning.

Synthetic Descriptions To explore whether the model predicts what we would expect, Table 2 shows synthetic descriptions that we applied to the model and observed the affect with the highest probability as assigned by the model.

Word-Emotion Associations Focusing on the word-level, we analyze what our model is learning by determining which words correspond to specific emotion labels (i.e., we applied single words of our vocabulary to the model and observed the resulting distribution over the emotion

Description	Predicted Affect
the robot looks down	fear
the robot looks up	interest
the robot moves away	boredom
the robot moves closer	interest
the robot shakes his head	frustration
the robot nods his head	understanding
the robot squeals	gratitude
squinting eyes	confusion

Table 2: Examples of predicting emotion labels from descriptions of robot actions.

labels). Table 3 shows several noteworthy examples of words with their corresponding top three ranking emotion labels in the model's resulting distribution after applying that word. As expected, *sorrow*'s most probable emotion label is `sorrow` (as was also the case for *understanding*, *confusion* and other emotion words), but other words that aren't directly related to emotion label also showed strong predictions. The word *eyes* corresponds to `confusion` and `alarm` (i.e., eyes furrow or widen, respectively) and when the robot lowers its lift, it is interpreted as `sorrow` (i.e., disappointment) or `desire` (i.e., lowering the lift keeps the lift out of the robot's camera view so it can better observe objects).

Word	Top Affects
sorrow	sorrow (0.62), relief (0.14)
eyes	confusion (0.47), alarm (0.3)
like	relief (0.49), understanding (0.48)
forward	hope (0.85), understanding (0.74)
quickly	alarm (0.83), understanding (0.56)
lowers	sorrow (0.51), desire (0.32)

Table 3: Words and our model's highest corresponding predicted emotions.

Ablations We performed an ablation analysis on our model where we modified the descriptions of the robot that our model uses to train. We did this by removing or retaining stop words, nouns, verbs, adjectives, and adverbs. The part-of-speech tags of the words in the descriptions were chosen using TextBlob, and the model was trained as explained in 5.1. We use the same metrics (i.e., accuracy, precision, recall, f1 score) as Experiment 3.

The results are displayed in Table 4, demonstrating that the base model and the version trained with no stop words perform the best according to all metrics, as expected. The version of the model trained with only stop words performed second-worst given the F1 score, as we expect that no relevant emotional information is conveyed in stop

	Acc	Prec	Recall	F1
Base	**90.4%**	**69.6%**	**31.4%**	**43.3%**
No Stop Words	89.9%	63.9%	30.3%	41.1%
Only Stop Words	87.8%	38.2%	7.4%	12.4%
Removed Nouns	88.8%	56.1%	18.3%	27.6%
Only Nouns	87.6%	41.2%	16.0%	23.0%
Removed Verbs	89.5%	64.4%	21.7%	32.5%
Only Verbs	87.5%	31.4%	6.3%	10.5%
Removed Adj	90.2%	68.5%	28.6%	40.3%
Only Adjectives	89.2%	67.6%	14.3%	23.6%
Removed Adverbs	90.3%	68.8%	30.3%	42.1%
Only Adverbs	87.9%	18.2%	1.1%	2.2%

Table 4: Results from ablation study, where descriptions of the robots's behaviors were modified before training.

words. The version of the model trained on only verbs had a slightly worse F1 score, indicating that verbs may not carry meaningful information for a model that is predicting perceived emotion. When comparing F1 scores of the versions of the model trained on only nouns or only verbs, it is clear that nouns are much more important to such a model than verbs, and this is further demonstrated by the high F1 score in the version of the model where verbs were removed. This shows that the composed descriptions represent complex language– the composed description is a better predictor of emotion.

Correlation with Sentiment We compared our results with a sentiment classifier to determine if simple sentiment classifiers capture nuances of emotion in different types of written text. We calculated sentiment scores with our model by first predicting the 16 emotions split into positive/negative valence pairs (as done in (McNeill and Kennington, 2019): *interest/alarm, understanding/confusion, relief/frustration, joy/sorrow, graditude/anger, hope/fear, surprise/boredom, desire/disgust*, then multiplied the probabilities of the 8 negative emotion labels by -1, then averaged the scores to obtain a sentiment score between -1 and 1, making our scores comparable to the sentiment scores given by the TextBlob sentiment classifier. We then calculated the correlation of determination between the sentiment scores from our model and TextBlob. The R^2 result was 0.11, indicating a poor correlation. We believe that our model learns robot-specific sentiment and context that a general model does not account for. We also note that the descriptions for this task represent an observer's interpretation of a behavior, whereas sentiment classification is tasked with the

emotional content within a written text; this shows that applying sentiment classification may not help in certain contexts. This point is further reinforced by a high R^2 value of 0.84 between the sentiment scores calculated from the true values in the data and our model, and a R^2 value of -0.51 between the sentiment scores calculated from the data and the sentiment scores from TextBlob.[3]

5.2 Experiment 2: Classifying Emotion Labels from Robot Behaviors

In this experiment, we use the robot states represented as Novikova features to train and evaluate a classifier to predict what people perceive about the emotional display of Cozmo's behaviors. The purpose of this experiment is two-fold: (1) to compare with the results of Experiment 1 to ascertain if robot behaviors represented as internal state updates contain comparable information to descriptions of those same behaviors (i.e., the concreteness assumption compared against the abstractness assumption), and (2) to build a model of affect that a grounded semantic model can use to ground into for Experiment 3.

Task & Procedure Given features that represent a robot behavior (i.e., the Novikova features described above) predict one or more of the 16 emotion labels. We perform a 10-fold cross validation and average the results of each fold. No individual behavior was represented both in the training and test set for each fold.

Model & Training Because our feature vectors are small, and because we do not have much training data, our approach uses a multi-label K-Nearest Neighbor (KNN) classifier (Zhang and Zhou, 2007) to map from features to a distribution over class labels (tests using other classifiers, such as neural networks, resulted in worse performance, likely due to the nature of the features and small amount of training data).[4] The only parameter that was needed for the KNN classifier was number of neighbors, which we set to 5 to balance generalizability and performance in our task.

Metrics To compare directly to prior work and Experiment 1, we report two metrics to give an overall understanding of how our model performs this classification task in decreasing degrees of constraint:

average accuracy: Because any given behavior could have multiple labels, we take the prediction distribution and for every affect that received a probability of more than 0.5 in that distribution (0.5 was determined using the development set), we counted that as a positive guess for that affect. We compared this to the labels (i.e., by comparing two binary vectors) and compute the accuracy for each behavior, then take the average. This will seem inflated as many zeros in the binary vectors will increase the accuracy, but it allows our model to predict multiple labels, which better follows what people do when interpreting emotions (i.e., this is the same as micro averaging, as done in Experiment 1). ***F1 fscore***: We compute the F1 score using micro averaging. We also report precision, recall, and F1 score for each individual affect.

The baseline we are comparing against is the best resulting setting and ablation reported in McNeill and Kennington (2019) (termed M&K below). Note that their work used the raw features from the robot including facial, audio, and internal state features. Their model was a multi-layer perceptron.

5.3 Results

The results are shown in Table 5 and Figure 3. The former shows a direct comparison to M&K; we are reporting state-of-the-art results in both metrics despite only using internal state updates (their work included facial and audio features, as produced by the robot) and our classifier is much simpler. Moreover, the use of a KNN classifier suggests that there are vectors (at least for the Cozmo robot) which are prototypical for a particular affect, even if individual behaviors for each affect appear to be very different. This is further evidenced by the individual results in Figure 3, which show high overall for each individual emotion (with the exception of boredom). Comparing the results to Experiment 1, we find that, given a succinct representation like the Novikova features, mapping from internal states to emotions performs comparably to a mapping from words (i.e., in the descriptions) to emotion. We are therefore able to capture an equivalent amount of information by applying either the concreteness as-

[3] We opted to not compare with other fine-grained emotion models like Abdul-Mageed and Ungar (2017) because because the goal is to compare the model from Experiment 1 with the model from Experiment 3.

[4] Multi-label KNN uses Bayesian Inference to compute a probability for each label.

Approach	Assumption	Feature Set	Avg accuracy	Avg f1
M&K	Concrete	raw internal states	71.0%	55.0%
Exp. 1 (BERT)	Abstract	text description	90.4%	43.3%
Exp. 2 (KNN)	Concrete	Novikova features	**91.0%**	**57.0%**

Table 5: We predict emotion labels using various feature sets derived from either the abstractness or concreteness assumption. Results show that our KNN model using the Novikova features (exp 2) performs better than the McNeill and Kennington (2019) prior work on the same task. Results of using the concreteness assumption in Exp. 2 are comparable to the abstractness assumption using text description as the features set in Exp. 1.

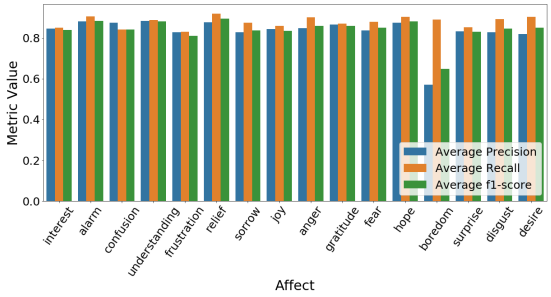


Figure 3: Average precision, recall, and F1 score for each emotion label for Experiment 2.

sumption or the abstractness assumption to this task. Taken together, the model from this experiment which maps from internal states (represented as Novikova features) to a distribution over 16 emotions could potentially be used as a simulated affective representation that a grounded classifier could ground into, which we explore in Experiment 3.

5.4 Experiment 3: Grounding Descriptions into Robot Behaviors and Affective States

In this experiment, we seek to answer the question: *Can a model for grounded lexical semantics ground into internal robot states as well as a representation of affect derived from those internal states?* We use the WAC model explained above as the model of grounded lexical semantics, the Novikova features as the robot state representation, and the trained model from Experiment 2 as a vector representation of affect (i.e., we treat it was an emotional substrate to ground into).

Task & Procedure Given the words in a description, we randomly select n distractor behaviors from the test set and apply the features of those distractors along with the gold behavior to the classifiers for each word in the description using the WAC model. Following Schlangen et al. (2016), the sum of the probabilities from these classifiers results in the semantic appropriateness score for each behavior. The model correctly identifies the gold behavior when that behavior is assigned the highest score.

We repeat this for 1-19 distractors d and report the accuracy for each d using a 10-fold cross-validation. Each behavior is represented twice in the dataset (once for each description), but we ensure that the same behavior never occurs in both train and test sets. We use words that are represented 5 or more times in the training data, resulting in a vocabulary of 347 words. Based on evaluations from one of the folds, we determined that for each positive instance of a word in the corpus, we should randomly choose 3 negative instances.

Model & Training Prior work using WAC has traditionally used logistic regression classifiers and multilayer perceptrons. Here, we take inspiration from Experiment 2 and use KNN for the WAC classifiers, as they are using the same set of features as the KNN classifier in Experiment 2 used. We trained an individual KNN for each word in our vocabulary with the nearest neighbor parameter set to 5 to balance generalizability and fitness.

A shortcoming of WAC is the independence assumption which fails to capture more complex semantic patterns within the text. However, the purpose of this experiment is not to capture the various semantic dependencies within the text, but to measure the comparative strength of our feature sets. WAC provides a simple and effective means by which to compare constructed affect to its lower-level source: robot state data.

We report results for when WAC only used the Novikova features directly (i.e., grounding into internal robot states represented as the 9 Novikova features), and also using the trained model from Experiment 2 as a representation of affect (16 features) for a total of 25 features.

Metrics The metric we report is accuracy of the model in choosing the correct retrieved behavior from the list of distractors, given the descrip-

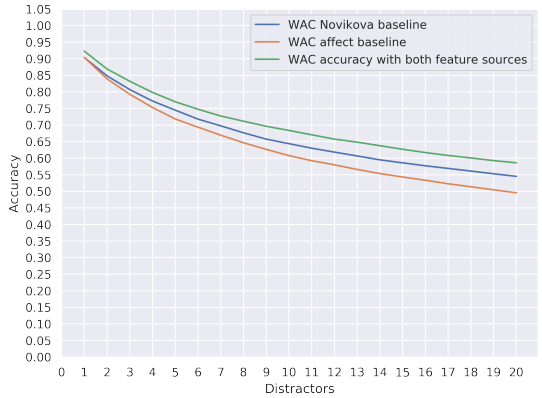


Figure 4: Experiment 3 Results: WAC applied to Novikova and affect feature types; accuracy for each ablation compared to number of distractors.

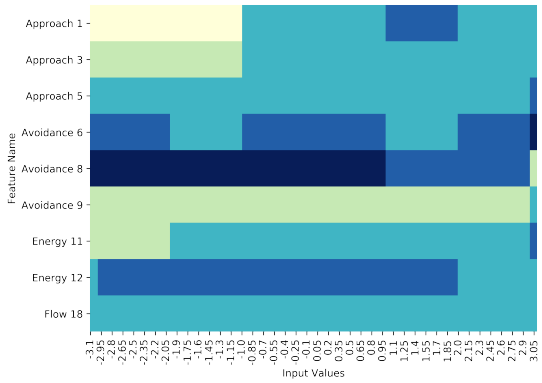


Figure 5: Heat map generated by ranging over all possible features and applying them to the WAC classifier for the verb *turns*; darker colors denote higher probabilities.

tion. The baseline for this experiment are the the other two ablations: when WAC model performance without emotion labels, and the WAC model that only uses emotion labels.

Results The results are shown in Figure 4. Overall, the model performs with high accuracy when there is only 1 distractor and we observe decreases in accuracy as the number of distractors increases. The WAC classifier performs well on its own given the 9 Novikova features alone, but it performs noticeably better when including our derived representation of affect from Experiment 2 (particularly when there are a larger number of distractors). Our results show that a model of grounded semantics can effectively ground into a predicted representation of affect. This is a very informative result: automated systems clearly don't have their own intrinsic biological / chemical affect, yet semantics can benefit from using a proxy of affect derived from human *perceptions*.

Analysis of Verbs In this section, we show how WAC learns verb semantics from the Novikova features by ranging over all possible values for each feature (i.e., we passed the Novikova features through a feature normalization transform, resulting in features ranging from -3.1 to 3.1), holding all other features to 0 when they are not being ranged over and using a WAC classifier to classify each possible combination of features.

Figure 5 illustrates what the *turns* WAC model learned by using a heatmap (darker colors denote higher probabilities returned by WAC). We can see from the figure that higher values in Avoidance 8 (i.e., wheel movement) is the most informative feature, where most values that denote some change in movement result in higher probabilities for the *turns* WAC classifier. From this we learn that a verb (used to denote a robot action, such as *turn*) will have the expected feature representation using the Novikova features in a KNN classifier. We noticed similar expected results from other verbs such as *tilts* (high sensitivity to Approach 3), *plays* (wide range of values for all features between Approach 3 and Flow 18) and *jerks* (high sensitivity to any feature relating to head and light movement with high energy).

6 Conclusion

This work represents a contribution to a growing shift towards human-centered, affective computing (Picard, 2000; Mohammad and Ovesdotter Alm, 2015). People who use natural language are also emotional beings; this has implications for the kinds of systems researchers develop that humans will interact with, as well as how the semantics of words will be acquired, represented, and applied in those systems; in particular, dialogue systems that are used for spoken interaction with multimodal agents, such as robots. This work helps shed light on how leveraging emotion and affect leads to more accurate semantic models.

Our work agrees with the claims in Lücking et al. (2019) which explains how distributional representations are not enough. The partial meaning of many words can be derived abstractly from nearby words (Firth, 1957), but words keep company with physical objects, entities, and situations as well as with other words. The complete semantic meaning of many words includes the

world, perceived and represented through physical modalities such as affect.

For future work, we will expand our repertoire of robot emotional displays by producing novel emotional behaviors for different robot platforms. We will perform studies to determine what kinds of emotional behaviors make a robot more amenable to performing collaborative tasks with humans, specifically tasks where the robot must learn new words from human collaborators and what those words denote. Our work has implications for how humans interact with robots and the semantics of the words that robots learn from humans: the meaning of a word is a function of what that word refers to in a scope of lexical and physical context.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning Emotion-enriched Word Representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018. Enabling Deep Learning of Emotion With First-Person Seed Expressions. In *Proceedings of the Second Workshop on Computational Modeling of People{'}s Opinions, Personality, and Emotions in Social Media*, pages 25–35, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural Language Acquisition and Grounding for Embodied Robotic Systems. In *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

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*.

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*.

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. *Behavior Research 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.

John R Firth. 1957. *Papers in Linguistics, 1934–1951*. Oxford University Press, Oxford, UK.

Ben Goertzel, Cassio Pennachin, Samir Araujo, Fabricio Silva, Murilo Queiroz, Ruiting Lian, Welter Silva, Michael Ross, Linas Vepstas, and Andre Senna. 2010. A general intelligence oriented architecture for embodied natural language processing. In *Artificial General Intelligence - Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010*, pages 13–18.

Krystyna Grabski, Laurent Lamalle, and Marc Sato. 2012. Contrle prédictif et codage du but des actions oro-faciales. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP*, pages 289–296.

Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.

C. Kennington and D. Schlangen. 2015. Simple learning and compositional application of perceptually groundedword meanings for incremental reference resolution. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 1.

Douwe Kiela and Stephen Clark. 2015. Multi-and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal.

Richard D Lane and Lynn Nadel. 2002. *Cognitive Neuroscience of Emotion*. Oxford University Press.

Andy Lücking, Robin Cooper, Staffan Larsson, and Jonathan Ginzburg. 2019. Distribution is not enough: going Firther. In *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, April, pages 1–10. Association for Computational Linguistics.

David McNeill and Casey Kennington. 2019. Predicting Human Interpretations of Affect and Valence in a Social Robot. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany.

Saif Mohammad and Cecilia Ovesdotter Alm. 2015. Computational Analysis of Affect and Emotion in Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Lisbon, Portugal. Association for Computational Linguistics.

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*.

Jekaterina Novikova, Gang Ren, and Leon Watts. 2015. It's Not the Way You Look, It's How You Move: Validating a General Scheme for Robot Affective Behaviour. In *Human-Computer Interaction – INTERACT 2015*, pages 239–258, Cham. Springer International Publishing.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

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.

Sichao Song and Seiji Yamada. 2018. Designing Expressive Lights and In-Situ Motions for Robots to Express Emotions. In *Proceedings of the 6th International Conference on Human-Agent Interaction - HAI '18*, pages 222–228, New York, New York, USA. ACM Press.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing " I Spy ". In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Gabriella Vigliocco, Stavroula Thaleia Kousta, Pasquale Anthony Della Rosa, David P. Vinson, Marco Tettamanti, Joseph T. Devlin, and Stefano F. Cappa. 2014. The neural representation of abstract words: The role of emotion. *Cerebral Cortex*.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298, Brussels, Belgium. Association for Computational Linguistics.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*.

Overcoming distances – the usage of calling contours to initiate dialogues*

Jakob Maché
Faculdade de Letras
Universidade de Lisboa
jakob.mache@letras.ulisboa.pt

Abstract

Based on the behaviour of calling contours (CC) in interaction with response particles such as *yes*/*no*, it will be shown here that intonation stripped off of any phonetic strings of word forms may convey propositional meaning, that they involve a great deal of compositionality and that they thus can be most accurately analysed as non sentential utterances along the lines of Ginzburg (2012): As such they are (i) sensitive to the speaker's (private) assumptions about the addressee's situation and (ii) sensitive to previously (publicly) established facts and that hence their semantics can only be understood if their interactive nature in dialogues as is fully acknowledged. Thus it will be demonstrated that a formalization of these intonation contours as construction-like phrasal signs which take names or definite descriptions as daughters, or as in the case of 'routine' calls, a whole range of utterance types. Such an analysis is necessary for other stylized intonation such as the Children's Chant and for languages in which illocutionary forced is largely determined by intonation. The analysis will be formulated within the *Type Theory with Records* (Cooper, 2005a,b) and *Conversation oriented Semantics* frame work (Ginzburg, 2012), which provide the required tools capable to account for any aspect of the behaviour of CC in discourse.

1 Introduction

Various languages have specific intonational means to attract the attention of a referent who is not engaged into the conversation with the speaker yet, such as Polish (Arvanati et al., 2016), Hungarian (Varga, 2008), French (Fagyal, 1997), Catalan (Borràs-Comes et al., 2015), Bengali (Hayes and Lahiri, 1992) and even some languages with lexical tone, such as Tianjin Mandarin (Zhang, 2018). This is typically done by calling the addressee's name, which involves in most (Indo-)European languages the use of a distinctive intonation contour. Such patterns are known as calling contours (alternative names: vocative chant, minor third, down-stepped level terminal contour/DLT).

Just as greetings they serve as initial moves to invite an addressee referent to engage into a communication. Despite the fact they are very brief, they involve a whole range of phonological and semantic idiosyncrasies.

They have received a vast amount of attention from phonetic (Ladefoged and Johnson 2011, 126–127 and Niebuhr 2013) and phonological perspective (Liberman 1975, 30–32, Gibbon 1976, 274–287, Ladd 1978, 520–524, Pierrehumbert and Hirschberg 1990, 293–294 and many more), but their precise effects and requirements in dialogues have not been addressed in a systematic way yet; the previous semantic studies by Truckenbrodt (2012, 2045–2048) and Condoravdi and Sunwoo (2017, 2018) restrict themselves on a small number of possible utterance types. This paper will provide an analysis that explains the semantic behaviour in dialogues of the complete range of known CC, including interrogative CC.

The Viennese variety of German provides an particularly interesting case, as it has contours which appear not to be present in other languages. It is highly likely that many aspects of the analysis sketched below can be applied to other (Indo-) European languages as well, but probably not all of them.

[0]I am grateful to the three anonymous reviewers for their enlightening hints, to Jonathan Ginzburg for comments on earlier versions, to Timo Buchholz for comments on intonation, to Felix Bildhauer, Antonio Machicao y Priemer and Ìyá Ọ̀ṣun.

2 Calling contours

In this paper, two major types of CC will be distinguished: Attention request calls (1), and interrogative calls (2). Apart from that there are also exclamations of surprise or disapproval, which are related to that phenomena, but exhibiting different effects on the discourse (3): The examples below demonstrate situations of the type in which the speaker calls the name *Friederike*.

In *Autosegmental Metrical Theory* developed by Ladd (2008, 116–119), calling contours are the paradigmatic case for the down-stepped boundary tone !H-%. *German ToBi* (*GToBI*) authored by Grice et al. 2005 favours this approach departing from original *ToBI* advocated by Pierrehumbert and Hirschberg (1990) and others, which only assumes two distinctive tonal levels for pitch accents and boundary tones: L and H. Likewise the non-canonical assumption of the complex boundary tone L-H% is motivated by Grice et al. (2005, 68–69) in their *GToBI* system for final fall-rise contours preceded by H tones. The analysis sketched here is neutral with respect to these two schools of thoughts, but for the sake of consistency Ladd's standard *GToBI* will be employed.

(1) a. fri:. də. ˈri:. kə
L+H* !H-%
(i) ...I am calling you!

b. fri:. də. ˈri:. kə
L+H* L-%
(i) ...Be careful!

(2) a. fri:. də. ˈri:. kə
H+L* H-%
(i) ...Is it you?
(ii) ...May I talk to you now?

b. fri:. də. ˈri:. kə
L*+H L-H%
(i) ...Will you stop doing that?

(3) a. fri:. də. ˈri:. kə
%H H*+H L-%
(i) ...I am so surprised to see you here!

b. fri:. də. ˈri:. kə
%H !H* L-%
(i) ...What have you done!

Starting with Ladd (1978, 520–524), the patterns of the type in (1-a) have been dubbed 'routine' calls (RCC), as they are limited to utterance situations which involve a certain familiarity between the participants and the situation and it has been attested in varieties of languages as diverse as Tianjin Mandarin (Zhang, 2018), a language with lexical tone, and Bengali (Hayes and Lahiri, 1992). Quiroz and Żygis (2017, 1211) noted that there are calls specific to situations in which the speaker has more urgent message to deliver (1-b) and others in which the nature of the call is more 'stern' and which they analyse as L*+H L-H% (2-b).

Unlike Quiroz and Żygis (2017), we will classify the latter 'stern' type as belonging to the class of interrogative calls. The first type of interrogative calls expressed by a H+L* H-% contour in (2-a) is reminiscent of reprise fragments (Ginzburg, 2012): they can either be requests to confirm the identity of the addressee (henceforth identity confirmation request calling contour, ICRCC), or they can be interpreted as requests for the addressee's availability to engage into a conversation with the speaker (henceforth, availability request calling contours ARCC). As indicated by the high boundary tone H-%, which is typical of polar questions in German and which serves often among other means to differentiate between questions and assertions (cf. Grice et al. 2005, 70–74, Peters 2018, 95), the most likely interpretation of the pattern in (2-b) is an interrogative one, as a request to receive the addressee's attention, a short answer interpretation is not possible, as the presence of the named referent is presupposed. As will be shown in Section 2.2.2, H-% boundary tones tend to signal the incompleteness of an utterance. The contour of (3-a) is an expression of a happy surprise and the one in (3-b) an expression of a disapproval of an act the addressee committed.

In what follows below, it will be demonstrated how intonation contours can occur as independent linguistic constructions with a specific content for the discourse supporting the analyses developed in (Beyssade and Marandin, 2006) and Portes et al. (2014) for different assertion contours. For reasons of space, exclamation contours will not be addressed here, as well as urgent calls, which are the least complex and whose analysis follows readily from what will be sketched below.

2.1 Phonological properties

As observed by Ladd (1978, 518), Hayes and Lahiri (1992, 78), CC in English may involve optional considerable lengthening of the chanted syllables. Hayes and Lahiri (1992, 81–83) point out that even short vowels may be exceptionally lengthened in English, too. A similar situation obtains

with German vowels. There are three different classes of vowels: Tensed, lax and reduced vowels. The latter two sets can under no condition be lengthened in canonical illocutions, such as assertions or questions (cf. Wiese 1996, 19–22, Eisenberg 2013, 89–95). However, in CC even reduced vowels in mono-moraic syllables can be lengthened, which cannot bear word stress under canonical conditions. Moreover, there are barely any contours in German which do not have either a high boundary tone H-% or a low boundary tone L-%. The RCC is one of the few instances which is considered to involve a down-stepped boundary tone !H-% (Grice et al., 2005, 74), other researcher such as (Peters, 2018, 89–90) postulate even an entirely new boundary tone level 0, called *level contours* in order to accommodate this contour.

The emergence of an intonation which is phonologically so distinct from the canonical contours is obvious: the lengthening of short vowels is necessary as vowels are the most audible sounds and the longer they are stretched the more their pronouncement ion is likely to attract attention. Likewise, the outstanding intonation pattern serves as a sort of alert for all the referents in space who are not engaged in a conversation with the caller yet. This would be much more difficult if the caller used standard intonation, which could just be part of a conversation they are leading with somebody else.

2.2 Semantic interpretation

2.2.1 The addressee's attention, benificial information and expectations

Ever since Pike (1945, 71–72), the down-stepped L+H* !H-% was associated with an utterance used to call a speech participant from a distance. The most detailed analysis is the one put forth by Gibbon (1976, 274–287), who shows that this contour does not only apply to names or definite descriptions but also to assertions. Based on experimental data, Condoravdi and Sunwoo (2017, 2018) observe some types of imperatives bearing that contour, in particular well-wishes and mnemonic requests – whereas it is not compatible with orders. This is applicable to German, too, as shown in (5). Furthermore, the L+H* !H-% contour can be applied to German *wh*-questions in highly restricted contexts (6), but it appears to be less felicitous with most polar questions or echo questions (7), but not not with all of them (8)

Finally, this contour is found with some verbless, non-sentential utterances as in *hallo* in (8) or (9) and (10).

(4) (das) Essen ist fertig!
L+H* !H-%
'Food is ready!'

(5) Grüß mir die Oma!
L+H*!H-%
'Send my regards to grandma!'

(6) Wer will noch Vanillekipferl?
L+H*!H-%
'Who wants more vanilla-flavored crescent cookies!?'

(7) *Will wer noch Vanillekipferl?
L+H*!H-%
'Does anybody want more vanilla-flavored crescent cookies!?'

(8) Hallo! Ist da jemand!?
L+H*!H-% L+H* !H-%
'Hello! Is there anybody?'

(9) Vanillekipferl!
L+H*!H-%
'(I have) vanilla-flavored crescent cookies (to share) !?'

(10) Ab ins Bett!
L+H*!H-%
'(go) in your bed'

However, as shown by Ladd (1978, 520–524): the usage of the L+H* !H-% contour is subject to much stricter limitations, as it would not proper to employ it in the contexts of emergencies such as given below:

(11) #Look out for the crevasse!

(12) #Daddy fell downstairs!

(13) #Help!/Fire!/Rape!

Moreover, he refers to contexts in which the addressee is in close proximity. He then concludes that these utterances imply a reference of some sort of routine actions rather than being calls.

All the examples discussed in the literature, including (4)–(6), have three things in common: (i) the speaker is not sure whether they (still) have the attention of the addressee, (ii) the utterance bearing the L+H* !H-% refers to some expectation that was either implicitly or explicitly addressed between the discussing parties and (iii) the encoded message implies some benefit for the addressee.

Aspect (ii) is not so obvious in example (8), which can be uttered in any space, in which the

speaker expects there to be some addressee with whom they have some familiarity. However, it would be out of place to utter this call in a governmental office or some authority. This pattern appears to be most natural in spaces in which the speaker is looking for somebody to solve an issue that was opened in a previous conversation. And it signals, too, that the speaker has no intention to cause any harm or worry to the addressee.

In contrast to claims made in previous research, no routine needs to be involved as the following example illustrated. At the 2006 election of the Austrian parliament the conservative right radical government lost its majority. There had not been any previous edition of that coalition. Nevertheless, it was possible then for somebody in favour of the opposition to pronounce the utterance (14) to some addressee whose attention they were not sure of. Might have been in a beer parlour at close proximity or not.

(14) Schwarz-Blau hat keine Mehrheit mehr!
L+H* !H-%
'The black blue coalition (conservative-right extreme government) no longer holds the majority of its seats anymore!'

This is a clear indicator that this calling contour is to be understood as a request for the addressee's attention, (cf. Pierrehumbert and Hirschberg 1990, 293–294 for a similar observation in English, and less explicitly Liberman 1975, 30) and at the same time the request refers to information, which is sensitive to previously established information in the discourse. This explains also why the L+H* !H-% always comes with some flavour of a routine, which can also be deconstructed as some sort of information that has entered the context of the two speech participants at some earlier point. In similar vein, Pierrehumbert and Hirschberg (1990, 298) argues that H* succeeded by L (in our analysis !H) usually expresses that the information is not new but it can be deduced from the shared beliefs of the interlocutors.

A further reason why examples (11)–(13) sound odd is that the L+H* !H-% contour may present the message bearing the contour as some sort of good news or the fulfilment of a wish or hope. For instance, example (14) would sound very cynical if it were uttered by a supporter of the conservative party or by the fascists.[1]

In contrast, the urgent type of calling contour is not subject to the requirement (ii).

As already mentioned above, the contour L*+H L-H% described in (2-b) will be considered as interrogative request for attention. *Pace* the cautious assumption of Quiroz and Żygis (2017, 1211) , this pattern is not restricted to female speakers at least in Austria, although it is very common in conversation between mothers and their children. The presence of the high boundary tone H-% suggests that it may be some sort of polar question, as this particular tone is only found in interrogative illocutions.

It is often seen as a sign of the speaker that they have not yet finished their turn. In Viennese German the pattern L*+H L-H% is actually ambiguous between a less pronounced and less frequent neutral request for attention: in this use it could be paraphrased as *Friederike, may I have your attention? I want to tell/ask you something*, but it is highly informal and it requires a certain amount of familiarity. The second more frequent interpretation is typically used when a parent catches their children doing something, they had been told not to do. So it is more a reprimand in disguise of a question as *Friederike, can you listen please? Stop doing that.*, henceforth reprimanding question calling contour (RepQCC). Just as with the RCC this contour makes reference to some previously established agreements, but here it is more a previously uttered wish of the speaker that the addressee should not do a certain thing.

2.2.2 Calling contours in dialogical interaction

So far we were able to identify three large ingredients to the semantic interpretation of these various CC, yet, we could not formulate the precise rules of interpretation. In other words, we do not know their semantic status and if they should be considers as belonging to a semantic type or not.

Some preliminary pioneer work in that field was put forward by Truckenbrodt (2012, 2045–2048), who assumes that vocatives introduce a proposition salient from contexts with the content 'I am talking

[1] The reason why the example discussed by Condoravdi and Sunwoo (2017) signals a negative perlocutionary effect may be due to cynical or sarcastic undertones, too.

(i) You can try. But he's not going to hear you. H* !H-L% (=15, Condoravdi and Sunwoo 2017)

Alternatively, it is possible that the calling contours do not convey exactly the same meaning in English and German.

to you'. Inspired by Pierrehumbert and Hirschberg (1990, 293), he then concludes that an H* pitch accent on vocatives means that the new information (the proposition salient from context) should be added to the common ground.

Truckenbrodt's (2012) approach is appealing but it faces some challenges. As pointed out by Ginzburg (2012, 69) and Krifka (2013), response particles such as *yes* or *no* involve a reference to a previously uttered or queried proposition, which will be accessible as question under discussion.

Despite the fact that CCs just come along with simple nouns or calls like *hallo* and occur discourse initially at the same time, they surprisingly are a licit target for *ja* and *nein* responses in German.

It deserves special attention that the usage of these particles with their canonical intonation to react to standard polar questions is not felicitous cf. (15)–(16). Moreover it appears that there are more options for affirmative response particles than for negative response particles.

(15)	#Ja H+L* L-%	(16)	#Nein H+L* L-%
(17)	Ja-a! L+H*!H-%	(18)	Nei-en! L+H*!H-%
(19)	Ja L*H- ˆH%	(20)	#Nein L*H- ˆH%

For the sake of simplicity, we are restricting ourselves here first to CCs with names. If the addressee is not sure either whether they have the attention of the original speaker, they would typically use *ja* with RCC to confirm the readiness to listen or to cooperate and *nein* with RCC to express that they confirm the reception of the message but that they are not ready to fulfill the speaker's intended perlocutionary act. Finally a rising *ja* appears to be the appropriate choice if the former addressee is certain to have the attention of the former speaker.

Despite the fact that rising *ja* is usually transcribed in everyday language as being a polar question *ja?*, there are serious doubts whether it can indeed be a polar question. The meaning of polar questions is to under-specify a proposition with respect to its truth value and to invite the addressee to tell whether the proposition is true or false. Querying a response particle would not result in any plausible interpretation, it would amount to a scenario where the former addressee asks the speaker: *Am I confirming the proposition you classified as question under discussion?*.

However, a high IP boundary tone H-% is not only limited to polar questions but it can occasionally signal incompleteness of the utterance, such as the speaker's readiness and expectation for a reaction (Peters, 2018, 95). In their analysis of boundary tones in English, Pierrehumbert and Hirschberg (1990, 305–307) show that all observed instances of H-% boundary tones make reference to a subsequent utterance, which can be either uttered by the same speaker or the addressee, depending on the utterance type.

In other words, rising *ja* signals the former addressee's receipt of the message and their readiness to listen to more. That's exactly the reason why rising *nein* is not felicitous, as it conveys conflicting information on lexical level it signals a lack of readiness to cooperate where as the intonation invites for more contributions.

Proceeding to RepQCCs, the most plausible answer would be rising *ja*, whereas RCC *ja* would be interpreted as sarcasm signaling reluctance. As for availability requests as in (2-a-ii), rising *ja* is the most plausible response, as they are usually employed in a situation where the speaker is has their attention already on the addressee, which the addressee would know by the time of their response.

At this point we are finally in a position to identify the content of the proposition to which responsive particles are referring, when used with CC on names: RCC and RepQCC query a proposition like *Are you ready to cooperate with respect to the content a certain message?*. This applies even for RCCs with *hallo*, as exemplified in (8). In contrast, ICRCC query a proposition like *Are you <name>?* and ARCC query propositions like *Are you available to join a conversation with me?*

Turning to RCC with more complex utterances, it turns out that the conveyed proposition is the same as with RCC on names. A *ja* response to an well whishing imperative (5) refers to question like: *are you ready to cooperate with respect to the content of the directive?*. Moving on to declaratives, in some cases a *ja* response would be straight forward as in (4), meaning that the addressee acknowledged that the food was ready or they would be preparing to come. However this declarative comes with a very strong perlocutionary effect. With more canonical declaratives such as (14), the use of RCC *ja* is more limited, as there is obviously an interference with the propositional content of the sentence, when it comes to the question what is going to be question under discussion. A RCC *ja*-response would somehow entail that the transmission of the

message would entail the addressee to carry out another action. Likewise, *wh*-questions as in example (6) specify another proposition as QUD, that is why RCC *ja* sounds less felicitous.. But an affirmative response is not totally excluded under conditions, if it is immediately followed by a short answer referring to the focused *wh*-phrase, like *Ja-a, i-ich!*. Note that this option is not at all available for *wh*-question with the canonical H*L-% intonation.

3 Analysis

There is a whole range of phenomena for which a successful analysis should be able to account. One of the most essential aspects is the question how it is possible that each sub-type of calling contours has a couple of semantic contributions which are present with all its realisations across utterance types: as shown above RCC are only felicitous if (i) the addressee has not confirmed their engagement, (ii) there is some unresolved issue between hearer and speaker (iii) if there is some information which is either beneficial to the hearer or to the addressee and (iv) they all convey the same QUD 'are you ready to cooperate with respect to the content of the directive?' (except with cases in which a further QUD is introduced and ranked higher on the QUD list).

The first approach could be to assume that the intonation contour itself contributes meaning to a under-specified illocution. Secondly, one could assume that the interpretation is somehow enriched by some socio-cultural factors from the utterance context by some sort of implicature in the Gricean sense (cf. Grice 1975). As we couldn't find any case, in which the aspects of meaning (i)–(iv) may be canceled, we conclude that they must be conveyed by the intonation it self. Note that intonation manifests itself as perceivable event and comes into consideration as a sign.

There are two more pieces of evidence which are in favour of the assumption that intonation contours can act as phrasal signs in languages of the Indo-European type. In extreme cases, RCC can convey the three aspects of meaning (i)–(iv) described even in cases in which the precise segmental phonetic form does not matter. Imagine a situation in which a group of people is taking a walk in a forest and suddenly they hear from afar somebody uttering the intonation contour (L+)H*!H-%. Even if the phonetic signal is too blurred in order to identify the precise phoneme segments, the group of hikers will immediately know that there is somebody who tries to catch somebody's attention, and that there is some degree of familiarity between the two of them. Sometimes the RCC can be articulated by whistling and its meaning would still be recognizable. Related, but more distinctive intonation contours like the taunting children's chant can alternatively be uttered as a repetitive sequence of meaningless syllables most typically as nə.nə.nə.nə.nə.nə or be applied to declarative sentences or imperatives (cf. Pike 1945, 71–72, Liberman 1975, 32–42).

If intonation contours can contribute meaning to existing utterances, the question arises how the meaning provided by that contour interacts with the meaning of the utterance to which the calling contour is applied? One strategy would be to assume that calling contours may override the phonological and semantic/pragmatic specification of the existing utterance, inspired by the analysis in terms of *composition by amendment* as proposed by (Zeevat, 2019). Alternatively, one could assume that there is a linguistic level in which utterances are under-specified with respect to their intonation on the one hand and under-specified with respect to their illocutionary force. Objects of the type *Locutionary Proposition* (*LocProp*) are a possible canditate to model these abstract linguisitic objects as they involve a description of their semantic and phonological properties (cf. Ginzburg 2012, 172–175). A precise formulation of such an analysis will be left to future research.

There is independent necessity for such an analysis that mediates between the specification of intonation contours and the meaning of utterances, as there are various languages in which the distinction between specific illocutionary acts is merely expressed by differences in the intonation, such as the distinction between Russian declarative clauses vs. polar questions.[2]In similar vein, intonation plays a distinctive role in the formation of polar questions in some lexical tone languages from the Benue subbranch such as Ìgbò (Uchechukwu 2008, Amaechi 2018) and Ẹ̀dó (Ọmọruyi 1988, 1989).

For the sake of space, the analysis focuses on RCC, as they involve the highest degree of complexity, the analysis for the remaining types follows from the former. A successful semantic analysis of CC has to be able to tackle the following puzzles: (i) Which type of illocutionary force do these calls

[2]For this insight I am indebted to Henk Zeevat (pers. commun.)

convey? Are they assertions? Questions? Directives/outcomes? Or an entirely new type on their own? (ii) How can semantic representations of 'empty' intonation contours be modeled? (iii) How can the positive undertone common to all RCC be represented? (iv) How can one model the flavour of routine/familiarity common to RCC? (v) How can one explain the fact that RCC can be applied to a big range of utterance types, conveying different effects in each case? (vi) How can the circumstance be expressed that the speaker is not sure whether the addresssee has their awareness? (vii) What is the actual propositional content of calls that only contain names? To what do the interlocutors say *yes* or *no* if they respond to a call of a name?

The analysis presented here consists of two major components: One the one hand, there are lexicon entries which model CC as phrasal signs that turn existing noun phrases or other utterance types in CCs. Secondly, there are conversational rules which license the use of CCs in conversations and which manage the updates of the dialogue game board, as suggested by Larsson (2002), Traum and Larsson (2003, 328) or Ginzburg (2012).

The format of the lexicon entries can be sketched as follows. Addressing (i), we tentatively drew the cautious conclusions that all the considered CCs are interrogatives, based on the discussion of the data in Section 2.2. As regards the aspects (ii)–(vii), *Type Theory with Records* (TTR) and *Conversation oriented Semantics* (KoS), as developed by Cooper (2005a,b), Cooper and Ginzburg (2015) and Ginzburg (2012) provides the right tools to account for all the desiderata mentioned above: CC can most accurately be analysed adopting Ginzburg's (2012) analysis for non-sentential utterances (NSU) as phrasal signs. The main assumption is that the *phon*-field of the mother node specifies the type of intonation contour and it inherits the precise phonological content (string of phonemes, prosodic structure) from its daughter, which can either be a name or definite description (cf. Figure 3) or a more complex type of utterance, as demonstrated in examples (4)–(14) (cf. Figure 4). It is plausible that the separate entries for daughters that are names and daughters that are more complex utterances could be unified in future research.

The *phon*-field will be considered as list of the type phonological word, inspired by the feature geometry employed by Klein (2000) and Bildhauer

phonological word $=_{def}$

$$\begin{bmatrix} \text{segs} & : & \text{list(phon)} \\ \text{pitch accent} & : & \text{Tone} \\ \text{boundary tone} & : & \text{Tone} \\ \text{utterance} & : & \text{Alignment} \\ \text{intonation phrase} & : & \text{Alignment} \\ \text{phonological phrase} & : & \text{Alignment} \end{bmatrix}$$

Figure 1: Type definition: phonological word

(2007, 161–186), as illustrated in Figure 1. The most relevant fields for the purpose here are pitch accent and boundary tone, which determine the intonation contour of the phrasal sign, the remaining fields will be inherited from the daughter's *phon*-field.

The fact that RCCs convey propositional content is reflected by the *cont*-field, which states the query whether the addressee is ready to cooperate with respect to a given message m. In cases where the daughter contains sentential utterances, m corresponds to the content of the proposition they introduce, in cases, where the daughter is a name, the content m is salient from the context. The complex type *m-ben* states that m is of benefit to the addressee and *m-exp* says that the addressee and/or the speaker expected m to happen at some point of time. And finally, the circumstance that the speaker is not sure whether they have the attention of the addressee is reflected by the fact that the *moves*-list is empty, which indicates that the interlocutors have not engaged into any conversation yet.

Note that A RCC does not have to be the first utterance of the speaker's turn but it has to be the speaker's first turn before it is even clear whether the addressee engages into the speaker's invitation or not. This can be seen from the authentic corpus example uttered by Farin Urlaub, a punk rock singer born in Berlin in 1963, during a reading of the cartoon *Didi & Stulle* in June 2020. In the scene under discussion, Angela Merkel just won the governmental elections and happens to see some random guy she beliefs to be David Bowie and tries to hit on him:

(21) Ey Bowie!! Lange nich jesehn und doch
hey Bowie long NEG see.PPP and yet
wiedaerkannt, wa? Rate ma wer
recognize.PPP QTAG guess.IMP PAR who
gerade die [$_{L+H*!H-\%}$ Wahl] jewonn
just the elections win.PPP

hat!![3]
has
'Hey Bowie! Long time no see. Guess who just won the elections!! '

As can be seen from example (21), the RCC is not restricted to dialogue initial utterances as long as they occur in the first turn. Assuming that that the addressee only starts to ground the utterances once they commit themselves to the conversation, the restriction of RCCs to occur in the first turn is modelled by an empty $Moves$ list. Following Ginzburg (2012, 175–189), we suggest that utterances which are not grounded yet are stored in the pending list. Note that it is not really felicitous to respond with *ja* to the imperative (21), unlike it is the case with the imperative (5), as the speaker expect the addressee to fulfil the directive immediately by giving him a response. Interestingly the RCC is incompatible with both calls like (*h*)*ey* and names which follow (*h*)*ey*.

The corresponding conversational rule for RCCs is outlined in Figure 2. Just like with greetings, RCCs occur discourse initially, as a consequence their conversational rule ressembles much to the format of greetings as suggested by Ginzburg (2012, 76).

The fact that the addressee has not committed to the dialogue yet is expressed by the empty $moves$-list, the fact that there is an open issue between speaker and addressee originating in an earlier conversation can be accommodated in the $issue$-list, which is meant to keep record of unsolved global QUD, as suggested by Cooper et al. (2000) and Larsson (2002, 163–164).

One of the reviewers wondered, whether the QUD which is introduced with calling contours could have entered the dialogue game board by accommodation. However, there is a crucial difference between the accommodation of unraised questions and the type of QUD introduced along with calling contours. Following Cooper et al. (2000) and Larsson (2002, 153–164), the scenario for accommodation of questions is characterised by a dialogue participant A who asks some question q_1 and and dialogue participant B who responds with a relevant answer p_1 and at the same time with a second answer p_2 for which A has not explicitly raised a question. In such a scenario, the question q_2, which was not explicitly asked by A, has to be accommodated by A.

The scenario with calling contours is very different. Because with what we are dealing here is not a situation in which a unraised question has to be reconstructed on the basis of an answer. Assume B is requestion the attention of A, a call for attention uttered by B is not a possible answer to a question A somehow planned to raise but they didn't raise explicitly. Thus it appears to be more accurate to assume that with attention requests, the QUD is introduced in virtue of the relevant conversational rule.

RepQCC can be analysed in a parallel manner, cf. Figure 5. As they only apply to names, the option with more complex utterances is left out. The main difference here is that the DGB contains an outcome o and the speaker referent does not want the addressee to actualize o. Moreover, this prohibition is an already established fact known to both the speaker and the addressee and hence part of the common *facts*-list.

The representation of ARCC are almost identical to the one of RepQCC, the main difference is that they lack the reference to a previously established prohibition on the addressee and the content value of the mother would be slightly different as in *?available(dgb.addr,dgb.spkr,dgb.utt-time*. Due to lack of space the precise phrasal structure will be ommitted here.

ICRCC differ in essential aspects, as it is seen in Figure 6. Unlike with the previously discussed CCs, the referent of the name bearer defined in the *dgb-params*-field of the daughter has not yet directly been identified with the addressee referent within the DGB. It presupposes the existence of a referent for the addressee and queries whether the addressee is identical to the name bearer, which is done in the *cont*-field of the mother node.

4 Summary and Outlook

In this paper, it was shown that contours stripped off of any lexical content may convey propositional meaning in German and propably many Indo-European languages too. CCs are be most accurately analysed as phrasal signs that select a proper name or in some instances a more complex phrases as their daughter. Moreover, they are combining with locutive propositions which are under-specified for their illocutionary force and for their intonation.

[3] youtube: *DIE ÄRZTE lesen für Berliner Liveclubs aus FILs „Didi & Stulle"*, https://youtu.be/bbCGqFaNo-Q, 32:40–50s; last access June 25 2020.

$$
\begin{array}{l}
\mathit{AttentionRequest} =_{def} \\
\left[\begin{array}{lll}
\text{pre} & : & \left[\begin{array}{lll} \text{moves=}\langle\rangle & : & \text{IllocProp} \\ \text{qud=}\{\ \} & : & \text{poset(Question)} \\ \text{issues} & : & \text{poset(Question)} \end{array}\right] \\
\text{issues} & : & \text{poset(Question)} \\
\text{effects} & : & \left[\begin{array}{lll} \text{LatestMove=?ready-to-cooperate(addr,m)} & : & \text{IllocProp} \\ \text{qud=}\{\ \} & : & \text{poset(Question)} \end{array}\right]
\end{array}\right]
\end{array}
$$

Figure 2: Conversational rule for L+H* !H-% ‘routine’ calls

$$
\begin{array}{l}
\textit{rout-call-name-ph} =_{def} \\
\left[\begin{array}{lll}
\text{phon=}\langle \left[\begin{array}{lll} \text{segs} & : & \text{list(phon)} \\ \text{PA=}\textit{low-high-star} & : & \text{tone} \\ \text{BT=}\textit{downstep} & : & \text{tone} \end{array}\right] \rangle & & \\
\wedge_{\textit{merge}}\ \text{LIST1(PHONWORD)} & : & \text{list(Phonword)} \\
\text{cat} & : & \left[\begin{array}{lll} \text{head=V[+fin]} & : & \text{PoS} \end{array}\right] \\
\text{dgb-params} & : & \left[\begin{array}{lll} \text{spkr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{m} & : & \text{IllocProp} \\ \text{m-exp} & : & \text{expect(addr,m.prop)} \\ \text{m-ben} & : & \text{benefit(addr,m.prop)} \\ \text{Moves=}\langle\rangle & : & \text{IllocProp} \\ \text{QUD=}\{\} & : & \text{Poset(Question)} \end{array}\right] \\
\text{cont=}\textit{?ready-to-cooperate(dgb.addr,dgb.m)} & : & \text{Question}
\end{array}\right] \\
\qquad\qquad\qquad\qquad\qquad\qquad \Big| \\
\textbf{dtr}\left[\begin{array}{lll}
\text{LIST1(PHONWORD)} & : & \text{list(RecType)} \\
\text{phon} & : & \text{LIST1(PHONWORD)} \\
\text{cat} & : & \left[\begin{array}{lll} \text{head=N} & : & \text{PoS} \end{array}\right] \\
\text{dgb-params} & : & \left[\begin{array}{lll} \text{y=cont.x=dgb.addr} & : & \text{Ind} \\ \text{rest}^{\textit{facts}} & : & \text{Named(dgb.addr,‘Name’)} \end{array}\right] \\
\text{cont} & : & \left[\begin{array}{lll} \text{x} & : & \text{Ind} \end{array}\right]
\end{array}\right]
\end{array}
$$

Figure 3: Lexicon entry of ‘routine’ calling contours with names

$$
\begin{array}{l}
\textit{rout-call-sent-ph} =_{def} \\
\left[\begin{array}{lll}
\text{phon=}\langle \left[\begin{array}{lll} \text{segs} & : & \text{list(phon)} \\ \text{PA=}\textit{low-high-star} & : & \text{tone} \\ \text{BT=}\textit{downstep} & : & \text{tone} \end{array}\right] \rangle & & \\
\wedge_{\textit{merge}}\ \text{LIST1(PHONWORD)} & : & \text{list(Phonword)} \\
\text{cat} & : & \left[\begin{array}{lll} \text{head=V[+fin]} & : & \text{PoS} \end{array}\right] \\
\text{dgb-params} & : & \left[\begin{array}{lll} \text{spkr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{m=dtr.cont} & : & \text{IllocProp} \\ \text{m-exp} & : & \text{expect(addr,m.prop)} \\ \text{m-ben} & : & \text{benefit(addr,m.prop)} \\ \text{Moves=}\langle\rangle & : & \text{IllocProp} \\ \text{QUD=}\{\} & : & \text{Poset(Question)} \end{array}\right] \\
\text{cont=}\textit{?ready-to-cooperate(dgb.addr,dgb.m)} & : & \text{Question}
\end{array}\right] \\
\qquad\qquad\qquad\qquad\qquad\qquad \Big| \\
\textbf{dtr}\left[\begin{array}{lll}
\text{LIST1(PHONWORD)} & : & \text{list(RecType)} \\
\text{phon} & : & \text{LIST1(PHONWORD)} \\
\text{cat} & : & \left[\begin{array}{lll} \text{head=V[fin+]} & : & \text{PoS} \end{array}\right] \\
\text{dgb-params} & : & \left[\begin{array}{lll} \text{y=cont.x=dgb.addr} & : & \text{Ind} \\ \text{rest}^{\textit{facts}} & : & \text{Named(dgb.addr,‘Name’)} \end{array}\right] \\
\text{cont} & : & \text{Prop} \vee \text{Wh-Question} \vee \text{Outcome}
\end{array}\right]
\end{array}
$$

Figure 4: Lexicon entry of ‘routine’ calling contours with more complex utterances

reprim-interr-call $=_{def}$

$$\begin{bmatrix}
\text{phon}=\langle \begin{bmatrix} \text{segs} & : & \text{list(phon)} \\ \text{PA}=\textit{low-star-high} & : & \text{tone} \\ \text{BT}=\textit{low-high} & : & \text{tone} \end{bmatrix} \rangle & & \\
\wedge_{merge}\ \text{LIST1(PHONWORD)} & : & \text{list(Phonword)} \\
\text{cat} & : & \begin{bmatrix} \text{head=V[+fin]} & : & \text{PoS} \end{bmatrix} \\
\text{dgb-params} & : & \begin{bmatrix} \text{spkr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{addr=cont.x} & : & \text{Ind} \\ \text{rest}^{facts} & : & \text{Named(addr,`Name')} \\ \text{o} & : & \text{Outcome} \\ \text{f=forbid(spkr,addr,o)} & : & \text{Prop} \\ \text{Moves=}\langle\rangle & : & \text{IllocProp} \\ \text{QUD=}\{\} & : & \text{Poset(Question)} \\ \text{Facts} & : & \text{Set(Prop)}\cup f \end{bmatrix} \\
\text{cont=}\textit{?ready-to-cooperate(dgb.addr,dgb.spkr)} & : & \text{Question}
\end{bmatrix}$$

$$\mathbf{dtr}\begin{bmatrix}
\text{LIST1(PHONWORD)} & : & \text{list(RecType)} \\
\text{phon} & : & \text{LIST1(PHONWORD)} \\
\text{cat} & : & \begin{bmatrix} \text{head=N} & : & \text{PoS} \end{bmatrix} \\
\text{dgb-params} & : & \begin{bmatrix} \text{y=cont.x=dgb.addr} & : & \text{Ind} \\ \text{rest}^{facts} & : & \text{Named(dgb.addr,`Name')} \end{bmatrix} \\
\text{cont} & : & \begin{bmatrix} \text{x} & : & \text{Ind} \end{bmatrix}
\end{bmatrix}$$

Figure 5: Lexicon entry of reprimanding question calling contours with names

identity-confirmation $=_{def}$

$$\begin{bmatrix}
\text{phon}=\langle \begin{bmatrix} \text{segs} & : & \text{list(phon)} \\ \text{PA}=\textit{high-low-star} & : & \text{tone} \\ \text{BT}=\textit{high} & : & \text{tone} \end{bmatrix} \rangle & & \\
\wedge_{merge}\ \text{LIST1(PHONWORD)} & : & \text{list(Phonword)} \\
\text{cat} & : & \begin{bmatrix} \text{head=V[+fin]} & : & \text{PoS} \end{bmatrix} \\
\text{dgb-params} & : & \begin{bmatrix} \text{spkr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{Moves=}\langle\rangle & : & \text{IllocProp} \\ \text{QUD=}\{\} & : & \text{Poset(Question)} \\ \text{Facts} & : & \text{Set(Prop)}\cup f \end{bmatrix} \\
\text{cont=}\textit{?identical(dgb.addr,dtr.cont.x)} & : & \text{Question}
\end{bmatrix}$$

$$\mathbf{dtr}\begin{bmatrix}
\text{LIST1(PHONWORD)} & : & \text{list(RecType)} \\
\text{phon} & : & \text{LIST1(PHONWORD)} \\
\text{cat} & : & \begin{bmatrix} \text{head=N} & : & \text{PoS} \end{bmatrix} \\
\text{dgb-params} & : & \begin{bmatrix} \text{y=cont.x} & : & \text{Ind} \\ \text{rest}^{facts} & : & \text{Named(y,`Name')} \end{bmatrix} \\
\text{cont} & : & \begin{bmatrix} \text{x} & : & \text{Ind} \end{bmatrix}
\end{bmatrix}$$

Figure 6: Lexicon entry of the of the H+L* H-% identity confirmation request for names

Ever since Pike (1945, 71–72), Liberman (1975, 32–42), Fónagy et al. (1983, 155–157) and Di Cristo (1999, 215–216) related phenomena have been observed, in which the contour alone appears to express a rather rich degree of semantic information, such as the popular Children's Chant to taunt others in a playful way. Or level contours in German, which consist of two tonal levels only flattening all curves and which signal a lack of interessent in the performed utterance, as discussed by Wöllstein (2016, 123–125), Peters (2018, 98–99). But these will be left to future studies.

References

Mary Chimaobi Amaechi. 2018. Yes-no question formation in igbo: the phono-syntax interface. *Journal of West African Languages*, 45(1):78–87.

Amalia Arvanati, Marzena Żygis, and Marek Jaskuła. 2016. The phonetics and phonology of the polish calling melodies. *Phonetica*, 73:338–361.

Claire Beyssade and Jean-Marie Marandin. 2006. French intonation and attitude attribution. In *Proceedings of the 2004 Texas Linguistics Society Conference: Issues at the Semantics-Pragmatics Interface*, Somerville, Mass. Cascadilla Proceedings Project.

Felix Bildhauer. 2007. *Representing Information Structure in an HPSG Grammar of Spanish*. Ph.D. thesis, Universität Bremen, Bremen.

Joan Borràs-Comes, Rafèu Sichel-Bazin, and Pilar Prieto. 2015. Vocative intonation preferences are sensitive to politeness factors. *Language and Speech*, 58(1):68–83.

Cleo Condoravdi and Jeong Sunwoo. 2017. Imperatives with the calling contour. In *Berkeles Linguistics Society (BLS) 43*, volume 1, pages 185–210, Somerville, Mass:. Cascadilla Proceedings Press.

Cleo Condoravdi and Jeong Sunwoo. 2018. Imperatives and intonation: the case of the down-stepped level terminal contour. In *West Coast Conference on Formal Linguistics (WCCFL) 35*, pages 214–223, Somerville, Mass. Cascadilla Proceedings Press.

Robin Cooper. 2005a. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3:333–362.

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

Robin Cooper, Elisabet Engdahl, Staffan Larsson, and Stina Ericsson. 2000. Accomodation questions and the nature of qud. In *Proceedings of Götalog 2000*, volume 5, pages 57–61, Gothenburg.

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*, pages 375–407. John Wiley, Malden, MA.

Albert Di Cristo. 1999. French. In Daniel Hirst and Albert Di Cristo, editors, *Intonation Systems. A survey of 20 languages*, pages 195–218. Cambridge University Press, Cambridge.

Thomas O. Ọmọruyi. 1988. On the formation of questions in Ẹdo. *Journal of African Languages and Linguistics*, 10:19–31.

Thomas O. Ọmọruyi. 1989. Focus and question formation in Ẹdo. *Studies in African Linguistics*, 20(3):281–300.

Peter Eisenberg. 2013. *Grundriß der deutschen Grammatik: Das Wort*, 4 edition, volume 1. Metzler, Stuttgart. GB 1602 E36??

Zsuszanna Fagyal. 1997. Chanting intonation in french. *University of Pennsylvania Working Papers in Linguistics*, 4(2).

Ivan. Fónagy, Eva. Bérard, and Judith Fónagy. 1983. Les clichés élodiques du français parisien. *Folia Linguistica*, 17:153–185.

Dafydd Gibbon. 1976. *Perspectives of intonation analysis*. Lang, Bern.

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

Martine Grice, Stefan Baumann, and Ralf Benzmüller. 2005. German intonation in autosegmental-metrical phonology. In Sun-Ah Jun, editor, *Prosodic Typology: the Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, Oxford.

Paul H. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, 3: Speech acts*. Academic Press.

Bruce Hayes and Aditi Lahiri. 1992. Durationally specified intonation in english and bengali. In *Proceedings of the 1990 Wenner Gren Center Conference on Music, Language, Speech and the Brain*, pages 78–91, Houndsmill. Macmillian.

Ewan Klein. 2000. Prosodic constituency in hpsg. In Ronnie Cann, Claire Grover, and Philip Miller, editors, *Grammatical Interfaces in HPSG*, volume 8 of *Studies in Constraint-Based Lexicalism*, pages 171–203. CSLI Publications, Stanford.

Manfred Krifka. 2013. Response particles as propositional anaphors. In *Proceedings of SALT 23*, pages 1–18, Santa Cruz, CA.

Robert D. Ladd. 1978. Stylized intonation. *Language*, 54(3):517–540.

Robert D. Ladd. 2008. *Intonational Phonology*. Cambridge University Press.

Peter Ladefoged and Keith Johnson. 2011. *A Course in Phonetics*, 6 edition. Wadsworth Cengage Learning, Boston, MA.

Staffan Larsson. 2002. *Issue-based dialogue management*. Ph.D. thesis, Göteborg University, Göteborg.

Mark Yoffe Liberman. 1975. *The intonational system of English*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Oliver Niebuhr. 2013. Resistance is futile – the intonation between continuation rise and calling contour in german. In *Proceedings on the 14th Interspeech Conference, Lyon*, pages 225–229. ISCA.

Jörg Peters. 2018. Phonological and semantic aspects of german intonation. *Linguistik online*, 88(1/18):85–107.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communications*, pages 271–311. MIT Press, Cambridge.

Kenneth L. Pike. 1945. *The intonation of American English*. Ann Arbor, Don Mills.

Cristel Portes, Claire Beyssade, Amandine Michelas, Jean-Marie Marandin, and Maud Champagne-Lavau. 2014. The dialogical dimension of intonational meaning: Evidence from french. *Journal of Pragmatics*, 74:15–29.

Sergio I. Quiroz and Marzena Żygis. 2017. The vocative chant and beyond: German calling melodies under routine and urgent contexts. In *Proc. Interspeech 2017*, pages 1208–1212.

David R. Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and new directions in discourse and dialogue*, chapter 15, pages 325–353. Kluwer Academic Publishers, Dordrecht.

Hubert Truckenbrodt. 2012. Semantics of intonation. In *Semantics. An International Handbook of Natural Language Meaning*, volume 3, pages 2039–2969. Walter de Gruyter, Berlin.

Chinedu Uchechukwu. 2008. The modal system of the igbo language. In Werner Abraham and Elisabeth Leiss, editors, *Modality-Aspect Interfaces. Implications and Typological Solutions*, volume 79 of *Typological Studies in Language*, pages 241–276. Benjamins, Amsterdam.

László Varga. 2008. The calling contour in hungarian and english. *Phonology*, 25:469–497.

Richard Wiese. 1996. *The phonology of German*. Phonology of the World's languages. Oxford University Press.

Angelika Wöllstein, editor. 2016. *Duden, die Grammatik: unentbehrlich für richtiges Deutsch*, 9 edition, volume 4 of *Der Duden*. Dudenverlag, Mannheim.

Henk Zeevat. 2019. Superlative quantification particles. In *Questions in Discourse*, volume 1, pages 286–306, Leiden. Koninklijke Brill.

Cong Zhang. 2018. Chanted call tune in tianjin mandarin: Disyllabic calls. In *th Interational Conference on Speech Prosody 2018*.

Comparing models of speakers' and listeners' head nods

Tom Gurion, Patrick G.T. Healey and Julian Hough
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
{t.gurion,p.healey,j.hough}@qmul.ac.uk

Abstract

Head movement, and head nods in particular, are important communicative signals in face to face conversations. Listeners' head nods can be characterised as acknowledgements to the speaker or as showing support when the speaker encounters problems in completing their turn. Speakers' head movement, on the other hand, is often explained as mimicry. These different explanations of the same observed phenomena are rarely compared, and often contradict each other. This study compares models of head nods in free conversation based on three theories: mimicry, backchannels, and listener responses to speaker trouble. Pairs of participants freely conversed for 15 minutes while fitted with a motion capture system. Head nods where automatically extracted from the data, and compared to predictions by the models. The results highlight the differences between head nod patterns of speakers and listeners, and suggest that the theories are not mutually exclusive. Comparing the theories yield inconclusive results and fail to determine which one best explains head nods. Lastly, the vast majority of the observed head nods are not accounted by any of the investigated models, implying that mimicry, backchannels, and responses to speech disfluencies alone cannot explain head nods exhaustively.

Face-to-face conversations are multimodal in nature. Verbal and non-verbal communication both contribute to the content of the conversation and coordinate its structure. Research of non-verbal communication shows a number of coordinating phenomena: listeners gesture to demonstrate their attendance to the speaker (Goffman, 1955); speakers respond to listeners gaze, and often change who they are looking at depending on eye-contact with their addressees (Goodwin, 1979); when listeners fail to provide the expected listening behaviours speakers get confused (Bavelas et al., 2000).

This study aims to compare three theories that explain head nods in conversation: nods as backchannels, nods as mimicry, and nods as listener responses to speaker trouble (specifically, disfluency). There are many studies about backchannels and mimicry, but to our knowledge, comparisons of these as competing explanations for the same observed phenomena are rare. Whereas two of these theories (backchannels and responses to speaker trouble) are specific to listeners, mimicry theory is agnostic to conversational roles. In other words, the mimicry literature in general ignores speaker-hearer roles. Whereas some studies investigate how speakers mimic listeners (Bailenson and Yee, 2005; Vrijsen et al., 2010) others look at listener mimicry of speakers (Maatman et al., 2005; Leander et al., 2012; Hale et al., 2018), while both apply the same reasoning and theory.

In this study we aim to test how well these theories explain observed head nods in free conversations in respect to speaker-hearer roles. We also address the question of overlap between the theories, and whether or not they provide an adequate account for head nods.

We conducted a study with pairs of participants having free dialogues, while fitted with a motion caption system. Their head movement was analysed for head nods and these were compared to the predicted behaviour according to models based on the three theories above.

The rest of the paper is as follows: Section 1 overviews the theoretical and empirical basis of the models of head nods in dialogue we implement; Section 2 describes our interactive head nod prediction models; Section 3 describes the methods we use for data collection and the automatic techniques we use for testing our model; Section 4 gives and discusses the results of applying our models to our data, before we conclude in Section 5.

1 Related work

1.1 Backchannels

Natural and engaging conversations rely on the ability of the interlocutors to achieve common ground. This process often involves feedback from listeners for their understanding, or misunderstanding. Timed listener responses, known as backchannels (Yngve, 1970), can be non-verbal (e.g. head nods), or para lingual (e.g. utterances like "uh-huh"). They are crucial for the speaker to assess the listener engagement and adapt to it (Bavelas et al., 2000), and can mediate turn-taking in conversation (Duncan et al., 1979).

Surface features like speaker-listener eye contact and speech prosody are often enough to model backchannels. Ward and Tsukahara (2000) suggeste a backchannel prediction model that is based on the speaker's prosody alone. Their model, summarised in Table 1, provides a set of simple hand-crafted rules that use the speaker "vocalisation state" (speaking or silent) and pitch, to predict listeners' backchannels. Later studies often rely on this model, either as a component for conversational systems (Maatman et al., 2005), or as a reference point for an alternative, sometimes more complex, backchannel predictor (Morency et al., 2008; Poppe et al., 2010).

1.2 Mimicry

A well known hypothesis is that conversational partners often mimic each other's postures, gestures, use of language, facial expression, and more. It is commonly claimed that mimicry in conversation has persuasive power and can increase likeability, empathy, and feeling of closeness (for a review see Chartrand and Lakin, 2013).

The automatic tendency to imitate others in social interactions is often described as the "chameleon effect" (Chartrand and Bargh, 1999). It is usually explained by the perception-behaviour link: a cognitive mechanism that link together perception and action and suggests that we have to act in order to perceive. This implies that mimicry is an unconscious automatic behaviour that is a by product of our ability to process social encounters.

The temporal properties of mimicry have been studied extensively. Early research in mimicry with virtual reality reported that head movement mirroring by a virtual agent, with a 4 second lag, increases rapport (Bailenson and Yee, 2005). A recent study (Hale et al., 2018) challenged this 4 second mimicry lag from Bailenson's work (Bailenson et al., 2004; Bailenson and Yee, 2005). The authors highlight that a 4 second lag implies that mimicry is not a reactive process and it relies on memory to operate, which is rather unlikely. One alternative model describes mimicry as a predictive process, similar to the mechanism that allows musicians to coordinate their playing, and implies no time lag. The third alternative describes mimicry as an immediate reactive response to other people, thus implies a time lag on the order of 300-1000 milliseconds. They used motion capture to measure the head pitch in dyadic conversations and found that listeners' low-frequency head movement follows the speaker's head movement after a 600 milliseconds delay.

1.3 Responses to speaker's troubles and disfluencies

The interactive nature of face-to-face conversations suggests that all of the participants in a conversation involve in the effort to keep it going. Therefore, listeners responses are especially important when the speaker encounters problems in producing a turn, as happen commonly in free conversations (Schegloff, 1992).

Disfluencies are often described as a 3-part structure of reparandom, interregnum, and repair (the terms were proposed by Shriberg, 1994). This structure is presented in the example bellow.

$$\text{John} \quad \underbrace{\text{likes}}_{\text{reparandum}} \quad \underbrace{\text{uh}}_{\text{interregnum}} \quad \underbrace{\text{loves}}_{\text{repair}} \quad \text{Mary}$$

A study exploring head movement after disfluent utterances found that addressees tend to nod more between 1 to 3 second after a disfluency (Healey et al., 2013). In addition, listeners gesture more when the speaker is disfluent, and their gestures tend to become similar to those of the speaker (Healey et al., 2015).

2 Models implementation

2.1 Backchannels

Ward and Tsukahara (2000) suggest a prosody-based backchannels prediction model. This model, summarised in Table 1, provides a set of simple hand-crafted rules that use the pitch of the speaker's voice, and the timing they started or stopped speaking, to predict hearers' backchannels. It is important to highlight that the original model was designed to predict verbal backchannels. Neverthe-

Provide a backchannel feedback upon detection of:
P1 a region of pitch less than the 26th-percentile pitch level and
P2 continuing for at least 110ms,
P3 coming after at least 700ms of speech,
P4 provided that no backchannel has been output within the preceding 800ms,
P5 after 700ms wait.

Table 1: Ward and Tsukahara (2000) backchannels prediction algorithm (copied from Poppe et al., 2010).

less, several studies have used it to predict non-verbal backchannels, and specifically head nods (e.g. Morency et al., 2008; Poppe et al., 2010), based on the assumption that verbal and non-verbal backchannels are generally interchangable (Bavelas et al., 2000).

This model was chosen for its simplicity and ease of interpretation. While machine learning based models often out-perform it (e.g. Morency et al., 2008), they are usually harder to interpret. The interpretability of the model is especially important when comparing competing models that represent theoretically different non-verbal behaviours.

The model used here has a slight deviation from the one in the original paper. The original model waits for a region of pitch with less than the 26th-percentile pitch level (rule P1 in table 1). This assumes access to the pitch data of the entire conversation, and therefore implies a non incremental operation. To adapt the algorithm to incremental processing appropriate for real-time applications a rotating 10 seconds buffer was introduced. This buffer always keeps the pitch profile of the last 10 seconds. Percentile calculation is done against this rotating buffer.

The model is implemented in the programming language PureData (Puckette, 1996)[1] and utilises the voice activation detection application py-webrtcvad[2] to decide if there is speech in the audio data or not (rule P3 in table 1). One instance of the model is run for each participant. It analyses the audio stream from the participant's microphone and predicts backchannels for their conversational partner.

[1]The backchannels detection server is available online at https://github.com/Nagasaki45/backchannels.

[2]https://github.com/wiseman/py-webrtcvad

2.2 Mimicry

To tackle the question of mimicry lag two mimicry models were implemented. They predict head nods for one participant 600 milliseconds (following Hale et al., 2018) and 4 seconds (following Bailenson and Yee, 2005) after a head nod by their partner. These models rely on a head nod detector that is described in detail in Section 3.2.

Note that the mimicry models use head nod mimicry, and not head movement mimicry as most mimicry studies suggested. There are two reasons for this decision: firstly, we aim to compare mimicry and backchannels as possible explanations for communicative head movement. Therefore, to obtain a meaningful comparison, both models should operate on the same physical movement. Second, the mimicry literature implies that mimicry is a general phenomenon that should operate on multiple levels. If the perception-behaviour link is a fundamental cognitive mechanism, as mimicry research suggests (Chartrand and Bargh, 1999), it should also operate on head nods. Furthermore, even when head movement mimicry is investigated, scholars often conclude that head nods are mimicked in conversation. For example, Hale et al. (2018) concluded that "the cognitive mechanisms generating mimicry of head nods act with a constant lag of around 0.588 msec".

2.3 Disfluency

Speech disfluencies are detected using IBM Watson's speech-to-text service, and the "deep_disfluency" software, a state-of-the-art incremental model that tags disfluencies in transcripts (Hough and Schlangen, 2017). The audio from each participant's microphone is sent to IBM Watson. The returned transcription is passed through the deep disfluency tagger. Interregnum and edit term tags (<e/>) and repair start tags (<rpS/>) are interpreted as disfluencies.

3 Methods

3.1 Participants

Thirteen pairs of native English speakers that knew each other in advance (14 women, 12 men), age 18-26 (mean: 20.8, std: 1.9) participated in the study. Most of the participants were undergraduate and master students in STEM, who were recruited through university mailing lists. Each of them received £10 compensation for their participation.

3.2 Apparatus

Two participants at a time participated in the study. They were seated in the same room, two metres apart, facing each other. Each participant was recorded by a dash microphone and a video camera.

To track their head movement the participants wear baseball caps fitted with HTC Vive trackers[3] on their visors. They also held HTC Vive hand-held controllers[4] that tracked their hands movement. Usually, the HTC Vive controllers and trackers work alongside the HTC Vive headset. Because the headsets were not necessary for this study they were set aside at a table near the participants and their movement data wasn't recorded. The logged movement data includes the head and hands position (Cartesian X, Y, and Z values), and rotation (Euler angles X, Y, and Z), sampled at the frame rate of the motion data capture application[5], which varies between 60Hz and 90Hz.

Head nods are detected based on the vertical position of the head tracker. To guarantee a constant rate, the vertical head position is up-sampled, linearly, to 100Hz. These samples are then filtered with two second order Butterworth filters: a low-pass at 4Hz followed by an high-pass at 1Hz. The model predicts a head nod if the result is smaller than -4 millimetres. The system won't report another head nod until the movement stabilise (a sample between -2 and 2 millimetres). This technique is based on Healey et al. (2013): "low frequency movements (1Hz and below) and high frequency movements (4Hz and above) were eliminated ... head nods were identified as vertical movements at a speed >0.3 mm/frame". The above study processed data at 60 frames per second, that is 0.18 millimetre per millisecond, or, translated to our 100hz sample rate, 1.8 millimetres per sample. This value, of 1.8 millimetres per sample, seemed too sensitive. We fine tuned it manually until the system was sensitive enough to detect most head nods but not too sensitive to be triggered by other movement.

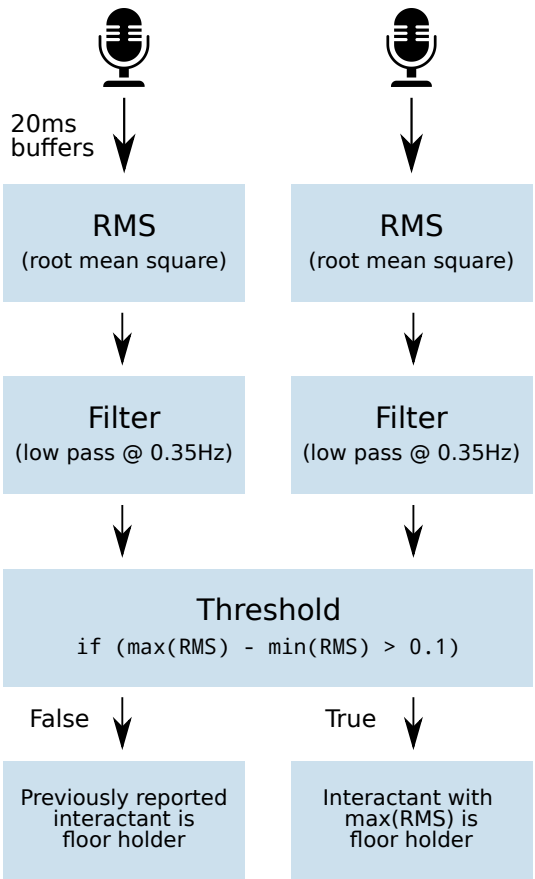


Figure 1: Schematic diagram of the floor control detection model.

Recording the participants' voices in one room introduced a significant audio bleed. That is, the microphone of one participant recorded a significant portion of their partner's voice. Therefore, and because of the models' dependency on audio, the iZotope RX 6 De-bleed[6] software was used to reduce the bleed as much as possible. The reduction strength was set to maximum and the artefact smoothing to 0. No other settings were tested. The reduction in bleed was not tested exhaustively, but a few measurements suggest that the bleed was reduced by up to 40dB. This version of the audio files was used instead of the original audio files for the floor control detection model (described next), and the backchannels and disfluency models.

A simple and accurate floor control detection

[3] https://www.vive.com/us/vive-tracker/

[4] https://www.vive.com/us/accessory/controller/

[5] The application for running this study is available online at https://github.com/Nagasaki45/F2F-study/tree/study-4.

[6] https://www.izotope.com/en/products/repair-and-edit/rx/features-and-comparison/de-bleed.html

model (Gurion et al., Unpublished) processed the audio from the participants' microphones to determine the floor holder at any given moment. This model, presented in Figure 1, relies on simple audio processing techniques of filters and thresholds to operate. Audio from the microphones of the two interactants is processed in buffers of 20 milliseconds. For each buffer the root mean square (RMS) value is calculated. These values are filtered by low pass filters with a cutoff frequency of $0.35Hz$. If the difference between the minimal and the maximal filtered RMS values is larger than 0.1 the interactant with the maximal filtered RMS value is the floor-holder. Otherwise the previously reported floor-holder is the floor-holder again.

Lastly, the backchannels, mimicry, and disfluency models described earlier logged predictions for partner nods.

3.3 Procedure

First, the participants filled a demographic questionnaire, followed by a questionnaire about their social relationship. Then they were introduced to the Dream Apartment design task (described in detail in Hough et al., 2016), in which they are asked to discuss the design of an apartment for them to share. The participants were then fitted with the motion capture system, discussed the task for 15 minutes, until the experimenter asked them to stop.

4 Results and discussion

4.1 Head nods frequency

First, head nods frequency was calculated separately for floor holders and non floor holders. For each participant, the number of head nods while holding the floor was divided by the total time they held the floor to find the nodding frequency. Nodding frequency for non floor holders was calculated in a similar fashion. Comparing these values suggest that floor holders nod more frequently (M=0.27Hz, SD=0.14) than non floor holders (M: 0.18Hz, SD=0.09; $t(25) = 4.11$, $p < 0.001$).

4.2 Overlap between models

Here we address the question of whether the models actually differ from each other. Whereas they are driven by different theories, there is no reason to assume that the models produce different predictions.

To test that we calculated the overlap for every combination of 2, 3, or all 4 models. A window of 400 milliseconds is defined around each prediction. The overlap for a combination of models is defined as the percentage of time for which the windows intersect, plus the percentage of time for which no window is reported (i.e. an XNOR logical operation on the windows). Let us consider an example with two models and a total duration of one second. If model A predicts a head nod at 0.2 seconds and model B predicts a head nod at 0.4 seconds the 400 milliseconds windows intersect on the interval 0.2-0.4 seconds and no window is reported for the interval 0.6-1 seconds. Therefore, they agreed on 60% of the time, and this is their overlap.

For each combination of models a chance overlap is also calculated. This is the overlap that is expected from models that output windows with the same summed duration as the windows produced by the actual models, but at random timestamps.

Figure 2 shows the difference between the overlap and the overlap expected by chance for all combinations of models. Each dot indicate a participant to visualise the distribution. Most of the combinations of models produce higher overlap than chance. Significantly higher than chance combinations are indicated with an asterisk (assessed by t-tests with Bonferroni correction for multiple tests thus $\alpha < \frac{0.05}{11}$).

The significant overlap between the backchannels and the disfluency models can be interpreted in a few ways. First, these models use speech audio as an input and design to predict head nods by listeners. The calculated chance, on the other hand, ignores conversational roles completely. Another possible explanation for the higher than chance overlap is that disfluencies shares prosodic characteristics with the backchannels model. This is discussed in depth by the authors that proposed the backchannels model (Ward and Tsukahara, 2000). Furthermore, they claim that disfluencies elicit backchannels.

The overlap between the mimicry600 and the disfluency models is also significantly above chance level. This finding is in line with previous studies that found increased gestures' similarity in free conversations during disfluent utterances (Healey et al., 2015).

We couldn't find a theoretical explanation, nor an explanation based on the models implementation, for the higher than chance overlap for the mimicry600 $\cap$ mimicry4000 combination. Lastly, given the overlaps found between combinations of

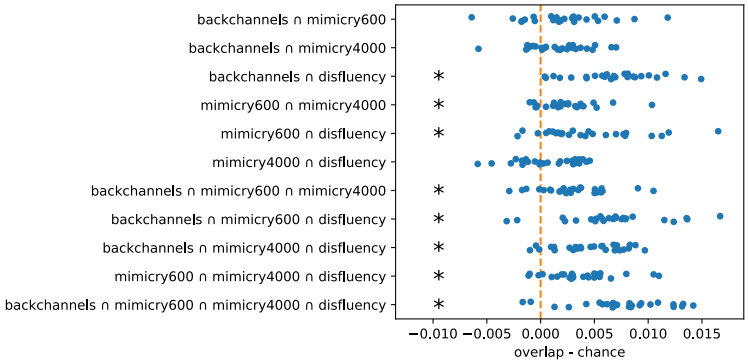


Figure 2: The difference between overlap and overlap expected by chance for combinations of models, per participant. Model combinations with overlap significantly higher than chance are marked with an asterisk.

two models, it is not surprising that combinations of 3 and 4 models also overlap.

4.3 Models comparison

We used precision, recall, and F1-scores (the harmonic mean of precision and recall), per participant, to assess how well the models perform. Because the models' predictions and the detected head nods are timed events we had to adapt the usual precision and recall definitions as follows.

- **True positives:** A prediction is correct if a head nod is detected in the 400ms window around it, as suggested by (Poppe et al., 2010). To make sure not to inflate the number of true positives no more than 1 true positive per head nod was allowed.
- **Precision:** Number of true positives divided by the number of predictions.
- **Recall:** Number of true positives divided by the number of detected nods.

This method is common for backchannel models evaluation (de Kok, 2013), and because the predictions here are similar (temporal point processes) we believe that this method is appropriate.

Figure 3 shows the precision, recall, and F1-score of the models while predicting for participants while holding and not holding the floor. The results suggest that there are no major differences between precision values for the different models. On the other hand, all models achieve higher precision when predicting speaker head nods than when predicting hearer head nods. This effect can be explained by the higher frequency of head nods while holding the floor, as discussed before, as this increase the chance of a prediction to match an actual head nod.

The backchannels and disfluency models achieve higher recall for hearers versus speakers. These models are designed for hearers, so it makes sense that they would perform better in that case. In addition, unlike the other models these rely on speech information to operate. The same input is also used to determine the floor holder. By relying on the same input for the models and for analysis the model can achieve arbitrary high recall. For example, predicting head nods in a high rate for hearers and none at all for speakers can produce an almost perfect recall for hearers and zero for speakers. Therefore, conclusions based on recall about the performance of the backchannels and disfluency models in relation to conversational role should be taken cautiously.

The mimicry600 model also has a higher recall for participants not holding the floor compared to floor holders. It might suggest that mimicry of speakers is more common than mimicry of hearers. Another possible explanation is the increased nodding frequency for floor holders discussed earlier. Because the head nods frequency of floor holders is higher than non floor holders, more predictions are generated for non floor holders, and therefore more detected head nods match with a prediction, inflating the recall value. The mimicry4000 model is not expected to be affected by this, as the average turn duration is 4.35 seconds. In other words, the predictions of the mimicry4000 model often happen beyond the turn that triggered them, so the influence of the floor on this model's recall is negligible.

Figure 4 can help explaining the differences in recall values between the models. The left bar graph

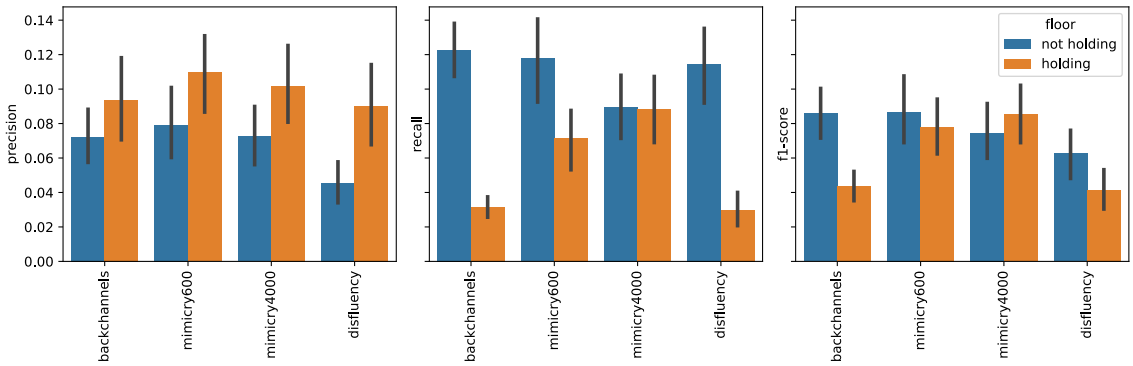


Figure 3: Precision, recall, and F1-score for the models' predictions for participants holding or not holding the floor. The error bards indicate a 95% confidence interval of the mean.

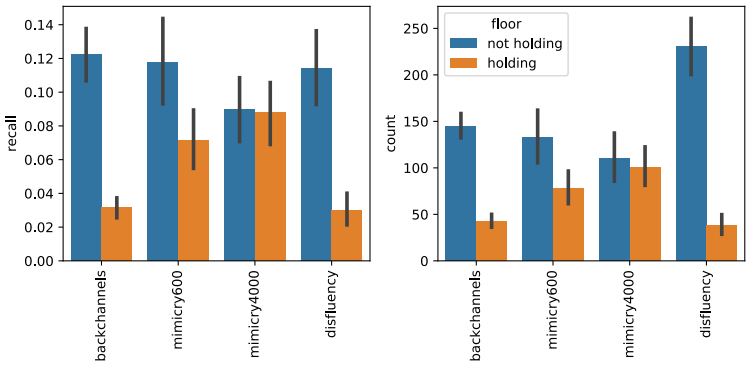


Figure 4: Recall performance of each model (left) and number of predictions generated by each model per participant, when holding and not holding the floor (right). The error bards indicate a 95% confidence interval of the mean.

shows the recall values of the models, for floor holders and non floor holders (same as the middle panel of Figure 3), while the right bar graph show the count of predictions in each of these conditions. For a fixed number of detected head nods, a higher number of predictions will result in an increased recall value. The similarity between the graphs suggests that the differences between the models in terms of their recall values is a direct result of the number of predictions in each condition. This might disprove the suggestion that different models capture speaker and hearer head nods behaviour better than others, as these differences could be due to prediction frequency alone.

4.4 Accounted head nods

Here we assess how many of the detected head nods are accounted by any of the models. A union of all four models is defined by all of the predictions of the models. Considering the definition of recall from the previous section, the recall of this union model indicates how many of the detected head nods are predicted by any of the models.

The average union model recall across all participants is 0.29. This low value indicates that the vast majority of the detected head nods are not accounted by any of the models we investigated. The recall for floor holders is 0.20 and for non floor holders is 0.38. As two out of the four models, the backchannels and disfluency models, are desgined specifically for listeners, the union recall for non floor holders is expected to be higher. Nevertheless, even for non floor holders most of the head nods in the data are left unexplained.

One possible explanation to this low number of accounted head nods is the 400 milliseconds window used for deciding if a prediction is correct. As discussed earlier, this value is common in backchannels research (Poppe et al., 2010), but it is on the shorter side of the spectrum in the literature (de Kok, 2013, page 70). A longer window should increase this number, though the window size choice should not be decided by this result. Another possibility is that the head nods detection

picks movements that are not necessarily head nods. This general issue with the current study is discussed further in the next section. Lastly, although the models discussed here cover a few different theories that are extensively studied, we are sure that they do not cover the entire range of head movement and head nods found in free conversations. As shown earlier, speakers nod more than listeners, and no model here is designed to address this nodding behaviour. A theory that specifically deals with speakers' head movement and nods might shed a light on the head nods that our models failed to predict.

4.5 Limitations

This study relies on an automatic head nod detector. This detector was never validated properly and it seems that it is a source of significant noise and inaccuracies in the analysis. Even on the theoretical level, there is no agreement on what head movements should be considered a head nod (e.g. is it related to pitch or height of head). Further research is required to either come up with a verified model or, perhaps, a broader definition of head nods that fits a concrete form of head movement.

In addition, this research relies on a method to evaluate predictions for timed events that does not produce precision and recall as commonly used, but a slight variations of them. The method looks for predictions around detected nods to estimate true positive, and derive the precision and recall from it. It was used before for similar task (Poppe et al., 2010), to our knowledge the method, and the window choice, were never properly evaluated.

5 Conclusion

This study compares predictions of models inspired by three theories to head nods of floor holders and non floor holders observed in free dialogues as annotated automatically by a head nod detector. According to our detector, the head nod frequency of floor holders is higher than that of non floor holders. We discover significant overlaps between the models, suggesting that the same head nods can often be explained adequately by multiple theories. The differences between the models and their performance for different conversational roles is not conclusive. Whereas some theories achieve higher recall for specific conversational roles the difference is usually a result of more predictions generated for these roles. Lastly, most of the observed head nods in the data went undetected by any of the models we investigated. This can be a result of technically insufficient methods of detecting head nods and matching these with predictions. Alternatively, it might suggest that the theories we investigate here are not enough to explain head nods by speakers and listeners in free conversation.

Acknowledgment

This work was supported by grant EP/L01632X/1 for the Centre for Doctoral Training in Media and Art Technology (MAT http://www.mat.qmul.ac.uk/) from the UK Engineering and Physical Sciences Research Council (EPSRC) and the UK Arts and Humanities Research Council (AHRC).

References

Jeremy N Bailenson, Andrew C Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. 2004. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 13(4):428–441.

Jeremy N Bailenson and Nick Yee. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941–952.

Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893–910.

Tanya L Chartrand and Jessica L Lakin. 2013. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308.

Starkey Duncan, Lawrence J Brunner, and Donald W Fiske. 1979. Strategy signals in face-to-face interaction. *Journal of Personality and Social Psychology*, 37(2):301.

Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, pages 97–121.

Tom Gurion, Julian Hough, Sophie Skach, and Patrick G.T. Healey. Unpublished. A simple, real-time model of turn-taking in dialogue. Manuscript in preparation.

Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. 2018. Are you on my wavelength? interpersonal coordination in naturalistic conversations. Preprint available at https://psyarxiv.com/5r4mj/.

Patrick GT Healey, Mary Lavelle, Christine Howes, Stuart Adam Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker's troubles. In *CogSci*, pages 2506–2511.

Patrick GT Healey, Nicola Plant, Christine Howes, and Mary Lavelle. 2015. When words fail: collaborative gestures during clarification dialogues. In *2015 AAAI Spring Symposium Series,(Chicago)*.

Julian Hough and David Schlangen. 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.

Iwan de Kok. 2013. *Listening Heads*. Ph.D. thesis, University of Twente. SIKS Dissertation Series No. 2013-29.

N Pontus Leander, Tanya L Chartrand, and John A Bargh. 2012. You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological science*, 23(7):772–779.

RM Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural behavior of a listening agent. In *International Workshop on Intelligent Virtual Agents*, pages 25–36. Springer.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pages 176–190. Springer.

Ronald Poppe, Khiet Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents*, pages 146–158. Springer.

Miller S Puckette. 1996. Pure data: another integrated computer music environment. *Proceedings of the second intercollege computer music concerts*, pages 37–41.

Emanuel A Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, 97(5):1295–1345.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Berkeley.

Janna N Vrijsen, Wolf-Gero Lange, Ron Dotsch, Daniël HJ Wigboldus, and Mike Rinck. 2010. How do socially anxious women evaluate mimicry? a virtual reality study. *Cognition and Emotion*, 24(5):840–847.

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

Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

Commitments in German Tag Questions: An Experimental Study

Yulia Clausen and **Tatjana Scheffler**
Department of Linguistics, SFB 1287
University of Potsdam
Germany
{yulia.clausen|tatjana.scheffler}@uni-potsdam.de

Abstract

While plain assertions commit the speaker to the expressed proposition, there are other speech act types that involve various degrees of tentative or projected commitment by both speaker and hearer, for example, tag questions. In this paper, we systematically test the behavior of German tag questions in contexts with different kinds of speaker/hearer commitments, and compare them to their English counterparts. Our experimental results paint a detailed picture of the felicity of German tag questions. We show that individual question tags vary not only on the commitment dimension, but also wrt. speaker certainty and the target of confirmation the speaker intends. Together these categories illustrate the variation of German tagged utterances and reveal aspects that need additional modeling.

1 Introduction

The discourse model by (Farkas and Bruce 2010) was developed for default assertions and questions. These constructions are associated with either direct commitment publicizing (assertions) or lack of commitments (questions). However, there exist "intermediate" constructions, such as tag questions, that involve tentative commitments.

It has been argued that these constructions modify the illocutionary force of an utterance (e.g., Reese and Asher 2007; Malamud and Stephenson 2014; Krifka 2015), and therefore affect which commitments, if any, are effected by the utterance. It has been noted that such modifiers (e.g., same polarity vs. reverse polarity tags in English) differ in meaning and function, and these semantic-pragmatic differences in turn arise from the varied relationships between speakers and hearers in the discourse situation, as well as the prejacent proposition (cf. Malamud and Stephenson 2014: 276).

To accommodate these constructions, Malamud and Stephenson (2014) proposed a modification to the conversational scoreboard model (Farkas and Bruce 2010) by introducing projected commitments, which allowed to distinguish between public and tentative commitments of participants. The extended model accounts for the differences between reverse polarity (RP) and same polarity (SP) tag questions in English. But would it also explain the differences one might encounter with other types of similar constructions? The German language has been shown to have a large inventory of invariant tag questions that on the one hand are largely interchangeable and on the other hand are subject to certain context restrictions, as we have shown in our recent corpus study (Clausen and Scheffler 2020).

We systematically test the commitments contributed by four German question tags in an online experiment, by varying the semantic and pragmatic context around the utterances along the dimensions proposed in previous theoretical and corpus-based literature. We translate the examples from (Malamud and Stephenson 2014) and (Wiltschko et al. 2018) to German, slightly adapting them where necessary. Using these as well as our own examples, we collect felicity judgments from native speakers to find out how suitable the tag variants are in certain contexts. We group the results based on the notions of previous *knowledge*, effected *commitments*, and *judgment* of the conversation participants from (Malamud and Stephenson 2014) and correlate them with the semantic/pragmatic categories *speaker certainty* and *target of confirmation* defined in (Clausen and Scheffler 2020).

After discussing the previous studies by Malamud and Stephenson (2014) on English tag questions and Clausen and Scheffler (2020) on German tag questions in more detail, we present the design and results of our experiment. We show that this kind of experiment is a reliable source of data for semantic/pragmatic analyses. For example, our

results for German declarative clauses, which we tested as a control, match the predictions from the literature. Further, while the German question tags often overlap in their felicity ratings, we identify particular constraints for certain tags, thus demonstrating that (1) "tag question" is not a uniform category with uniform semantics and pragmatics, and (2) German question tag variants differ along other dimensions than the English RP and SP constructions. We conclude by discussing what properties of German tag questions have not yet been taken into account in the existing discourse model.

2 Background

In this section we outline the idea behind the discourse model by Malamud and Stephenson (2014) and its criteria for distinguishing between English RP and SP tag questions. Further, we provide an overview of the variability of German tag questions based on previous work.

2.1 English tag questions

The discourse model (Malamud and Stephenson 2014) is composed of the following components:

1. DC_X: for each participant X, the set of X's public discourse commitments.
2. *Table*: stack of issues to be resolved (the top issue first), where issues are represented as sets of propositions.
3. *Common Ground (CG)*: the set of propositions that all speakers are publicly committed to.
4. *Projected CGs*: a set of potential CGs giving possible resolution(s) of the top issue on the Table in the expected (canonical) next stage of the conversation; this 'next stage' is typically reached within the next few moves responding to the current move.
5. *Projected* DC_X: for the speaker and the hearer(s); allow for moves that give tentative commitments (by adding propositions to the speaker's projected, rather than present, commitments), or to offer the speaker's best guess of commitments of other participants (by adding to others' projected commitment sets).

In this model, commitments are crucial for the analysis of the RP and SP tags. Along with further criteria, such as whose judgment is at issue (speaker's or hearer's) and who has the knowledge about the expressed prejacent proposition, they provide for a clear distinction between the use cases of these tags.

Examples (1) and (2) from (Malamud and Stephenson 2014: 279) show how the felicity conditions for RP and SP tags attached to the same utterance change depending on the situational context.

(1) 'Blushing/Innuendo': A and B are gossiping. A doesn't know anything about B's neighbor. B says, blushing, 'You've GOT to see this picture of my new neighbor!' Without looking, A replies:

A: He's attractive, isn't he?
OK *A: He's attractive, is he?*

(2) 'Seeking agreement': A and B are discussing various traits of their mutual acquaintances. B says, 'I think Bill, more than anything else, is just a really nice guy'. A replies:

OK *A: (But) he's attractive too, isn't he?*
A: He's attractive too, is he?

A typical use case of English RP tags is when the interlocutors are establishing points of agreement (2), so that independent commitments of both speaker and hearer are at issue. Both speaker and hearer are informed, but the speaker is expressing an opinion and seeks agreement (the judgments of both speaker and hearer are at issue). The commitments of the speaker can be definite or tentative, where tentativeness is associated with some lack of confidence.

English SP tags on the other hand involve independent commitments of the hearer, and may or may not involve dependent commitments of the speaker (1). The hearer is informed, whereas the uninformed speaker is making a guess about potential hearer commitment. Hearer judgment is at issue.

2.2 German tag questions

Contrary to English, German does not have variable tags, but offers a large inventory of invariant tags instead (cf. Clausen and Scheffler 2020). The tags show a considerable overlap in use, but are also subject to constraints and display preferences for certain (situational/conversational) contexts. For example, a clear contrast between the tags *oder* and *ne* has been repeatedly mentioned in the literature (see e.g. König 2017; Clausen and Scheffler 2020).

Whereas *oder* is mostly used to express speaker uncertainty and requests confirmation, *ne* prefers contexts where the speaker is certain and therefore the hearer's reaction is not necessarily required. However, there are also contexts that are easily available for both (cf. Example (3) from (Clausen and Scheffler 2020: 2)).

(3) Emma and her family returned from a vacation. It is Monday morning, her mother comes to the kitchen and finds Emma watching movies instead of getting ready for school. Mother says:

Du musst heute nicht in die Schule, {ne? | oder?}
'You don't have to go to school today, right?'

Further, in the previous corpus study we have shown that tag questions expressing confirmation requests for acceptance of a command most frequently occur with the tag *ja* (cf. Clausen and Scheffler 2020: 24).

In the corpus study, we investigated the usage patterns of the five most common German tags based on the syntactic and pragmatic properties of the utterance a tag attaches to (the anchor). We identified that speaker certainty and target of confirmation are crucial in distinguishing between the types of German tags. Speaker certainty reflects the speaker's epistemic status regarding the expressed anchor proposition, i.e. either sure or unsure. The target of confirmation encodes the speaker's expectation for the hearer's reply to the tag question. It includes four types: *p is true* (hearer is asked to confirm whether the anchor proposition is true); *H believes p* (hearer is asked to confirm that they (also) believe the proposition, i.e. show agreement with it); *H perceives p* (hearer is asked to express their awareness of the proposition); *H accepts p* (hearer is asked to confirm that they accept the command expressed in the anchor).

In this study, we test whether the model by (Malamud and Stephenson 2014) is applicable to German tags. First, we conduct an experiment to test in which contexts different tags are (dis)preferred or may not occur at all. Based on the elicited judgments, we then determine the kinds of commitments different tags involve. To group different types of tag questions and contexts, we additionally adopt the categories from (cf. Clausen and Scheffler 2020) which play an important role in the differentiation of German tag variants.

3 Experiment

We conducted an experiment in which the felicity of German tag questions in certain contexts was assessed by native German speakers. We used the contexts from (Malamud and Stephenson 2014), which involve varying degrees of speaker and hearer commitment and knowledge. Additionally, we tested the contexts from (Wiltschko et al. 2018), where the use of declaratives and tag questions with Canadian *eh* is compared on the basis of speaker and hearer knowledge. We translated these examples into German and adapted them where necessary, replacing the utterances in question with the German tag questions. Additionally, we included two of our own examples.

Our experiment contained 20 items, each consisting of a description of a conversational situation (i.e., context) and an utterance produced by one of the conversation participants (all items are listed in Appendix A).

We tested every item in five variants, each time with a different ending: a period (declarative) or one of the common German question tags (*ja*, *ne*, *nicht*, or *oder*). Every participant saw each item only once, i.e. with only one of the five possible endings. To ensure this, we created five questionnaires containing the same utterances each with one of the five different endings, which we randomly distributed among participants (a Latin square design). Figure 1 shows an example of the participants' view of the experiment item WI-06 with the ending *ja* (see Section A.3 of Appendix A for translation). The same item with the other endings was shown to different participants.

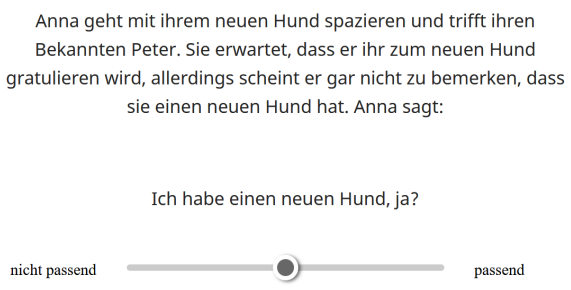


Figure 1: Example of the participants' view of the experiment item WI-06.

Along with the actual items, each questionnaire contained 20 fillers that were constructed analogously to the test items and consisted of a context and an utterance to be judged. For the fillers, we

used various sentence types, as well as tag questions with tags other than the four we are interested in (e.g. *nicht wahr*?), varying the utterances between well-fitting and more awkward.

The questionnaires were created using _magpie[1] and participants were recruited via the crowd-sourcing platform Prolific[2] and paid for participation. We selected the participants based on the following criteria: born and currently living in Germany, monolingually raised speakers of German, German nationality, no language related disorders. They were instructed to carefully read the contexts, paying attention to which of the interlocutors is the speaker of the utterance in question. The participants were asked to rate on a continuous scale how well each utterance fits into the given context. The scale ranged internally from zero to 100 which were displayed as *nicht passend* 'not fitting' and *passend* 'fitting' (see Figure 1). The items were presented in a randomized order one at a time. The participants could control how quickly they wanted to switch to the next item.[3] Once they made a decision, they could not go back and change it.

We obtained data from 149 participants.[4] We used MACE (Hovy et al. 2013) to detect unreliable annotators and excluded 21 submissions with MACE score below 0.5.[5] This left us with 128 submissions with the following number of judgments per questionnaire: 21, 23, 23, 29 and 32.

We represent the results per item aggregated over all participants as box plots. Within the results for each item, we perform pair-wise comparisons of the ending variants with the independent t-test[6] and measure the effect size of differences between the variants with Cohen's d[7] (Cohen 1988).

[1]Minimal Architecture for the Generation of Portable Interactive Experiments: https://github.com/magpie-ea

[2]https://www.prolific.co/

[3]Participants who took longer than the time allotted by the platform were automatically replaced by new participants. Additionally, we rejected submissions that were completed exceptionally fast.

[4]One participant experienced a technical problem with submission, so that their data is not available for us.

[5]MACE clusters annotators based on the overlap between their annotations, providing a quality score for individual annotators even in the absence of gold labels; https://github.com/dirkhovy/MACE.

[6]We used the t-test implementation in statsmodels package v0.11.1: https://www.statsmodels.org/stable/generated/statsmodels.stats.weightstats.ttest_ind.html

[7]We used the implementation in pingouin package v0.3.3: https://pingouin-stats.org/generated/pingouin.compute_effsize_from_t.html

4 Results

In this section, we summarize the results of our experiment. We find that, as expected from the discussion of English tag questions in the literature, the contexts and utterances form groups based on the involved speaker and hearer commitments, knowledge, and judgments. However, these distinctions alone are insufficient to differentiate the German tag variants. Adding speaker certainty and target of confirmation allows for clearer separation between the tag variants, although there are still several items that do not fit into any of the groups and require further study.

We describe the identified groups according to the involved commitments including one representative example per group. A full list of contexts, utterances, box plots, and statistics of participants' ratings is provided in Appendix A.

4.1 Speaker and hearer commitments

The utterances in this group involve independent commitments of both speaker and hearer. Both interlocutors are informed and establishing points of agreement. Therefore, both their judgments are at issue. The speaker is sure about their own opinion and wants to know whether the hearer has the same belief regarding the proposition (target of confirmation *H believes p*). We use item MS-03 (Example (4) and Figure 2) as representative of this group. Items MS-04, WI-05 and WI-11 are listed in Appendix A.

(4) MS-03: Eva and Laura are talking about their mutual friend. Laura says: "I think Mark is, first and foremost, a very nice guy.". Eva answers:

Aber er ist auch attraktiv{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'But he's attractive too{. | right?}'

In English, RP tags, *eh* and declaratives are fine in these contexts (cf. Malamud and Stephenson 2014; Wiltschko et al. 2018). Similarly, in our data, declaratives as well as *ne*, *nicht* and *oder* were rated high. However, *ja* received significantly lower ratings than the other tags (across all items tested in this group) and declaratives (in all items but MS-03). Note that the overall felicity rating for *ja* is still high, even though significantly lower than for the other tags. This example represents a classic use

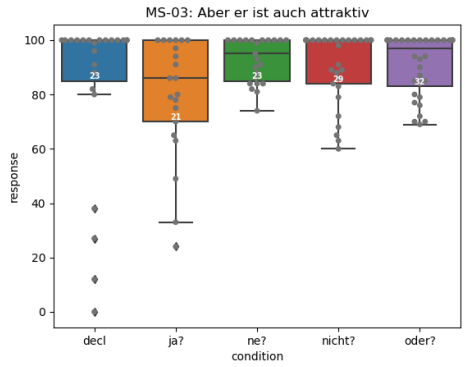


Figure 2: Ratings for item MS-03.

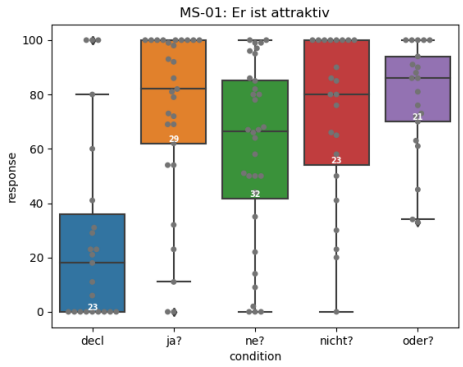


Figure 3: Ratings for item MS-01.

of tag questions – establishing agreement between interlocutors. Consequently, all tags are felicitous here.

4.2 Hearer commitments

This group comprises utterances in contexts where the speaker is uninformed and the hearer's commitment, judgment and knowledge are at issue. The speaker is unsure and wants to obtain confirmation regarding the truth of the expressed proposition (target of confirmation *p is true*). Item MS-01 serves as a representative example of this group (Example (5) and Figure 3). Items MS-02, WI-04 and WI-07 are listed in Appendix A.

(5) MS-01: Anne and Lena are chatting. Anne doesn't know anything about Lena's new neighbor. Lena says: "You HAVE to see this picture of my new neighbor!" and blushes. Anne answers without looking at the picture:

Er ist attraktiv{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'He is attractive{. | right?}'

In (5), the speaker Anne is uninformed about the physical appearance of Lena's neighbor, but is making a guess based on Lena's blushing. She wants Lena to confirm whether the neighbor actually is attractive (target of confirmation). The judgment involved would necessarily be Lena's (the hearer's), because the speaker has not seen the neighbor. Therefore, any (projected) commitment wrt. the neighbor's attractiveness must also be the hearer's.

In English, SP tags and *eh* are preferred over declaratives in these contexts. We observe the same tendency in our data: tag questions were rated higher than the declaratives (statistically significant in three items: MS-01, MS-02, WI-04). Among the tags, *oder* is preferred over *ne* (statistically significant in two items: MS-01, MS-02).

The item WI-07 is less straightforward. In this context, the speaker Ben knows that his friend Maria has wanted to adopt a dog for a long time, and when he finally sees her with a dog, he says "You have a new dog{. | right?}". Here, the declarative was rated relatively high (contrary to declaratives in other items in this group), and there are no significant differences between the tags. The felicity of the declarative in this context can be explained by the fact that Ben knew about the adoption plan and merely states the obvious. However, since he doesn't know this for sure, he still expects a reply, which is indicated by the high ratings of the tags.

4.3 Speaker commitments

In this group, the utterances are speaker centered: the speaker is informed and sure regarding the expressed proposition. Speaker's commitment and judgment are at issue, whereas the hearer is merely asked to express their awareness of the anchor proposition (target of confirmation *H perceives p*).

In the English variants of these examples, declaratives are felicitous. In our German test items, declaratives have continuously been rated higher than the variants with the tags (significant in all items). However, the distributions of the tag ratings are not homogeneous among the six items in this group. Rather, the items form three subgroups: In the first subgroup, the tags were rated consistently lower than declaratives (see Example (6) and

Figure 4 for WI-10 and Appendix A for ON-01).

(6) WI-10: Tom and Laura have decided to attend a lecture together. Tom came on time and has already heard the beginning of the lecture. As Laura arrives late, Tom says to her:

Dem kann man gar nicht folgen{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'It's not really possible to follow him{. | right?}'

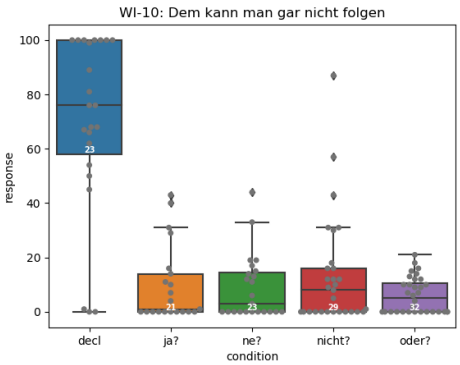


Figure 4: Ratings for item WI-10.

In the second subgroup, the tags received higher ratings compared to the first subgroup, but they display large variation along almost the whole scale, indicating dissent or uncertainty among participants (see Example (7) and Figure 5 for WI-08 and Appendix A for ON-02).

(7) WI-08: Elisa has organized a surprise party for Louis. A friend was meant to get Louis to the party at a certain time. As Louis opens the door, everyone shouts "Surprise!". Louis says:

Das ist vielleicht eine Überraschung{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'Well, that is a surprise{. | right?}'

In the third subgroup, *ne* is rated significantly higher than the other tags, even though with large variation. Additionally, in both items, *ja* is significantly better than *oder* (see Example (8) and Figure 6 for WI-03 and Appendix A for WI-06).

(8) WI-03: Eric is taking a stroll. Suddenly, a bird starts talking to him. At first he is confused. Then a woman approaches him and says:

Überraschung! Du bist bei "Versteckte Kamera" {. | *ja?* | *ne?* | *nicht?* | *oder?*}

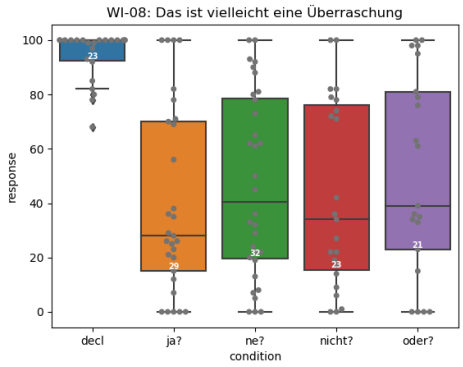


Figure 5: Ratings for item WI-08.

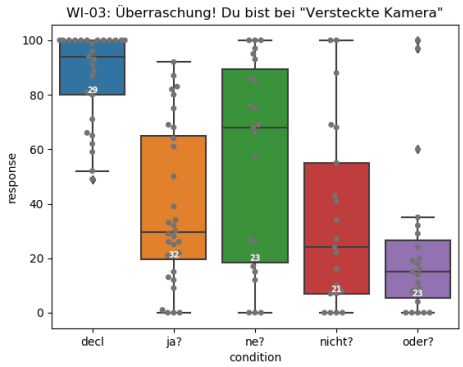


Figure 6: Ratings for item WI-03.

'Surprise! You're on "Candid Camera" {. | right?}'

We attribute these effects to the allocation of knowledge. In the first subgroup, the speaker conveys completely new information to the hearer, which is why the plain declarative is the best option. Since the hearer has no prior knowledge of the proposition at issue, which the speaker is just bringing up for the first time, all kinds of question tags are infelicitous.

In the second subgroup, there exists a certain amount of common ground about the issue raised by the speaker (the surprise for Louis planned by the hearers in WI-08). For this reason, the speaker can assume that the expressed proposition is not completely new to the hearer. This makes the use of tags more appropriate for some language users, who use them to ask the hearer for their opinion/confirmation.

In the third subgroup, there is also (visual) evidence (the weird bird incident in WI-03), for which

the utterance serves as a kind of explanation. Here, the speaker wishes not a confirmation or reply, but merely asks the hearer to take note of what has been said. The tags *ja* and, especially, *ne* are more strongly associated with this target of confirmation, *H perceives p*, than the other tags (cf. Clausen and Scheffler 2020: 26), and this is confirmed by the felicity ratings here, as well.

4.4 Directives

We tested the felicity of German tags in directives (see Examples (9) and (10) for items WI-01 and WI-02, respectively and Figure 7). These utterances are characterized by the speaker requesting confirmation from the hearer about the acceptance of the expressed command (target of confirmation *H accepts p*).[8] Consequently, the hearer's judgment is at issue. The question of commitments is not straightforward with this type of utterances, and they are not discussed in (Malamud and Stephenson 2014). However, in the subsequent work on the Chinese question tag *ba*, which can occur with imperatives, Ettinger and Malamud (2015) ascribe the commitments in directives to the hearer.

(9) WI-01: Lukas and Robin are students and roommates in a dorm. One evening, they are sitting on the couch and watching a movie. Lukas says:

Holst du mir mal ein Bier{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'Will you bring me a beer{. | right?}'

(10) WI-02: Tobias and Max are students and roommates in a dorm. One evening, they are sitting on the couch and watching a movie. Max gets up and heads to the kitchen. Tobias says:

Bringst du mir mal ein Bier mit{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'Will you bring me a beer{. | right?}'

Both items describe similar situations, where two fellow students are spending an evening in a dorm, probably the way they usually do. In both cases, one student is asking the other one to bring him a beer. The difference in WI-02 is that upon hearing the utterance Max is already on his way to the kitchen, and it is more likely that he will comply with it.

As Figure 7 shows, in both cases, declaratives and the tag *ja* were rated higher than the other tags. The use of the untagged variant seems to be the most preferred one when expressing a command in both situations. However, whereas in WI-02, where the hearer is already on his way to the kitchen, almost all participants agree, in WI-01, their consensus is somewhat lower. A possible reason is a lack of courtesy resulting with the use of a plain declarative in this context.

The felicity of *ja* in these items is not surprising. It is significantly better than *ne* (WI-01) and *nicht* and *oder* (both items). This confirms the findings in our corpus study that *ja* strongly correlates with directives (cf. Clausen and Scheffler 2020: 26). The situation with *ne* is more interesting – the ratings are widely spread across the middle range of the scale in WI-01 (both interlocutors are sitting on the couch). In WI-02 (the hearer is on the way to the kitchen), the use of *ne* becomes much more felicitous. It is significantly better than *nicht* in WI-01 and both *nicht* and *oder* in WI-02.

The ratings for the untagged and the *ne* utterances go in line with the English version of these examples, where the declarative is fine in both cases, but Canadian *eh* is felicitous only in WI-02 (cf. Wiltschko et al. 2018: 581,582). Since *ne* prefers situations where the speaker is sure (cf. Clausen and Scheffler 2020), it becomes more acceptable in WI-02, where the speaker has more certainty that his friend will bring him a beer, but wants it to be confirmed.

4.5 Remaining cases

There are four remaining items that did not fit into any of the described groups. For reasons of space, we discuss only two of the items here, but see Section A.5 in Appendix A for all items.

The items MS-06 and MS-07 (Examples (11) and (12) and Figure 8) both involve tentative speaker commitments and independent hearer commitments and are associated with speaker uncertainty. However, they differ with respect to the target of confirmation, as well as involved judgments and knowledge. In these examples, the speaker is making a guess about some issue they are unsure about, while the hearer could resolve this issue in their reply.

(11) MS-06: Tom and Ole are working at the store

[8] Although, strictly speaking, these utterances do not express commands, we adopt this phrasing from (Clausen and Scheffler 2020) to distinguish these utterances from the usual requests essentially encoded in all tag questions.

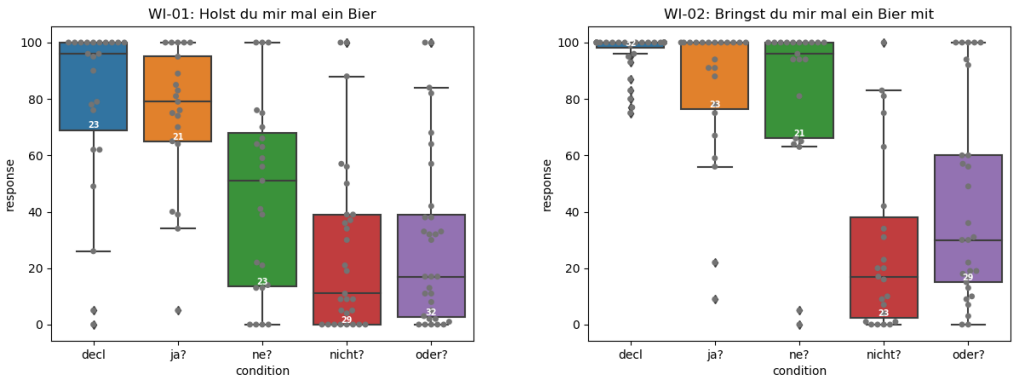


Figure 7: Ratings for items WI-01 and WI-02.

part time, they are sorting the paint buckets in two categories: "red" and "orange". Tom points at one of the buckets and asks: "How would you describe this color?". Ole says:

Es ist rot{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'It's red{. | right?}'

(12) MS-07: Maik is being questioned about the state capitals during an examination. The teacher says: "What is the capital of Hessen?'. Maik is not sure, but thinks it might be Wiesbaden. He answers:

Das ist Wiesbaden{. | *ja?* | *ne?* | *nicht?* | *oder?*}
'This is Wiesbaden{. | right?}'

In English, both utterances are felicitous with RP tags and declaratives (cf. Malamud and Stephenson 2014: 281,282), although declaratives fail to convey the uncertainty of the speaker. In German, we find that differences wrt. target of confirmation prevent them from being put into one group. In MS-06, the interlocutors strive to come to an agreement about an opinion, and both their judgments are at issue (target of confirmation *H believes p*). In MS-07, only the hearer's judgment and knowledge matter (target of confirmation *p is true*), whereas the speaker lacks knowledge.

In both items, declaratives and tags are mostly fine. However, in MS-06 *ja* is significantly worse than other options. In MS-07, *ne* was rated somewhat lower. The tag *oder* was rated consistently high and is significantly better than *ne* and *ja*. This is a typical use case for *oder*, where an uncertain speaker is seeking confirmation from an informed hearer.

4.6 Summary

Table 1 summarizes the distribution of commitments in the English and German variants of the examples discussed in this study.[9] In the most typical use of tags, i.e. establishing agreement where both speaker and hearer commitments are at issue, all German tags are felicitous. Only the tag *ja* is somewhat degraded here. With cases where only hearer commitments are at issue, again, all German tags can be used. However, the preferences for specific tag variants seem to depend on the nature of the context: the tags *ja*, *nicht* and *ne* become less felicitous in situations where the speaker is not just asking for confirmation, but is making a guess regarding the potential hearer commitments based on their behavior. Finally, in the contexts where only speaker commitments are at issue, the only available German tag is *ne*, and, in case of directives, also (and especially) *ja*.

This overview shows that based on the commitments alone we can confirm the difference between the tags *ne* and *oder* shown in our corpus study (Clausen and Scheffler 2020) as well as previous literature: *oder* is infelicitous unless hearer commitments are involved (S+H or H), and *ne* is virtually the only available tag when only speaker commitments are at issue. Also, the special preference for *ja* in directives is confirmed by the results of the experiment.

However, since *ne*, for example, can also be used where hearer commitments are involved, further differentiating criteria are needed to pin down the more subtle differences between the tags. In fact,

[9] Judgments for English reverse and same polarity tags (RP/SP) are taken from (Malamud and Stephenson 2014), and judgments for the Canadian *eh* are taken from (Wiltschko et al. 2018).

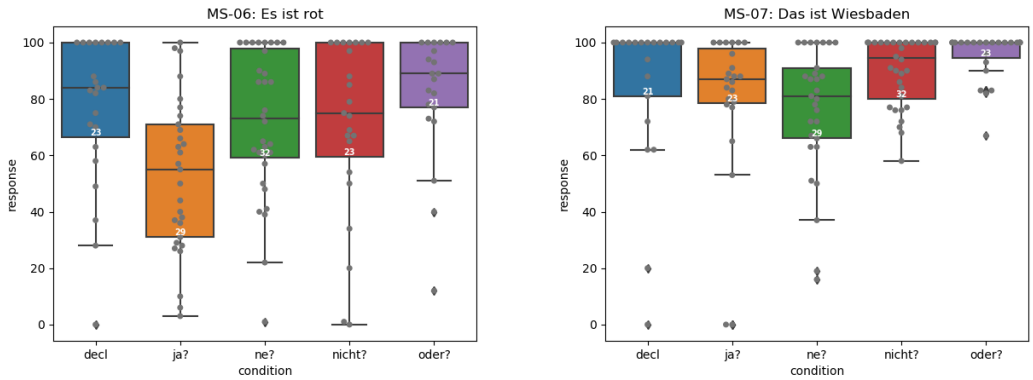


Figure 8: Ratings for items MS-06 and MS-07.

item	RP	SP	*eh*	*ja*	*ne*	*nicht*	*oder*
S+H							
MS-03	✓	#		✓	✓	✓	✓
MS-04	✓	#		✓?	✓	✓	✓
WI-05			✓	✓?	✓	✓	✓
WI-11			✓	✓?	✓	✓	✓
H							
MS-01	#	✓		✓	✓?	✓	✓
MS-02	#	✓		?	?	?	✓
WI-04			✓	✓	✓	✓	✓
WI-07			✓	✓	✓	✓	✓
S							
WI-03			#	#	✓?	#	#
WI-06			✓	#	✓?	#	#
WI-08			#	#?	#?	#?	#?
WI-10			#	#	#	#	#
ON-01				#	#	#	#
ON-02				?	?	?	?
S_{dir}							
WI-01			#	✓	?	#	#
WI-02			✓	✓	✓	#	#

Table 1: Commitments in English and German question tags for all of the studied items except for the remaining cases in Section 4.5: both speaker's and hearer's (S+H), hearer's (H), speaker's (S), and speaker's in directives (S_{dir}).

commitments seem to also not suffice to account for the use of the Canadian tag *eh*, which is fine both where RP tags as well as SP tags are used. Additionally, its felicity conditions vary within the group(s) of speaker commitments.

To resolve the unclear cases, several items need further testing: items WI-08 and ON-02 might be ambiguous and need to be reformulated to restrict the context interpretation.[10] Further, all unclear cases (marked with ✓? and #? in Table 1) should be retested with more participants.

5 Discussion

The method of an online experiment used in this study proved to be a reliable way of assessing the subtle differences in the use of tag variants (and other utterances). The high/low ratings for certain well-studied utterances were conform with the expected outcomes in most cases. For example, declaratives were consistently rated high where speaker commitments are involved. In typical situations, where the tags are used to establish agreement (e.g. item MS-03), the experiment results followed the patterns demonstrated in previous work. This indicates general diligence of participants' replies, that can additionally be filtered for an even higher reliability with an automatic method (MACE).

Adopting various discourse contexts from previous studies on English allowed us to test the use of German tagged utterances under similar conditions. Malamud and Stephenson (2014) differentiate English RP and SP tags based primarily on speaker/hearer commitments. RP tags are felicitous where both speaker and hearer commitments are at issue, and a speaker can make either definite or tentative commitment. SP tags are used in utterances where only hearer commitments are at issue. For German, it is not possible to make a clear distinction between different tag variants based on the involved commitments, which is why we also make use of speaker certainty and target of confirmation. We find that the distinction sure/unsure roughly corresponds to definite/tentative commitments. All our items where speakers make tentative commitments coincide with 'unsure' and cases involving definite

[10] We thank our reviewers for pointing to this issue.

commitments with 'sure'. Target of confirmation is not tied to any other category. Items with the same commitments, such as MS-06 and MS-07, where in English RP tags are felicitous, involve different targets of confirmation, and the German tag ratings differ between these items (see Section 4.5).

Due to a large degree of overlap among the German tags, no direct correspondences to the English RP tags and SP tags can be established. Both English tag variants are matched (to varying degrees) by all German tags we tested. However, it stands out that *ja* is the least felicitous tag in the contexts where RP tags are used. Whereas the judgments for the other tags show good consensus among the raters, *ja* displays considerably more variation.

We find that certain German tags behave consistently regarding commitments: *oder* is felicitous whenever hearer commitments are at issue (both speaker and hearer or only hearer). Overall, it appears that when no hearer commitments are involved, the use of German tags becomes much more restricted.

Finally, we observe that many items show great inter-speaker variation in ratings, indicated by the long boxes in plots. This may point to participants' different interpretations of the contexts, or to differences in opinions on the uses of tags (when we observe a 'gap' between the ratings that seem to split the opinions into positive vs. negative camps, e.g. *ne* in WI-03). Alternatively, even if participants have the same interpretations, these might be vague or they might be uncertain about them.

As future work, we plan to set up a follow up experiment with a larger number of participants to retest the problematic and less clear items using reformulated as well as new contexts. Further, we will look into adapting the described discourse model to the German tags. There are several things that need to be considered. First, there exist utterances not discussed in (Malamud and Stephenson 2014) – with the speaker centered contexts and target of confirmation *H perceives p*. Second, directives expressing requests for acceptance of commands might involve a different kind of speaker commitments. Finally, and perhaps most importantly, it needs to be tested whether the variability of German tags can be explained by the existing model.

Funding Information

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287, project A03.

References

Yulia Clausen and Tatjana Scheffler. 2020. A corpus-based analysis of meaning variations in German tag questions: Evidence from spoken and written conversational corpora. *Corpus Linguistics and Linguistic Theory*, pages 1–31. Published online ahead of print.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*, 2 edition. Lawrence Earlbaum Associates, Hillsdale, NJ.

Allyson Ettinger and Sophia A. Malamud. 2015. Mandarin utterance-final particle ba in the conversational scoreboard. In *Proceedings of Sinn und Bedeutung*, volume 19, pages 232–251.

Donka Farkas and Kim Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics*, 27.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Katharina König. 2017. Question tags als Diskursmarker? – Ansätze zu einer systematischen Beschreibung von *ne* im gesprochenen Deutsch. In H. Blühdorn, A. Deppermann, H. Helmer, and T. Spranz-Fogasy, editors, *Diskursmarker im Deutschen. Reflexionen und Analysen*, pages 233–258. Verlag für Gesprächsforschung, Göttingen.

Manfred Krifka. 2015. Bias in commitment space semantics: Declarative questions, negated questions, and question tags. In *Proceedings of SALT 25*, 0, pages 328–345.

Sophia A. Malamud and Tamina Stephenson. 2014. Three Ways to Avoid Commitments: Declarative Force Modifiers in the Conversational Scoreboard. *Journal of Semantics*, 32(2):275–311.

Brian Reese and Nicholas Asher. 2007. Prosody and the interpretation of tag questions. In *Proceedings of Sinn und Bedeutung*, volume 11, pages 448–462.

Martina Wiltschko, Derek Denis, and Alexandra D'Arcy. 2018. Deconstructing variation in pragmatic function: A transdisciplinary case study. *Language in Society*.

A Experimental items

The items labeled with MS have been adapted from (Malamud and Stephenson 2014), the items labeled with WI have been adapted from (Wiltschko et al. 2018), the items labeled with ON were created by the authors of the paper. We present the items in groups according to the description in the paper. For each item, we present the German context and utterance with their translation, and the box plot of results. Thresholds and effect sizes for significant differences between variants are summarized for each item group in a table at the end of the group.

A.1 Speaker and hearer commitments group

MS-03 Eva und Laura diskutieren über ihren gemeinsamen Bekannten. Laura sagt: "Ich finde Mark ist, vor allem, ein sehr netter Typ.". Eva antwortet:

Aber er ist auch attraktiv{. | ja? | ne? | nicht? | oder?}

'Eva and Laura are talking about their mutual friend. Laura says: "I think Mark is, first and foremost, a very nice guy.". Eva answers:

But he's attractive too{. | right?}'

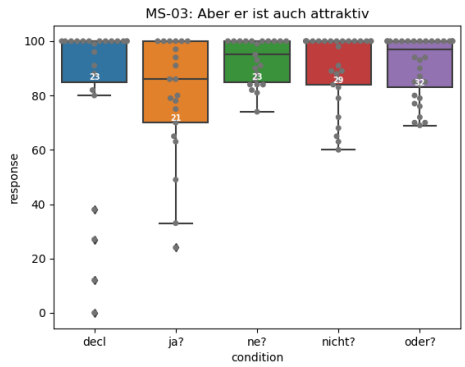


Figure 9: Ratings for item MS-03.

MS-04 Marie und Antonia sind Grundschullehrerinnen. Sie besprechen, was jeder von ihren gemeinsamen Schülern bei dem Tag der offenen Tür vorführen wird. Marie schlägt vor: "Nina könnte ein Gedicht vortragen, weil sie das gut kann.". Antonia sagt:

Aber sie tanzt auch schön{. | ja? | ne? | nicht? | oder?}

'Marie and Antonia are primary school teachers. They are talking about what each of their mutual students will present at the open day at school. Marie suggests: "Nina could read a poem, since she's good at it." Antonia says:

But she dances well too{. | right?}'

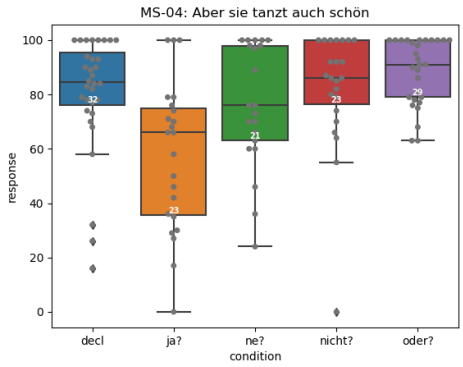


Figure 10: Ratings for item MS-04.

WI-05 Emma und Simon haben zusammen im Kino einen Film gesehen. Als sie das Kino verlassen, sagt Emma:

Der Film war gut{. | ja? | ne? | nicht? | oder?}

'Emma and Simon have seen a movie together. As they leave the theatre, Emma says:

The movie was good{. | right?}'

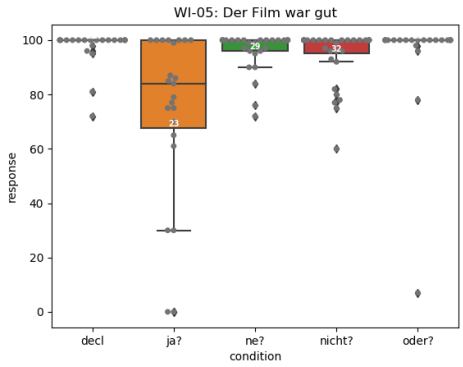


Figure 11: Ratings for item WI-05.

WI-11 Iris und Liam sitzen in einem Vortrag. Iris versteht nicht, worum es geht. Sie guckt Liam an – er sieht ebenso verwundert aus. Iris sagt:

Dem kann man gar nicht folgen{. | ja? | ne? | nicht? | oder?}

'Iris and Liam are attending a lecture. Iris does not understand what it is about. She looks at Liam – he looks equally bewildered. Iris says:

It's not really possible to follow him{. | right?}'

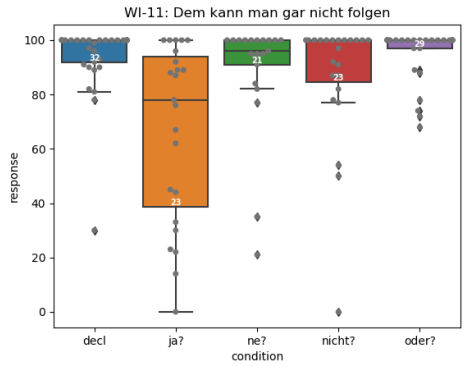


Figure 12: Ratings for item WI-11.

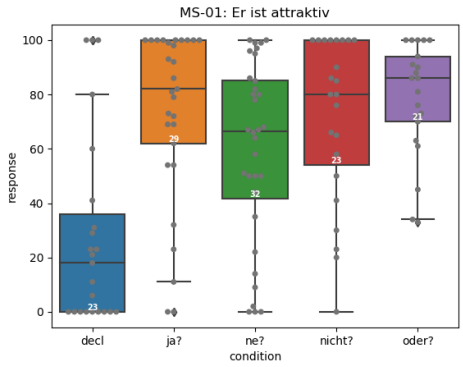


Figure 13: Ratings for item MS-01.

	item ID	*t*	*p*	df	*d*
decl vs. *ja?*	MS-04	3.55	=.001	53	1.0
	WI-05	3.35	=.002	42	1.0
	WI-11	4.25	<.001	53	1.2
ne? vs. *ja?*	MS-03	2.63	=.012	42	0.8
	MS-04	2.48	=.017	42	0.7
	WI-05	3.75	<.001	50	1.0
	WI-11	2.61	=.013	42	0.8
nicht? vs. *ja?*	MS-03	2.11	=.040	48	0.6
	MS-04	3.38	=.002	44	1.0
	WI-05	3.49	=.001	53	1.0
	WI-11	2.37	=.022	44	0.7
oder? vs. *ja?*	MS-03	2.43	=.019	51	0.7
	MS-04	5.53	<.001	50	1.5
	WI-05	2.71	=.010	44	0.8
	WI-11	4.45	<.001	50	1.2

Table 2: Significance tests: t-test and Cohen's *d* for the group **speaker and hearer commitments**.

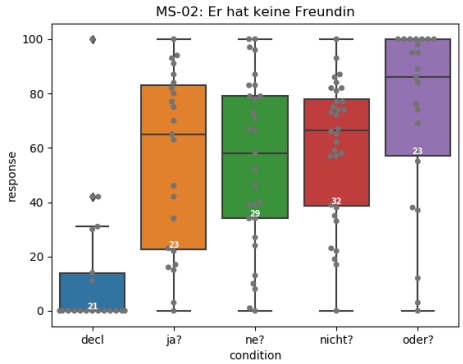


Figure 14: Ratings for item MS-02.

A.2 Hearer commitments group

MS-01 Anne und Lena unterhalten sich. Anne weiß noch nichts über Lenas neuen Nachbarn. Lena sagt: "Du MUSST dir mal dieses Bild von meinem neuen Nachbarn ansehen!" und wird dabei rot. Anne antwortet, ohne auf das Bild zu schauen:

Er ist attraktiv{. | ja? | ne? | nicht? | oder?}

'Anne and Lena are chatting. Anne doesn't know anything about Lena's new neighbor. Lena says: "You HAVE to see this picture of my knew neighbor!" and blushes. Anne answers without looking at the picture:

He is attractive{. | right?}'

MS-02 Sarah und Julia unterhalten sich. Sarah weiß noch nichts über Julias neuen Nachbarn. Julia sagt: "Du MUSST meinen neuen Nachbarn kennenlernen" und wird dabei rot. Sarah sagt:

Er hat keine Freundin{. | ja? | ne? | nicht? | oder?}

'Sarah and Julia are chatting. Sarah doesn't know anything about about Julia's new neighbor. Julia says: "You HAVE to meet my knew neighbor!" and blushes. Sarah says:

He doesn't have a girlfriend{. | right?}'

WI-04 David und Finn sitzen in einer Bar und tauschen sich über die letzten Neuigkeiten aus. Finn glaubt David bei der gestrigen Episode der "Versteckten Kamera" gesehen zu haben. Er möchte wissen, ob es wirklich David war und sagt:

Du warst gestern bei "Versteckte Kamera"{. | ja? | ne? | nicht? | oder?}

'David and Finn are sitting in a bar and discussing the latest news. Finn believes he saw David on last night's episode of "Candid Camera". He wants to know whether it really was David and says:

You were on "Candid Camera" yesterday{. | right?}'

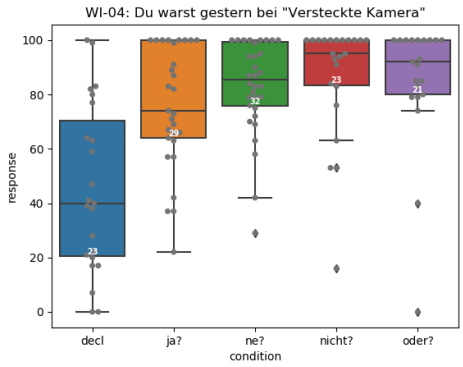


Figure 15: Ratings for item WI-04.

WI-07 Bei einem Spaziergang trifft Sven seine Bekannte Maria. Maria hat ihren Hund dabei. Sven weiß, dass sie schon seit langem einen Hund adoptieren wollte. Er nimmt an, dass sie das nun endlich getan hat. Sven sagt:

Du hast einen neuen Hund{. | ja? | ne? | nicht? | oder?}

'During a stroll Sven meets his acquaintance Maria. Maria has her dog with her. Sven knows that she's been wanting to adopt a dog for a while. He assumes that she has finally done so. Sven says:

You have a new dog{. | right?}'

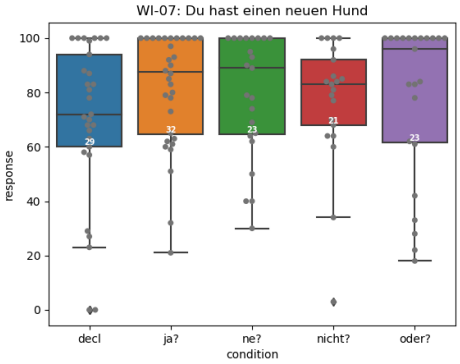


Figure 16: Ratings for item WI-07.

	item ID	t	p	df	d
ja? vs. decl	MS-01	4.86	<.001	50	1.4
	MS-02	4.82	<.001	42	1.5
	WI-04	4.09	<.001	50	1.1
ne? vs. decl	MS-01	3.33	=.002	53	0.9
	MS-02	5.03	<.001	48	1.4
	WI-04	5.74	<.001	53	1.6
nicht? vs. decl	MS-01	4.48	<.001	44	1.3
	MS-02	6.74	<.001	51	1.9
	WI-04	5.27	<.001	44	1.6
oder? vs. decl	MS-01	5.66	<.001	42	1.7
	MS-02	6.71	<.001	42	2.0
	WI-04	4.68	<.001	42	1.4
oder? vs. *ne?*	MS-01	2.32	=.025	51	0.7
	MS-02	1.99	=.052	50	0.6

Table 3: Significance tests: t-test and Cohen's d for the group **hearer commitments**.

A.3 Speaker commitments group

WI-10 Tom und Laura haben beschlossen, zusammen zu einem Vortrag zu gehen. Tom war rechtzeitig da und hat bereits den Anfang des Vortrags angehört. Als Laura verspätet reinkommt, sagt Tom zu ihr:

Dem kann man gar nicht folgen{. | ja? | ne? | nicht? | oder?}

'Tom and Laura have decided to attend a lecture together. Tom came on time and has already heard the beginning of the lecture. As Laura arrives late, Tom says to her:

It's not really possible to follow him{. | right?}'

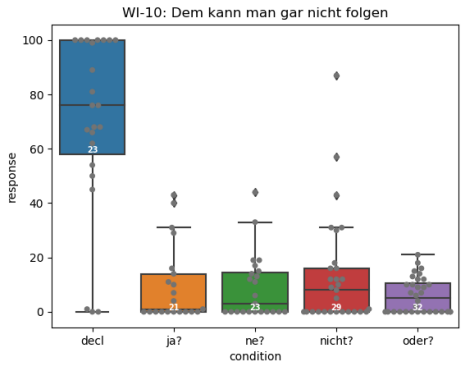


Figure 17: Ratings for item WI-10.

ON-01 Tanja war im Kino und hat einen neuen Film gesehen. Als sie sich später am Abend mit Jana trifft, die diesen Film noch nicht gesehen hat, sagt Tanja:

Dieser Film hat einen Oscar gewonnen und der war auch total gut{. | ja? | ne? | nicht? | oder?}

'Tanja was at the theatre and has seen a new movie. As she later meets up with Jana, who hasn't seen this movie yet, Tanja says:

This movie has won an Oscar and it was very good indeed{. | right?}'

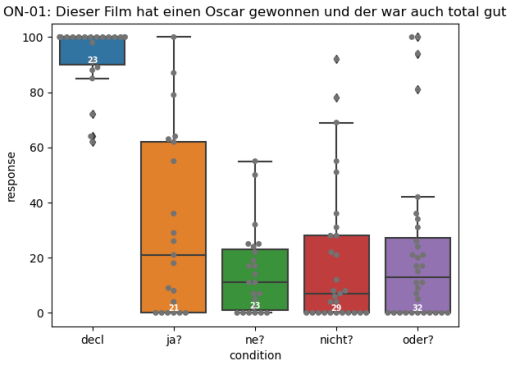


Figure 18: Ratings for item ON-01.

WI-08 Elisa hat für Louis eine Überraschungsparty organisiert. Ein Freund sollte Louis zu einer bestimmten Uhrzeit zu der Party bringen. Als Louis die Tür aufmacht, schreien alle auf einmal "Überraschung!". Louis sagt:

Das ist vielleicht eine Überraschung{. | ja? | ne? | nicht? | oder?}

'Elisa has organized a surprise party for Louis. A friend was meant to get Louis to the party at a certain time. As Louis opens the door, everyone shouts "Surprise!". Louis says:

Well, that is a surprise{. | right?}'

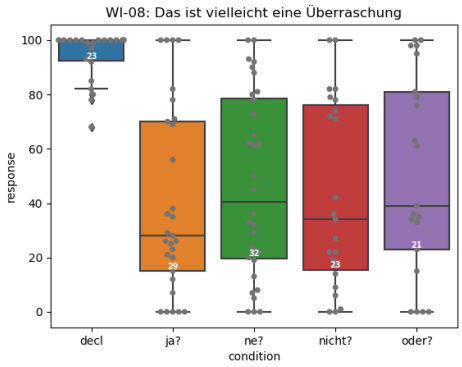


Figure 19: Ratings for item WI-08.

ON-02 Fabian hat sich früher um die Pflanzen im Garten seiner Eltern gekümmert. Jetzt lebt er in einer kleinen Wohnung und möchte dort auf dem Balkon etwas pflanzen. Dafür hat er einige pflegeleichte Pflanzensorten gekauft. Am Telefon mit seinen Eltern sagt Fabian:

Die brauchen natürlich nicht so viel Arbeit{. | ja? | ne? | nicht? | oder?}

'Fabian used to take care of the plants in his parents' garden. Now he lives in a small apartment and wants to plant something on the balcony. For this he bought some easy-care plant varieties. On the phone with his parents Fabian says:

They obviously don't need that much work{. | right?}'

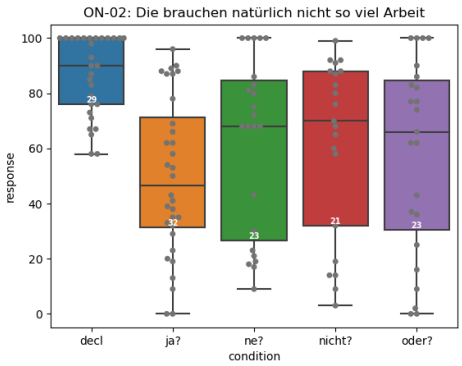


Figure 20: Ratings for item ON-02.

WI-03 Eric macht einen Spaziergang. Plötzlich spricht ihn ein Vogel an. Erst ist er völlig verwirrt. Dann kommt eine Frau zu ihm und sagt:

Überraschung! Du bist bei "Versteckte Kamera"{. | ja? | ne? | nicht? | oder?}

'Eric is taking a stroll. Suddenly, a bird starts talking to him. At first he is confused. Then a woman approaches him and says:

Surprise! You're on "Candid Camera"{. | right?}'

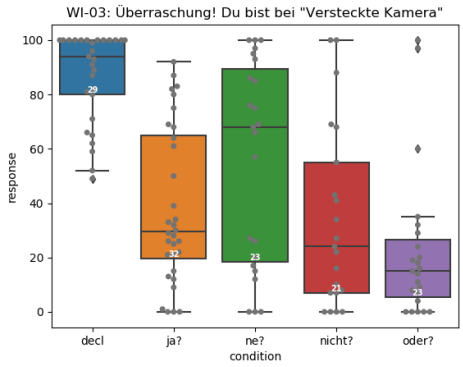


Figure 21: Ratings for item WI-03.

WI-06 Anna geht mit ihrem neuen Hund spazieren und trifft ihren Bekannten Peter. Sie erwartet, dass er ihr zum neuen Hund gratulieren wird, allerdings scheint er gar nicht zu bemerken, dass sie einen neuen Hund hat. Anna sagt:

Ich habe einen neuen Hund{. | ja? | ne? | nicht? | oder?}

'Anna goes for a walk with her new dog and runs into her acquaintance Peter. She expects him to congratulate her on getting a new dog, however, he seems not to notice, that she has a new dog. Anna says:

I have a new dog{. | right?}'

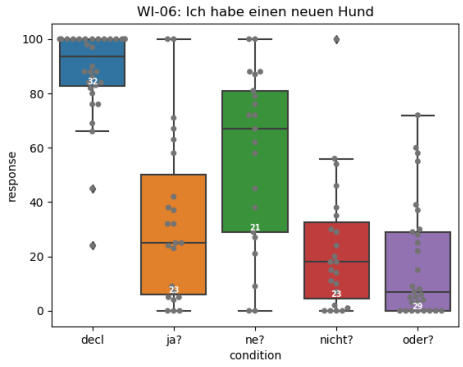


Figure 22: Ratings for item WI-06.

A.4 Directives group

WI-01 Lukas und Robin sind Studenten und Mitbewohner einer WG. An einem Abend sitzen sie auf der Couch und schauen einen Film. Lukas sagt:

Holst du mir mal ein Bier{. | ja? | ne? | nicht? | oder?}

'Lukas and Robin are students and roommates in a dorm. One evening, they are sitting on the couch and watching a movie. Lukas says:

Will you bring me a beer{. | right?}'

	item ID	*t*	*p*	df	*d*
decl vs. *ja?*	WI-03	7.95	<.001	59	2.0
	WI-06	8.36	<.001	53	2.3
	WI-08	7.35	<.001	50	2.1
	WI-10	7.73	<.001	42	2.3
	ON-01	8.45	<.001	42	2.6
	ON-02	6.32	<.001	59	1.6
decl vs. *ne?*	WI-03	3.99	<.001	50	1.1
	WI-06	4.46	<.001	51	1.3
	WI-08	6.65	<.001	53	1.8
	WI-10	8.35	<.001	44	2.5
	ON-01	19.3	<.001	44	5.7
	ON-02	3.70	<.001	50	1.0
decl vs. *nicht?*	WI-03	7.30	<.001	48	2.1
	WI-06	11.6	<.001	53	3.2
	WI-08	6.95	<.001	44	2.0
	WI-10	7.55	<.001	50	2.1
	ON-01	12.6	<.001	50	3.5
	ON-02	3.77	<.001	48	1.1
decl vs. *oder?*	WI-03	10.3	<.001	50	2.9
	WI-06	14.0	<.001	59	3.6
	WI-08	5.48	<.001	42	1.7
	WI-10	10.7	<.001	53	2.9
	ON-01	10.7	<.001	53	2.9
	ON-02	4.03	<.001	50	1.1
ne? vs. *ja?*	WI-03	1.96	=.055	53	0.5
	WI-06	2.51	=.016	42	0.8
	ON-01	-2.17	=.036	42	-0.7
ne? vs. *nicht?*	WI-03	2.01	=.051	42	0.6
	WI-06	4.00	<.001	42	1.2
ne? vs. *oder?*	WI-03	3.42	=.001	44	1.0
	WI-06	5.16	<.001	48	1.5
ja? vs. *oder?*	WI-03	2.04	=.046	53	0.6
	WI-06	2.14	=.037	50	0.6

Table 4: Significance tests: t-test and Cohen's *d* for the group **speaker commitments**.

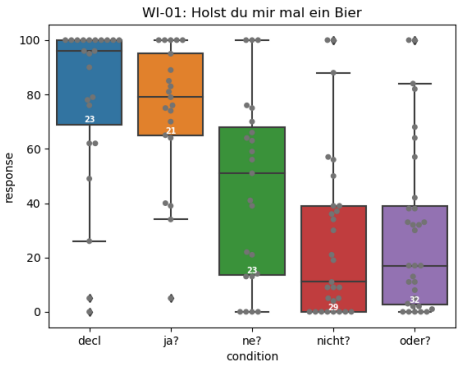


Figure 23: Ratings for item WI-01.

WI-02 Tobias und Max sind Studenten und Mitbewohner einer WG. An einem Abend sitzen sie auf der Couch und schauen einen Film. Max steht auf und geht in Richtung Küche. Tobias sagt:

Bringst du mir mal ein Bier mit{. | ja? | ne? | nicht? | oder?}

'Tobias and Max are students and roommates in a dorm. One evening, they are sitting on the couch and watching a movie. Max gets up and heads to the kitchen. Tobias says:

Will you bring me a beer{. | right?}'

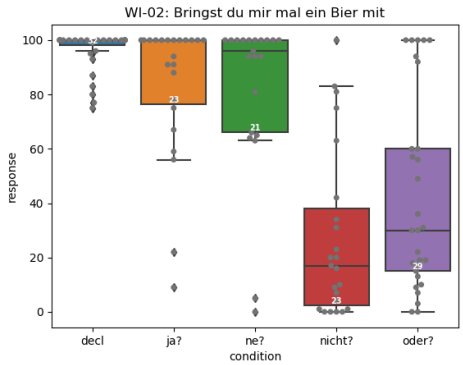


Figure 24: Ratings for item WI-02.

A.5 Remaining cases

MS-05 Sophia und Lisa unterhalten sich. Lisa hat den Nachbarn von Sophia noch nie getroffen und fragt: "Was hältst du von deinem neuen Nachbarn?". Sophia ist sich nicht sicher, was genau damit gemeint ist: seine Tauglichkeit als Nachbar oder als mögliches Date. Sophia sagt:

Er ist attraktiv{. | ja? | ne? | nicht? | oder?}

'Sophia and Lisa are talking. Lisa has not met Sophia's neighbor yet and asks: "What do you think of your new neighbor?". Sophia is not sure, what exactly is meant by that: his neighborliness or suitability for dating. Sophia says:

He's attractive{. | right?}'

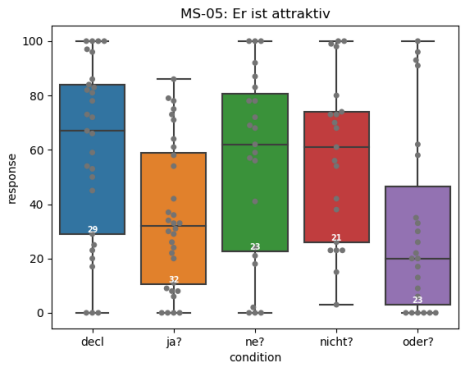


Figure 25: Ratings for item MS-05.

	t	*p*	df	*d*
decl vs. *ja?*	3.18	=.002	59	0.8
decl vs. *oder?*	3.00	=.004	50	0.8
ne? vs. *ja?*	2.35	=.022	53	0.6
ne? vs. *oder?*	2.29	=.027	44	0.7
nicht? vs. *ja?*	2.71	=.009	51	0.8
nicht? vs. *oder?*	2.58	=.013	42	0.8

Table 5: Significance tests: t-test and Cohen's *d* for item MS-05.

MS-06 Tom und Ole haben einen Nebenjob in einem Geschäft und sortieren gerade Farbeimer in zwei Kategorien: "rot" und "orange". Tom zeigt auf einen Eimer mit orange-roter

Farbe und fragt: “Wie würdest du diese Farbe bezeichnen?”. Ole sagt:

Es ist rot{. | ja? | ne? | nicht? | oder?}

‘Tom and Ole are working at the store part time, they are sorting the paint buckets in two categories: “red” and “orange”. Tom points at one of the buckets and asks: “How would you describe this color?”. Ole says:

It’s red{. | right?}’

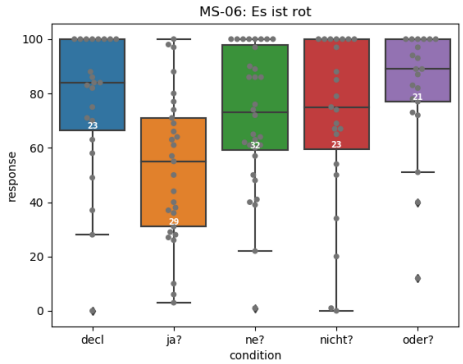


Figure 26: Ratings for item MS-06.

	t	p	df	d
decl vs. *ja?*	3.15	=.003	50	0.9
ne? vs. *ja?*	2.79	=.007	59	0.7
nicht? vs. *ja?*	2.21	=.031	50	0.6
oder? vs. *ja?*	3.99	<.001	48	1.1

Table 6: Significance tests: t-test and Cohen’s d for item MS-06.

MS-07 Maik wird bei einer Prüfung nach den Landeshauptstädten gefragt. Der Lehrer sagt: “Was ist die Hauptstadt von Hessen?”. Maik ist sich nicht sicher, denkt aber, dass es Wiesbaden sein könnte. Er antwortet:

Das ist Wiesbaden{. | ja? | ne? | nicht? | oder?}

‘Maik is being questioned about the state capitals during an examination. The teacher says: “What is the capital of Hessen?’. Maik is not sure, but thinks it might be Wiesbaden. He answers:

This is Wiesbaden{. | right?}’

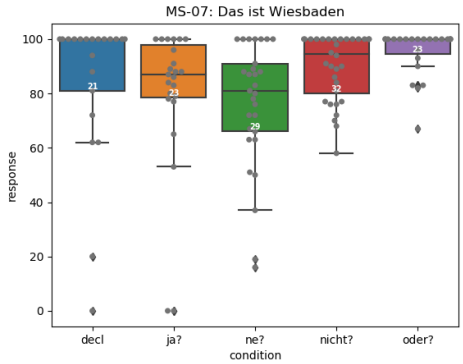


Figure 27: Ratings for item MS-07.

	t	p	df	d
nicht? vs. *ne?*	2.92	=.005	59	0.7
oder? vs. *ne?*	3.77	<.001	50	1.1
oder? vs. *ja?*	2.66	=.011	44	0.8

Table 7: Significance tests: t-test and Cohen’s d for item MS-07.

WI-09 Ines hat für Ben eine Überraschungsparty organisiert. Ein Freund sollte Ben zu einer bestimmten Uhrzeit zu der Party bringen. Als Ben die Tür aufmacht, schreien alle auf einmal “Überraschung!”. Ines sagt:

Das ist vielleicht eine Überraschung{. | ja? | ne? | nicht? | oder?}

‘Ines has organized a surprise party for Ben. A friend was meant to get Ben to the party at a certain time. As Ben opens the door, everyone shouts “Surprise!”. Ines says:

Well, that is a surprise{. | right?}’

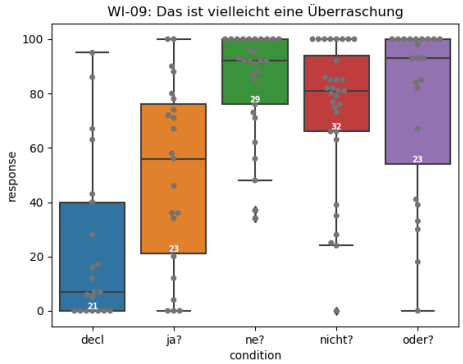


Figure 28: Ratings for item WI-09.

	t	p	df	d
ja? vs. decl	2.71	=.010	42	0.8
ne? vs. decl	8.68	<.001	48	2.5
nicht? vs. decl	6.38	<.001	51	1.8
oder? vs. decl	5.62	<.001	42	1.7
ne? vs. *ja?*	4.69	<.001	50	1.3
nicht? vs. *ja?*	2.97	=.005	53	0.8
oder? vs. *ja?*	2.75	=.009	44	0.8

Table 8: Significance tests: t-test and Cohen’s d for item WI-09.

The contribution of dialogue act labels for convergence studies in natural conversations

Simone Fuscone[1,2], **Benoit Favre**[2] and **Laurent Prévot**[1,3]
[1] Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[2] Aix Marseille Univ, CNRS, LIS, Marseille, France
[3] Institut Universitaire de France, Paris, France

Abstract

Speakers convergence in conversation had been studied because of its relevance for cognitive models of communication as well as for dialogue systems adaptation to the user. Convergence effects have been established on controlled datasets. Tracking interpersonal dynamics on generic corpora has provided positive but more contrasted outcomes. We propose here to enrich large conversational corpora with dialogue act information and to use acts as filters to create sub sets featuring homogeneous conversational activity. Those sub sets allow a more precise comparison between speakers' speech variables. Our experiences consist of comparing convergence on low level variables (Energy, Pitch, Speech Rate) measured on raw datasets, with human and automatically labelled datasets. We found that such filtering does help in observing convergence suggesting that future studies should consider such high level dialogue activity types and the related NLP tools as important aspects for analyzing conversational interpersonal dynamics.

1 Introduction

Individual speech characteristics of the participants engaged in a conversation depend on speech activities of both participants and on their interaction. This intuition has taken shape in numerous studies and even resulted in general models of communication as described in the *accommodation theory* (Giles et al., 1991) and in the *interactive alignment* model (Pickering and Garrod, 2004). The phenomenon has been studied under different angles such as speech or phonetic *convergence* (Street, 1984; Pardo, 2006), prosodic *entrainment* (Levitan and Hirschberg, 2011; Truong and Heylen, 2012).

Earlier work from (Edlund et al., 2009) found that participants to a conversation tend to be more similar (in terms of gaps and pauses duration) to their partner than chance would predict. This is in line with the more recent study on speech rate from (Cohen Priva et al., 2017). However, the absence of significant results in comparing the inter-speaker distance in the first and second halves of the conversation make them conclude that convergence "cannot be captured by sampling the distance between speakers at two different stages of the dialogue". The study in (Truong and Heylen, 2012) conducted a similar experiment (investigating intensity and pitch) on English MapTask (Anderson et al., 1991) with little results as well and conclude that "the measure of alignment remains a complicated matter due to its dynamic nature and to the social factors that can influence the amount of convergence". Although the literature has proposed many different ways of approaching these questions, they all rely on extracting features from a certain period or point in time and comparing them.

In this paper we would like to push further the investigation of interpersonal dynamics in real-life corpora. Considered from the angle of speech and linguistic variables, an essential aspect of conversational corpora is their huge variability. This variability is due to a large extent to the different conversational activities speakers can participate in. For instance, they can enter in a storytelling sequence in which one interlocutor become less active, enter into a heated debate, etc... We propose here to use the latest dialogue tagging techniques to create subsets obtained from dialogue acts filtering in which frequent dialogue acts are used as a proxy to characterize the conversational activity of a given turn, and therefore compare data points that are more homogeneous.

Our approach combines three ingredients that were not used together in the past and current literature to our best knowledge. First, we recognize

the social factor and hypothesis that the lower level variables (such as energy) are less prone to social factor or strategic adaptation that other variables. This is justified by the fact that production and perception processes are deeply linked in conversation and that entrainment at those lower level seems to be automatic. Second, we consider that some amount of convergence should be observed within the time frame of a conversation even though we acknowledge that the "two halves" approach is extremely crude in this respect. Third, similarly to (Cohen Priva et al., 2017) our approach is based on a large conversational corpus with the intention of overcoming the noise and the small size of the effect by increasing the amount of data considered.

The paper starts with a review (Section 2) of related works describing previous studies that focus on convergence. Then in Section 3 we introduce the dataset that we used in our analysis and we describe the DA tagger. Section 4 shows the feature extraction of the variables we scrutinized in this work and the methodology we apply to measure convergence. The results are illustrated in Section 5. Finally, we discuss the results and possible future improvements and open aspects in Section 6.

2 Related work

Conversational Interpersonal Dynamics have been approached at different granularity levels and for a large range of variables. In terms of granularity, studies can be (i) **Inter-conversation** comparisons ; (ii) **Intra-conversation** (focusing on the dynamics within conversations). Inter-conversation comparison could be simple correlation studies between the speakers (Edlund et al., 2009) or, when the data allows, comparison between values of a speaker and its conversational partners vs. a speaker and all other non-partner corpus participants (Cohen Priva et al., 2017). Intra-conversation studies vary a lot in term of approaches ranging from coarse-grained "difference-in-difference" *convergence* (Edlund et al., 2009; Truong and Heylen, 2012; Cohen Priva and Sanker, 2018) approaches consisting in comparing differences between speakers in different intervals to fine-grained *synchrony* approaches using sliding window in order to compare local speaker similarities (Truong and Heylen, 2012).

While a large body of carefully controlled experiments on lab speech provided results on convergence, the results on real corpora (from the studies listed in the previous paragraph) provide a more complex pictures, the exhibition of some effect for some variables but with a relative fragility of the magnitude of these effects (Fuscone et al., 2018) and arise overall many comments on the methodological subtleties and mentions to room for improvements (See (Truong and Heylen, 2012; Cohen Priva and Sanker, 2018) for instance).

Finally in regards to the scrutinized features of speech, (Natale, 1975) has shown that modulating experimentally the intensity of voice of a shadow partner during the conversation led the participants to change their intensity converging to the partner's values. (Edlund et al., 2009) targeted interactional variables of pauses and inter-speakers gaps duration showing convergence between the speakers. Pitch was the focus of (Truong and Heylen, 2012; Bonin et al., 2013) while speech rate had been investigated for example in (Cohen Priva et al., 2017).

Our hypothesis, derived from both the empirical results above and the theoretical models (Pickering and Branigan, 1998; Giles and Powesland, 1997), is that automatic *entrainment* and strategic *adaptation* are blending in to produce *convergence* and *synchrony* phenomena. This blend is complex but intuitively, respectively low-level variables (such as intensity) and high-level variables (such as lexical or syntactic choices) may be more impacted by automatic entrainment and strategic adaptation. This could explain why firmer and more results seems to be obtained on low-level variables (Natale, 1975; Levitan, 2014).

To summarize conversational interpersonal dynamics, that is supposed to be an established phenomenon, can be surprisingly difficult to track in real conversations. The main issue is the heterogeneity of the speech activities both within a conversation (huge and rapid variation across time depending on what interaction is doing) and across conversations. We propose here to use dialogue acts to "organize" and filter conversational activities within large generic corpora. Moreover, to account for *adaptation* one must take precise care of speaker profiles. Our approach therefore focuses on relatively low level variable to avoid as much as possible the "adaptation" part of the interpersonal dynamics.

Our question is whether we can observe more reliably interpersonal dynamics in raw, manually

	All	St.	Opi.	Bc.
SWBD	180h	-	-	-
SWDA	41h	17h	7 h	1h
SW-Auto	119h	47h	19h	3h

Table 1: Results Summary : Hours of speech excluding noise and silence

DA-tagged (smaller) or automatically DA-tagged (larger) dataset even if we reduce the data size. An underlying question being whether the noise introduced by the DA-tagging uncertainty and the data size reduction is compensated by the gain in homogeneity between the material that is compared. Finally, a side question is whether some DA-specific subsets (e.g backchannel) are interesting sub datasets for studying specific interpersonal effects.

3 Dialogue Act Filtering and DataSets

In conversational speech, a large amount of variability that can be observed for a given speech variable across time is due to the different speech activities that are performed by the speakers. For instance, participants can remain rather passive, be very narrative or inquisitive, etc.; We attempt to deal with this crucial issue by using dialogue acts (DA) as a proxy for creating subset corresponding to different "dialogue activity". We only used the acts dominating the DA distribution: *Statement, Opinion, Backchannel*

3.1 Raw dataset

Switchboard (Godfrey et al., 1992) is a corpus of telephonic conversations between randomly assigned speakers[1] of American English discussing a preassigned topic. The corpus consists of 2430 conversations (of an average duration of 6 minutes) for a total of 260 hours, involving 543 speakers. The corpus has audio, time aligned transcripts and a segmentation into *utterances*.

3.2 Manually tagged datasets

642 Switchboard conversations have been segmented and annotated for dialogue acts (DA) (Calhoun et al., 2010). The DA-tagged set has been simplified to 42 tags but a few of them (Statement: 36%, Backchannel: 19%, Opinion: 13%) are dominating the distribution, illustrated in Table 2. See (Stolcke et al., 1998) for the details.

[1] Speakers therefore do not know each other.

DA type	Example
STA	*"And that was pretty heartrending for her"*
OPI	*"money seems to be too big of an issue."*
BAC	*"Uh-huh."*
AGR	*"you're right"*

Table 2: Examples for the DA types used.

	Prediction Score		
Class	Precision	Recall	F1
BAC	0.78	0.84	0.81
STA	0.78	0.73	0.75
OPI	0.47	0.61	0.53
OTH	0.71	0.68	0.69

Table 3: Prediction score of the Turn Tagger for the classes: Backchannel (BAC), Statement (STA), Opinion (OPI).

3.3 Automatically tagged dataset

Since Switchboard provides human annotation for a subset of the entire corpus, we propose to use a **turn tagger** on each conversation.

We considered to first try the categories that contain the majority of classes, said *Statement* (STA), *Backchannel* (BAC), *Opinion* (OPI) and *Other* (OTH) as shown in Figure 1. We used as train, development and test sets the NXT Switchboard corpus that contains annotated Dialog Acts for 642 conversations. As the dialog acts do not match the turn segmentation of the conversations, we label each turn of the corpus by assigning one of the majority class, among the DA tags used in the turn. The model we used is described in (Auguste et al., 2018) and inspired by the model of (Yang et al., 2016). It is a two level hierarchical Neural Network (with learning rate = 0.001, batch size = 32, max length of each turn = 80, embeddings words dimension = 200). In the first level each turn is treated singularly taking into account the words that form the turn while the second level is used to take into account the whole turn in the context of the conversation. Each level is a bidirectional Long Short Term Memory (LSTM). We used 80% of switchboard data as training set, 10% for development and 10% for the test set.

Score for the classes are reported in Table 3. In particular the class Opinion presents a low score in both precision and recall. As expected the class Other as well, that contains all other DA types, has a low score. In order to improve

	Prediction Score		
Class	Precision	Recall	F1
BAC+AGR	0.88	0.85	0.86
STA+OPI	0.84	0.92	0.87
OTH	0.62	0.49	0.55

Table 4: Prediction score of the Turn Tagger for the 3 classes.

the precision and recall of the tagger, we try to reduce the noise of the tagging task by adopting a grouping of similar categories, corresponding to *Statement+Opinion* (STA+OPI), *Backchannel+Agreement* (BAC+AGR) and *Other* (OTH) which includes all the other dialog acts. This grouping was obtain by first considering only the DA dominating the distribution. Then we manual inspect many examples of each dialogue act and figure that although functionally different *statements* and *opinions* on the one hand; *backchannel* and *Agreement* on the other hand corresponded to very similarly conversational activities. More precisely, the former have clear *main speaker* feeling with a lot of semantic content while the later have a much more *listener* nature with various kinds of feedback related lexical items (see Table 2). The resulting distribution is 52% STA+OPI, 25% BAC+AGR and 23% of OTH. The Accuracy of the DA tagger is 81% on the test set, the details for each category is reported in table 4. The accuracy of the class OTH, as expected, is lower compared to the other 2 classes, taking into account that it is formed by heterogeneous DA acts.

4 Analysis

4.1 Feature processing

Energy (E) and **Pitch (P)** are computed from the audio files with the help of *openSMILE* (feature extraction and audio analysis tool) by (Eyben and Schuller, 2015) while **Speech Rate (R)** is computed using time aligned transcripts.

Energy (E): One of the issue of telephonic conversation is the distance mouth-microphone that affects measured values of voice intensity. This adds noise to the distribution of values across speakers, even for the same speaker across different conversations. So to reduce this effect we introduce a normalization factor consisting in dividing each speaker E values by the average E produced by that speaker in the entire conversation. In addition, to reduce the environmental noise, we computed the average E using the temporal windows where the probability of voicing is above 0.65. Then we computed for each conversational unit (turn or Dialog act as provided by Switchboard transcripts) the average E.

Pitch (P): We computed the average in each conversational unit for each speaker.

Speech Rate (R): We used the approach proposed by Cohen-Priva (Cohen Priva et al., 2017) that defines R for an utterance as the ratio between the actual duration of the utterance and its expected duration (computed by estimating every word duration into the whole corpus, for all speakers). Values above / below 1 correspond respectively to fast / slow speech compare to the average of the corpus. In order to make the measure R more reliable we consider only utterances having more than 5 tokens.

4.2 Convergence

We divide each conversation into two halves and compare the distance between the average values of the target variables coming from each speaker. (Truong and Heylen, 2012; Edlund et al., 2009).

We computed the difference between the mean value of target variables for the two speakers in both halves. This provided us two values (first and second interval) for each variable and each conversation:

$$\Delta\overline{V}_i = \mid \overline{V}_{Ai} - \overline{V}_{Bi} \mid \quad (1)$$

, where $i = 1, 2$ refers respectively to the first and second interval, A and B indicate the speakers who take part in the conversation while V could be E (Energy), $F0$ (Pitch) and R (Speech rate). Our aim is to test the hypothesis that convergence, on such rather low-level variables, occurs during the interaction. We therefore computed the difference between both intervals, resulting in a distribution of these values in both intervals for the whole corpus. We then fitted a linear mixed regression model to this distribution to test if there is a significant difference across the intervals. Moreover, the sign of the estimate of the model provides us the direction of the evolution. We use the `lme4` library in `R` (Bates et al., 2014) to fit the models and provide t-values. The `lmerTest` package (Kuznetsova et al., 2014), which encapsulates `lme4`, was used to estimate degrees of freedom. (*Satterthwaite approximation*) and calculate p-values. In the model, the $\Delta\overline{V}_i$ is the predicted value, the A and B identities as well as the topic of the conversation are set

	All	St.	Opi.	Bc.
SWBD	**E**R	X	X	X
SW-DA-Man.	ER	ER	-	E
SW-DA-Auto.	X	-	-	E

Table 5: Difference-In-Difference Results Summary: E: Energy; P: Pitch ; R: Speech Rate; -: any significance; X: no performed experiments. normal font : p-value≤0.05 ; **bold** : p-value ≤ 0.01

as random intercepts. The model, in R notation, is $\Delta\overline{V}_i \sim t_i + (1 \mid topic) + (1 \mid speaker_A) + (1 \mid speaker_B)$.

We computed the mean target variables in the two intervals following the same method as for the whole dataset. Excluding conversations with undefined values (miss one or more average interval values), the number of conversations is 593 for *Statement*, 581 for *Backchannel* and 381 for *Opinion*.

5 Results

We report the results in the case of the whole dataset without DA (SWBD) and the DA manually tagged (SW-DA-Man.), the results for the Dialog acts grouping that refers to SW-DA-Man. and the dialog acts automatically tagged (SW-DA-Auto). The results are summarized in Tables 5 for Inter-Speaker correlations and 6 for convergence.

	All	St. + Opi.	Bc. +Agr.
SWBD	**E**R	X	X
SWBD-Auto	X	ER	E

Table 6: Difference-In-Difference Results Summary: E: Energy; P: Pitch ; R: Speech Rate; -: any significance; X: no performed experiments. normal font : p-value≤0.05 ; **bold** : p-value ≤ 0.01

5.1 No Dialog acts filtering (SWBD and SW-DA-Man.)

The use of the whole dataset without considering DA filtering is shown in Table 7. Convergence arises for Energy and Speech Rate. When using the DA-tagged part of Switchboard (642 conversations, corresponding to 41 hours when excluding silence and noise) as reported in table 8, there is a weaker effect both on E and R while F0 doesn't show any significant effect. The less strong significance compared to the SWBD case is expected, considering the huge data size reduction. For R instead, the effect seems to be stronger in this case despite the reduction of conversations.

SWBD	*Entire Corpus (180 hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.063**	**0.012**	$\mathbf{7 \times 10^{-7}}$
P-Mean	-0.044	0.021	0.490
R-Mean	**-0.049**	**0.024**	**0.046**

Table 7: Parameters our linear model for energy, pitch and speech rate for the raw corpus and for the manually tagged corpus. Speech rate was not considered for backchannel.

SW-DA-Man.	*Whole DA-tagged subset (41 Hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.054**	**0.021**	**0.026**
P-Mean	-0.057	0.040	0.158
R-Mean	**-0.106**	**0.047**	**0.026**

Table 8: Parameters our linear model for energy, pitch and speech rate for the raw corpus and for the manually tagged corpus. Speech rate was not considered for backchannel.

5.2 DA filtering (SW-DA-Man.)

The results for the categories statement, opinion and backchannel are reported in Table 9.

The *backchannel* subset exhibits convergence for E despite the huge data size reduction. However, no significant effects in F0 and R were found in category most likely because *backchannel* are made of very short utterances, including frequently one word and for which estimated duration is a problematic question.

Opinion DA-tag filtering seems to reduce too much both the data size as well as the samples number participating to the mean calculation for each halves and provides more noisy data in which significance cannot be attained.

The *statement* subset shows convergence for energy and speech rate. Considering just statement seems to clean the dataset by feedback-related differences as well as strong disfluencies (type *abandoned* in Switchboard). This helps observing the effect for speech rate. Contrarily, the wide variety of *statements* in terms of utterance duration could be an issue for pitch since contours and physiological-related decreasing slope could result in a lot of noise for this variable. Overall the comparison of the results of the non-filtered datasets and of the

SW-DA-Man.	*Backchannel (1 Hour)*		
Feature	Estimate	std	p-values
E-Mean	**-0.082**	**0.041**	**0.045**
P-Mean	0.043	0.022	0.491
SW-DA-Man.	*Statement (17 Hours)*		
Feature	Estimate	std	p-values
E-Mean	**-0.071**	**0.023**	**0.032**
P-Mean	-0.025	0.038	0.653
R-Mean	**-0.123**	**0.049**	**0.012**
SW-DA-Man.	*Opinion (7 Hours)*		
Feature	Estimate	std	p-values
E-Mean	-0.061	0.033	0.627
P-Mean	-0.032	0.053	0.552
R-Mean	-0.096	0.061	0.115

Table 9: Results for **SW-DA manually tagged**: Parameters our linear model for energy, pitch and speech rate for the raw corpus and for the manually tagged corpus. Speech rate was not considered for back-channels.

SW-DA-Auto	*Backchannel (3 Hours)*		
CLASS	Estimate	std	p-values
E-mean	**-0.079**	**0.031**	**0.009**
P-mean	0.003	0.002	0.815
SW-DA-Auto	*Statement (47 Hours)*		
CLASS	Estimate	std	p-values
E-mean	-0.040	0.025	0.2396
P-mean	0.009	0.004	0.611
R-mean	-0.004	0.004	0.332
SW-DA-Auto	*Opinion (19 Hours)*		
CLASS	Estimate	std	p-values
E-mean	0.011	0.048	0.811
P-mean	-0.014	0.032	0.663
R-mean	-0.008	0.009	0.369

Table 10: Results for **SW-DA automatically tagged**: Parameters of our linear model for energy, pitch and speech rate for the corpus automatically tagged considered statement, opinion and backchannel as DA classes. Speech rate was not considered for backchannel

filtered subsets suggests interesting patterns. For whole dataset, the data on which we compute the mean is much more dense and therefore the calculation robust, this seems to be enough to establish a trend for E and R. However, F0 doesn't show any significant effect for all the dataset (entire, subset and DA-acts of the Switchboard corpus). This probably can be explained considering that pitch is a more complex variable in the natural conversations framework and an average approach can't get the more strategic behavior of pitch ((Reichel et al., 2018)).

Backchannel is a strong filter (that controls for lexical content, duration and even the utterance function which is precisely delineated – the most frequent items are "Yeah" $= 35\%$, "Uh-huh" $= 18\%$, "Um-hum" $= 16\%$, "Right" $= 9\%$). However estimates for speech rate for *backchannel* is extremely problematic because of its linguistic form. In the case of *opinion* the strong reduction of data size is likely the cause to not find any effect. Finally, about *statements* it seems that filtering out very different utterances helps in bringing a more coherent dataset and therefore observing the convergence effect on E and R.

5.3 DA automatically tagged (SW-DA-Auto)

Here we report the results we obtained using the DA, grouped as statement, opinion and backchannel in Table 10 and the grouping of statement + opinion, backchannel + agreement in Table 11.

Comparing the results of the dataset automatically tagged and the subset with human annotation 5.2 we note that the only variable that shows convergence is energy in the case of *backchannel*. Even the number of conversations increases compared to the manual annotated dataset, the tagging stage could introduce noise derived by the turn that are non correctly labeled. As consequence this could have been affected the measure of distances between speakers.

Using indeed the classes derived by grouping similar classes (statement + opinion and backchannel + agreement) results in an emerging conver-

SWDA-Auto	*Backchannel + Agreement*		
CLASS	Estimate	std	p-values
E-mean	**-0.079**	**0.028**	**0.006**
P-mean	0.053	0.028	0.192
SWDA-Auto	*Statement + Opinion*		
CLASS	Estimate	std	p-values
E-mean	**-0.055**	**0.011**	$\mathbf{4 \cdot 10^{-6}}$
P-mean	-0.035	0.038	0.353
R-mean	**-0.075**	**0.021**	**0.008**

Table 11: Parameters of our linear model for energy, pitch and speech rate for the corpus automatically tagged considered statement + opinion and backchannel + agreement as DA classes. Speech rate was not considered for backchannel.

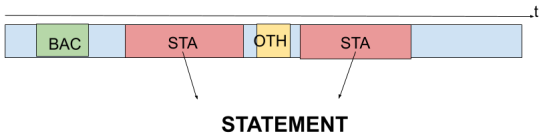


Figure 1: Each turn, following the provided segmentation by SW, is labeled considering the majority of dialogue acts that forms the turn.

gence effects for energy and speech rate. The reduction of distance between speakers in the second half of the conversation results in both the STA+OPI and BAC+AGR, showing that energy is the variable that mostly exhibits convergence. These results compared to the one of Table 10 could be partially explained by the fact that the performances of the tagger are better in this case. Increasing the precision reduces the noise of mislabeled turns that influence the filtering of DA.

6 Discussion

In this study we scrutinized *convergence* during the course of a conversation and in a real world setting (Switchboard corpus). Our work is a step toward using conversational corpus in more controlled studies on specific phenomena by using dialogue acts as a speech activity filter that reduces noise of real conversational dialogues and at the same time study convergence that happens within conversations (most of previous work establishes convergence by comparing speakers' distance in a conversation to other conversations and not what is happening in the time course of a conversation).
Our experiment consists in comparing the speakers difference between average values of a given variable in the first and second half per conversation, for the entire corpus and for the subsets made by DA annotations.

Our results show that speakers tend to have a more similar average values in the second half of the conversation for E and R. This confirms the results obtained on experimental lab speech and it is compatible with the results at corpus level.
The second idea we developed in this study is that dialog acts can have different behaviors in regard of convergence. We then split the whole datasets into sub set corresponding to specific frequent dialogue acts. Also in this case every significant or nearly significant difference correspond to a reduction of the distance between the speakers in the second part.

This result is interesting considering that the magnitude of the effect is still present even if in the dialog acts subsets the number of data to compute average values dramatically decreases (Table 1). The same trend is also found using a DA tagger of the turns produced by the speakers. Also in this case even if the automatic tagging introduce some noise the averages distance in second half of conversation decreases as shown in 11 and 10.

Our results complement and strengthen the picture provided by the literature. They also open up the possibility to a range of studies on large corpora, including new studies taking advantage on large corpora partially controlled *a posteriori* thanks to dialogue act tagging.

As for future work, we would like to confirm our results by replicating them on the larger Fisher corpus (Cieri et al., 2004). We also plan to tag the entire corpus in order to increase the statistical power as done for Switchboard. Finally, we would like to articulate out findings with the result on more local interpersonal dynamics such as *synchrony* (Levitan and Hirschberg, 2011).

References

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

Jeremy Auguste, Robin Perrotin, and Alexis Nasr. 2018. Annotation en actes de dialogue pour les conversations d'assistance en ligne. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, page 577.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7):1–23.

Francesca Bonin, Céline De Looze, Sucheta Ghosh, Emer Gilmartin, Carl Vogel, Anna Polychroniou, Hugues Salamin, Alessandro Vinciarelli, and Nick Campbell. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. In *Proceedings of 14th Annual Conference of the International Speech Communication Association*, Lyon, France.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus:

a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

U Cohen Priva and C Sanker. 2018. Distinct behaviors in convergence across measures. In *Proceedings of the 40th annual conference of the cognitive science society. Austin, TX*.

Uriel Cohen Priva, Lee Edelist, and Emily Gleason. 2017. Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, 141(5):2989–2996.

Jens Edlund, Mattias Heldner, and Julia Hirschberg. 2009. Pause and gap length in face-to-face interaction. In *Tenth Annual Conference of the International Speech Communication Association*.

Florian Eyben and Björn Schuller. 2015. opensmile:): the munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, 6(4):4–13.

Simone Fuscone, Benoit Favre, and Laurent Prevot. 2018. Replicating speech rate convergence experiments on the switchboard corpus. In *Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*.

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

Howard Giles and Peter Powesland. 1997. Accommodation theory. In *Sociolinguistics*, pages 232–239. Springer.

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*, volume 1, pages 517–520. IEEE.

A Kuznetsova, P Bruun Brockhoff, and R Haubo Bojesen Christensen. 2014. lmertest: tests for random and fixed effects for linear mixed effects models. *See https://CRAN. R-project. org/package= lmerTest*.

Rivka Levitan. 2014. *Acoustic-prosodic entrainment in human-human and human-computer dialogue*. Ph.D. thesis, Columbia University.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011*.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790.

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

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

Uwe D Reichel, Katalin Mády, and Jennifer Cole. 2018. Prosodic entrainment in dialog acts. *arXiv preprint arXiv:1810.12646*.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al. 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

Richard L. Street. 1984. Speech convergence and speech evaluation in fact-finding iinterviews. *Human Communication Research*, 11(2):139–169.

Khiet P Truong and Dirk Heylen. 2012. Measuring prosodic alignment in cooperative task-based conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Towards the score of communication

Andy Lücking
Université de Paris
Laboratoire de Linguistique Formelle
Goethe-Universität Frankfurt
Text Technology Lab
luecking@em.uni-frankfurt.de

Jonathan Ginzburg
Université de Paris
Laboratoire de Linguistique Formelle
Laboratoire d'Excellence LabEx-EFL
Institut Universitaire de France
yonatan.ginzburg@u-paris.fr

Abstract

The exchange of verbal and non-verbal communication signals in face-to-face dialogue is complexly organised in several ways: each contribution is produced and processed incrementally, contributions may be consecutive (e.g. question-answer pairs) or overlapping (e.g. backchannelling), and the contributions themselves may be multimodal. Contributions nonetheless exhibit pairwise utterance coherence, and in two respects: across tiers and across discourse co-texts. For these reasons, we propose to distribute dialogue agents across different tiers and to 'incrementalize' the sequential notion of turns according to the model of music-inspired communication scores.

1 Motivation

It is a truism that natural language communication is multimodal, or as we will also say, proceeds on different *tiers*.[1] What is meant by this slogan is that dialogue agents in addition to speech exchange a great variety of non-verbal communication means like manual gestures, facial expressions, gaze, laughter or suprasegmental signals (the latter two are vocal but not verbal). The chief significance of nonverbal communcation means usually resides in a *dialogue-oriented meaning* (Ginzburg and Poesio, 2016). For instance, backchannelling signals such as nodding or vocalisations such as 'mhm' influence the development of discourse (Bavelas et al., 2000). With *communication scores*, we aim to address these aspects of the fabric of communication by distributing dialogue agents across several communication tiers and allow tiers to be shared by different participants (i.e., giving up a strict sequential notion of turns). We hypothesise that the resulting (conceptual and formal) intricacies are regimented by a fundamental dialogical constraint of coherence.

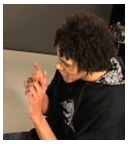
[1] We conceive a tier to be a layer of communication in the semiotic triad spanned by a transfer *medium*, an interpretive *code*, and a receiving sense *modality*. For instance, the phonetic events encoding natural language expressions, which are produced with the articulatory organ and spread through air and are perceived by the ears, constitute a tier, namely the vocal tier.

2 Some data

The most striking feature of multimodal discourse is the *binding problem* (Feldman, 2012). Interpretation is often guided by the heuristics 'if multiple signs occur simultaneously, take them as one' (Enfield, 2009, 9). For instance, a drawing gesture co-occurring with a shape description is understood as expressing a single *idea unit* (McNeill, 1992): *the house [has a RECtangular]* ⌊⌋ *shape* (here the gesture temporally overlaps with the underlined portion of speech, the stretch of the gesture's stroke is indicated by square brackets, the trajectory of the drawing gesture is sketched after the closing stroke bracket; capital letters indicate main stress). However, since simultaneity is not the only temporal pattern of multimodal communication, the simple heuristics has to qualified. A point in case is post-stroke holds, as argued by Rieser (2015) – we will come back to such issues shortly.

A multimodal turn grabbing is exemplified in (1), which stems from the (German) SaGA dialogue V4, at 8:39 (Lücking et al., 2010). Here, the route-giver R wants to continue her route-giving and produced the definite article *die* 'the [fem.]'. At this point, the follower F cuts in by index finger raising and exclaiming *Moment* 'wait' twice.

(1) R: *[die]*

F: [moment] moment 

The turn grab was multimodal, but a rectangular gesture would be inappropriate in this case, a point-

ing or a 'stop!' gesture seems to be called for, if any. This might indicate a multimodal interface of certain discourse-related expressions to interactive gestures as has been argued for wrt. demonstrative expressions and deixis (Frege, 1918) and spatial expressions and iconic (Schegloff, 1984) gestures, among others.

Boarding the turn of the speaker doesn't necessarily lead to a turn grab, rather, interlocutors can also produce *joint utterances* (Poncin and Rieser, 2006), undermining a strict notion of speaker and addressee even more than constant backchannelling.

With respect to the latter, an interesting asymmetry can be observed (one example is given in (35) below): it is possible that A is speaking and B is agreeing or disagreeing at the same time by nodding or head shaking, respectively. It is also possible that A is speaking and agreeing (with his-/herself) at the same time: namely when A speaks and nods. But it only seems to be possible that A is speaking and *dis*agreeing (shaking head) at the same time under special circumstances, for instance, if A's utterance contains a negative particle (*not*, *n't*, *no*) or is part of a *hostile* (Gregoromichelaki et al., 2011) move (i.a., A is not committed to the content of the utterance). We take up some examples in Sec. 5.

We hypothesise that one can make sense out of this by a (multimodal extension of a) fundamental dialogical constraint, namely *(pairwise) utterance coherence* (Ginzburg, 2012, Sec. 6.7.1). Co-occurring signals as well as consecutive utterances constitute a kind of adjacency pair in the sense that they can be embedded in a dialogue gameboard in a *relevant* way. This embedding has two aspects: (i) there is a grammar rule that licenses the combination of the signals, and (ii) the resulting dialogue gameboard can be connected to the actual one by means of one or more conversaitonal rules. In order to investigate the constraints that apply to largely simultaneous and multimodal interactions, one needs to 'compartmentalize' agents into articulators, as is outlined in the following.

3 Formal background

Communication scores (Sec. 4) rest on a couple of previous works which are briefly introduced subsequently.

3.1 TTR

Communication scores are formulated within *Type Theory with Records* (TTR, Cooper, 2005; Cooper and Ginzburg, 2015). TTR integrates logical techniques such as the lambda calculus and the expressiveness of feature-structure like objects (namely records and record types). A typing *judgement* $a : T$ is true iff object a is of type T. Types constructed from n-ary predicates ($n > 0$) are *dependent* on the values assigned to the *labels* that appear as arguments. Thus, if $a_1 : T_1$, $a_2 : T_2(a_1)$, ..., $a_n : T(a_1, a_2, \ldots, a_{n-1})$, then the record on the left in (2) is of the record type on the right in (2):

(2) $$\begin{bmatrix} l_1 = a_1 \\ \vdots \quad \vdots \\ l_n = a_n \end{bmatrix} : \begin{bmatrix} l_1 : T_1 \\ \vdots \quad \vdots \\ l_n : T(l_1, l_2, l_{n-1}) \end{bmatrix}$$

The notation $[l = a : T]$ represents a *manifest field* (Coquand et al., 2003). It is a notational convention for a *singleton type* T_a, where for any $b, b : T_a$ iff $b = a$.

Merge types correspond to unification in feature-structure formalisms. A merge '$\wedge$' is exemplified in (3):

(3) a. $A = \begin{bmatrix} l_1 : T_1 \\ l_2 : T_2(l_1) \end{bmatrix}$ and $B = \begin{bmatrix} l_3 : T_3 \end{bmatrix}$

b. $A \wedge B = \begin{bmatrix} l_1 : T_1 \\ l_2 : T_2(l_1) \\ l_3 : T_3 \end{bmatrix}$

For more (on) TTR see Cooper (2012).

3.2 Strings

Drawing on work of Fernando (2007, 2011), TTR comes with a string theory of events. For three events e_1, e_2 and e_3, the string $e_1e_2e_3$ represents a course of events, namely the succession of e_1, e_2 and e_3, in that order. The notation $e_1e_2e_3$ is an abbreviation for a time-indexed record:

(4) $\begin{bmatrix} t_0 = e_1 \\ t_1 = e_2 \\ t_3 = e_2 \end{bmatrix}$, where time indices t_i are in $\mathbb{N}$.

If $e_1 : T_1$, $e_2 : T_2$ and $e_3 : T3$, then $e_1e_2e_3 : T_1^\frown T_2^\frown T_3$ – the type constructor '$^\frown$' builds string types out of types. In order to exploit feature structure expressiveness in string types, a string of record types can be build by the same means, but is notationally enclosed in brackets. For instance, the first move of a chess player playing a Sicilian opening is represented in (5):

(5)

$$\begin{bmatrix} \mathrm{x} & : \mathit{Ind} \\ \mathrm{c}_x & : \text{chess-player}(\mathrm{x}) \\ \mathrm{y} & : \mathit{Ind} \\ \mathrm{c}_y & : \text{pawn}(\mathrm{y}) \\ l_1 & : \mathit{Loc} \\ \mathrm{c}_{l_1} & : \text{field-c7}(l_1) \\ l_2 & : \mathit{Loc} \\ \mathrm{c}_{l_2} & : \text{field-c5}(l_2) \\ \mathrm{c}_{on} & : \text{on}(\mathrm{y},l_1) \\ \mathrm{e} & : (\begin{bmatrix}\mathrm{e} : \text{grab}(\mathrm{x},\mathrm{y})\end{bmatrix} \frown \begin{bmatrix}\mathrm{e} : \text{move-to}(\mathrm{x},\mathrm{y},l_2)\end{bmatrix} \frown \begin{bmatrix}\mathrm{e} : \text{drop}(\mathrm{x},\mathrm{y})\end{bmatrix}) \end{bmatrix}$$

3.3 Incrementality

The feature-structure set-up of record types can be used in order to 'import' constraint-based grammars such as a *Head-driven Phrase Structure Grammar* (HPSG; Pollard and Sag, 1994). An emulation of HPSG within TTR is HPSG_{TTR} (Cooper, 2008; Ginzburg, 2012). Modelling the derivation of a constituent structure within a chart parser, an incremental version of HPSG_{TTR} can be obtained (Ginzburg et al., 2020). Incremental processing rests on perceptual classification: an acoustic event on the speech tier is classified as the realisation of the phonological part of a sign, that is, as an instance of a *lexical resource*. An example appropriate to Beethoven's anniversary year 2020 is given in (6), which draws on the NP format of Lücking and Ginzburg (2019):

(6) a. Lex('Beethoven', NP)

b. $$\begin{bmatrix} \text{s-event} : \begin{bmatrix} \mathrm{e} & : \texttt{beethoven} \\ \text{spkr} & : \mathit{Ind} \\ \text{addr} & : \mathit{Ind} \\ \mathrm{c}_{\text{sp}} & : \text{addressing}(\text{spkr},\text{addr},\mathrm{e}) \end{bmatrix} \\ \text{phon} : /\text{Beethoven}/ \\ \text{syn} : \begin{bmatrix} \text{cat=np} & : \mathit{Cat} \\ \text{dtrs=}\langle\,\rangle & : \text{list}(\mathit{Sign}) \end{bmatrix} \\ \text{q-params} : \begin{bmatrix} \text{refind} & : \mathit{Ind} \\ \mathrm{c}_{\text{nm}} & : \text{named}(\text{refind}, \text{'Beethoven'}) \end{bmatrix} \\ \text{cont=q-params.refind} : \mathit{Ind} \end{bmatrix}$$

We can proceed from a speech event to the associated sign information by dint of the phonological classification constraint in (7) (slightly modified from Ginzburg et al., 2020):

(7) If Lex(T, C) is a lexical resource, then for any speech event u such that $u : T$, it is licensed to merge resource and speech event: Lex(T, C) $\wedge$ [s-event : [e=u : T]].

An example of an outcome of this sign classification is shown in (6b).

Since 'Beethoven' is a proper name (a full NP), it raises the expectation that a VP will follow. This expectation is backed by grammar which supplies a syntactic resource in form of the rule 'S $\rightarrow$ NP VP'. String types (cf. Sec. 3.2) tie up these things as follows. The acoustic speech event 'Beethoven' gives rise to the following initialisation of a chart:

(8) $$\begin{bmatrix} \mathrm{e}_1 = \texttt{beethoven} : \mathit{Phon} \\ \mathrm{e} \ : \ (\begin{bmatrix}\mathrm{e}_1 : \text{start}(\mathrm{e}_1)\end{bmatrix} \frown \begin{bmatrix}\mathrm{e}_1 : \text{end}(\mathrm{e}_1)\end{bmatrix}) \end{bmatrix}$$

The start-end string in (8) corresponds to an edge in chart parsing (Earley, 1970), see also Fig. 1. Applying lexical resource classification adds sign information:

(9)

$$\begin{bmatrix} \mathrm{e}_1 = \texttt{beethoven} : \mathit{Phon} \\ \mathrm{e}_2 : \ \text{Lex('Beethoven', NP)} \wedge \begin{bmatrix}\text{s-ev} : \begin{bmatrix}\mathrm{e}=\mathrm{e}_1 : /\text{Beethoven}/\end{bmatrix}\end{bmatrix} \\ \mathrm{e} \ : \ (\begin{bmatrix}\mathrm{e}_1 : \text{start}(\mathrm{e}_1) \\ \mathrm{e}_2 : \text{start}(\mathrm{e}_2)\end{bmatrix} \frown \begin{bmatrix}\mathrm{e}_1 : \text{end}(\mathrm{e}_1) \\ \mathrm{e}_2 : \text{end}(\mathrm{e}_2)\end{bmatrix}) \end{bmatrix}$$

Syntactic resources allow us to project hypotheses of possible continuations (what is required (req) for a rule to apply), given what has been found (fnd). Such a move has been been developed for incremental processing in *Dynamic Syntax* (Gregoromichelaki et al., 2011). However, the closest precursor is Poesio and Traum (1997), where micro-conversational events generate (in the sense of Goldman, 1970) conversational moves. We adopt such an approach here to the HPSG_{TTR} format, following Ginzburg (2012, Sec. 8) and Ginzburg et al. (2018):

(10) $$\begin{bmatrix} \mathrm{e}_1 = \texttt{beethoven} : \mathit{Phon} \\ \mathrm{e}_2 : \ \text{Lex('Beethoven', NP)} \wedge \begin{bmatrix}\text{s-ev} : \begin{bmatrix}\mathrm{e}=\mathrm{e}_1 : /\text{Beethoven}/\end{bmatrix}\end{bmatrix} \\ \mathrm{e}_3 : \ (\begin{bmatrix}\text{rule=S}\rightarrow\text{NP VP} : \text{NP}\frown\text{VP fnd=}\mathrm{e}_2 : \mathit{Sign} \\ \text{req=VP} : \mathit{Sign} \\ \mathrm{e} : \text{required(req,rule)}\end{bmatrix}) \\ \mathrm{e} \ : \ (\begin{bmatrix}\mathrm{e}_1 : \text{start}(\mathrm{e}_1) \\ \mathrm{e}_2 : \text{start}(\mathrm{e}_2)\end{bmatrix} \frown \begin{bmatrix}\mathrm{e}_1 : \text{end}(\mathrm{e}_1) \\ \mathrm{e}_2 : \text{end}(\mathrm{e}_2) \\ \mathrm{e}_3 : \text{start}(\mathrm{e}_3)\end{bmatrix} \frown \begin{bmatrix}\mathrm{e}_3 : \text{end}(\mathrm{e}_3)\end{bmatrix}) \end{bmatrix}$$

If the string chart processes an input of the category marked as required, a sentential parse will be achieved:

(11)

$$\begin{bmatrix} e_1 = \texttt{beethoven} : \mathit{Phon} \\ e_2 : \text{ Lex('Beethoven', NP)} \wedge \begin{bmatrix} \text{s-ev} : \begin{bmatrix} e{=}e_1 : /\text{Beethoven}/ \end{bmatrix} \end{bmatrix} \\ e_3 : \left(\begin{bmatrix} \text{rule=S}\rightarrow\text{NP VP} : \text{NP}^\frown\text{VP} \\ \text{fnd}{=}e_2 : \mathit{Sign} \\ \text{req=VP} : \mathit{Sign} \\ \text{e : required(req,rule)} \end{bmatrix} \frown \begin{bmatrix} \text{fnd}{=}e_5 : \mathit{Sign} \\ \text{req}{=}\varepsilon : \mathit{Sign} \\ \text{e : complete(rule)} \end{bmatrix} \right) \\ e_4 = \texttt{rocks} : \mathit{Phon} \\ e_5 : \text{ Lex('rock', VP)} \wedge \begin{bmatrix} \text{s-event} : \begin{bmatrix} e{=}e_4 : /\text{rocks}/ \end{bmatrix} \end{bmatrix} \\ e \ : \left(\begin{bmatrix} e_1 : \text{start}(e_1) \\ e_2 : \text{start}(e_2) \end{bmatrix} \frown \begin{bmatrix} e_1 : \text{end}(e_1) \\ e_2 : \text{end}(e_2) \\ e_3 : \text{start}(e_3) \\ e_4 : \text{start}(e_4) \\ e_5 : \text{start}(e_5) \end{bmatrix} \frown \begin{bmatrix} e_3 : \text{end}(e_3) \\ e_4 : \text{end}(e_4) \\ e_5 : \text{end}(e_5) \end{bmatrix} \right) \end{bmatrix}$$

Semantic composition proceeds in parallel to chart construction as specified in the lexical resources and the syntactic rules. The result of a chart parse is a HSPG_{TTR} sign defined by the (completed) rule that covers the whole speech event. The chart in (11), for instance, derives the sentential (i.e. cat=vp and dtrs=$\langle\rangle$) sign given in (12), where "cont" labels a structure of type *Prop*(osition), the type of an Austinian (Austin, 1950) proposition, pairing a situation (record) and a situation type (record type):

(12)

$$\begin{bmatrix} \text{phon} : /\text{Beethoven}/^\frown/\text{rocks}/ \\ \text{syn} : \begin{bmatrix} \text{cat=vp} \ : \mathit{Cat} \\ \text{dtrs}{=}\langle\ \rangle : \text{list}(\mathit{Sign}) \end{bmatrix} \\ \text{q-params} : \begin{bmatrix} \text{refind} : \mathit{Ind} \\ c_{\text{nm}} : \text{named(refind, 'Beethoven')} \end{bmatrix} \\ \text{cont}{=} \begin{bmatrix} \text{sit} = s_0 : \mathit{Rec} \\ \text{sit-type} = \begin{bmatrix} \text{nucl} : \text{run(q-params.refind)} \end{bmatrix} : \mathit{RecType} \end{bmatrix} \end{bmatrix}$$

We will refer to the maximal sign processed by a chart for a speech event *e* as *Sign*(*e*). Since *e* can be either a simple acoustic event (corresponding to a lexical element), or a string of acoustic events, two cases have to be distinguished:

(13) Sign(e)

a. $$\begin{bmatrix} \text{sit=s-event} : \begin{bmatrix} e : \mathit{Phon} \end{bmatrix} \\ \text{sit-type} : \text{Lex}(T, \mathit{Sign}) \wedge \text{sit.e} : T \end{bmatrix}$$

b. $$\begin{bmatrix} \text{sit=s-event} : \begin{bmatrix} e : (\mathit{Phon}^\frown\mathit{Phon})^+ \end{bmatrix} \\ \text{sit-type} : \mathit{Sign} \wedge \begin{bmatrix} \text{syn} : \begin{bmatrix} \text{dtrs=e} : \mathit{String}(\mathit{Sign}) \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Furthermore, given the mechanism of rule projection also an anticipatory version of Sign(e) can be defined, *ProjSign(e)* ('projected sign'). The processing and production of natural language sentences in interaction is highly anticipatory, as is evinced by timing relations from turn-taking (Levinson and Torreira, 2015).

Given a non-sentential utterance *u* processed so far, and the series of "req" rules σ leading from the syntactic parse of *u* to S, then

(14) ProjSign(e) is

$$\begin{bmatrix} \text{sit=s-event} : \begin{bmatrix} u : \text{Phon} \end{bmatrix} \\ \text{sit-type} : \text{Sign} \wedge \begin{bmatrix} \text{syn} : \begin{bmatrix} \text{dtrs}{=}\sigma : \mathit{Sign}(\mathit{Sign}) \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

ProjSign(e) can further be refined in terms of the number, depth and probabilities of the syntax rules involved. The latter can be derived from *data-oriented parsing* (Bod and Scha, 1996) and implemtented in a probabilistic version of TTR (Cooper et al., 2015). We leave such refinements and their empirical testing to future work.

We assume that ProjSign(e) is part of a dialogue agent's private share of its information state. ProjSign(e) is constantly compared to Sign(e) from the public information state via monitoring processes. ProjSign(e) can lead to turn-overlapping backhchannelling and joint utterances, among others; or to anticipatory errors, in case of which a special kind of correction will occur (*mhm, mhm . . . Oh wait! No! That's not what I expected*). So far, ProjSign(e) is an 'experimental feature', but one that is needed not least for accounting for some puzzling, short timing relations and joint utterances. In any case, Sign(e) and ProjSign(e) connects incremental charts to information states and locutionary propositions utilized in KoS, where it can also lead to 'turn projection', as argued by Ginzburg et al. (2018) with regard to forward-looking disfluencies.

3.4 KoS

Within language use, the signs of a natural language are part of 'mechanisms of interaction' (Kempson et al., 2016). In order to analyze natural language interactions, we make use of *dialogue gameboards* (Ginzburg, 1994). A dialogue game board (DGB) is an information-state based structure for modeling *communicative interactions*. The DGB from KoS tracks the interlocutors (*spkr* and *addr* fields), a record of the dialogue history (*Moves*), dialogue moves that are in the process of grounding (*Pending*), the question(s) currently under discussion (*QUD*), the assumptions shared among the interlocutors (*Facts*) and the dialogue participant's

view of the visual situation and attended entities (*VisualSit*). The TTR representation of a DGB is given in (15), where *LocProp* is the type of a *locutionary proposition* (see (16) below) and *poset* abbreviates 'partially ordered set'.

(15)
$$\begin{bmatrix} \text{spkr} & : \mathit{Ind} \\ \text{addr} & : \mathit{Ind} \\ \text{utt-time} & : \mathit{Time} \\ \text{c-utt} & : \text{addressing(spkr, addr, utt-time)} \\ \text{facts} & : \text{set}(\mathit{Prop}) \\ \text{visualsit} & : \mathit{RecType} \\ \text{pending} & : \text{list}(\mathit{LocProp}) \\ \text{moves} & : \text{list}(\mathit{LocProp}) \\ \text{qud} & : \text{poset(Question)} \end{bmatrix}$$

A special kind of propositions are *locutionary propositions* (*LocProp*, Ginzburg, 2012, 172):

(16)
$$\mathit{LocProp} =_{def} \begin{bmatrix} \text{sit} & : \mathit{Sign} \\ \text{sit-type} & : \mathit{RecType} \end{bmatrix}$$

Locutionary propositions are the link between dialogue gameboards and incremental processing. In terms of chart parsing (Sec. 3.3), a *LocProp* is the classification of an acoustic speech event by means of a sign type. That is, *LocProp* corresponds to Sign(e) from (13). Locutionary propositions are sign objects required to explicate clarification potential and grounding (Ginzburg, 2012).

Dialogue dynamics is regimented by *conversational rules*. A conversational rule is specified in terms of its preconditions (preconds) and its effects. A simple example is free speech, where, given an empty QUD list, any dialogue participant can make a contribution, that is turn(holder)Underspec(ified) (Ginzburg, 2012):

(17)
$$\begin{bmatrix} \text{preconds} : \begin{bmatrix} \text{qud}=\langle\rangle : \text{poset}(\mathit{Question}) \end{bmatrix} \\ \text{effects} \quad : \text{turnUnderspec} \wedge \\ \qquad \begin{bmatrix} \text{a} : \mathit{Prop} \\ \text{R} : \mathit{IllocRel} \\ \text{moves.latest=R(spkr,addr,a)} : \mathit{IllocProp} \end{bmatrix} \end{bmatrix}$$

We will make use of free speech in the analysis of (35) in Fig. 2 below.

3.5 Multimodal charts

In order to process input on several tiers, multimodal chart parsing has been devised as an extension of unification-based grammar parsing (Johnston, 1998). A graphical illustration is given in Fig. 1, where speech and gesture input can be parsed into several *multicharts*, namely $\{(s,0,1),(g,3,4)\}$, $\{(s,1,2),(g,3,4)\}$, or $\{(s,0,2),(g,3,4)\}$.

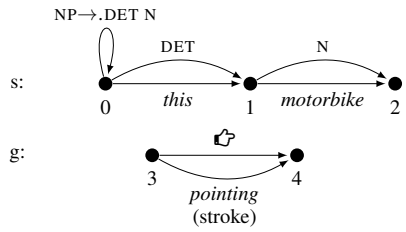


Figure 1: Multimodal chart parser

Within HPSG grammars, speech-gesture integration rules have been defined based on multichart parsing (Alahverdzhieva et al., 2017; Lücking, 2016).[2] In these approaches, the range of possible multicharts is further constrained by temporal and prosodic information. The original temporal constraint says that 'the time of the speech [...] must either overlap with or start within 4 seconds of the time of the gesture' (Johnston, 1998, 627). Since such temporal constraints are always a bit arbitrary, we propose a communication-based constraint instead. To this end we make use of the (still time-based) account of speech-gesture of Lücking (2013), respectively its HPSG_{TTR} reformulation (Lücking, 2016), since this account follows a perceptual classification approach as already used for speech interpretation (cf. Subsec. 3.3). Speech-gesture integration on this account is modelled in terms of a *speech-gesture ensemble* (Kendon, 2004), where a gesture (G-DTR) attaches to a phonetically marked *affiliate* (AFF; Schegloff, 1984) from speech (S-DTR), which is required to exhibit a feature called "CVM".

(18)
$$\mathit{sg\text{-}ensemble}$$
$$\begin{bmatrix} \text{phon=s-dtr.phon} : \mathit{Phon} \\ \text{cat=s-dtr.cat} : \mathit{SynCat} \\ \text{cont=g-dtr.traj} \wedge \text{s-dtr.cont.cvm} \end{bmatrix}$$
$$\mathit{s\text{-}dtr} \qquad \mathit{g\text{-}dtr}$$
$$\begin{bmatrix} \text{phon.accent} : \mathit{Marked} \\ \text{cat} : \mathit{SynCat} \\ \text{cont} : \mathit{SemObj} \end{bmatrix} \qquad \begin{bmatrix} \text{aff=s-dtr} : \mathit{Sign} \\ \text{traj} : \mathit{Vec} \end{bmatrix}$$

The underlying rationale of (18) is that a gesture movement is a trajectory that is mathematically described as a sequence of vectors in three dimensions ($\mathbb{R}^3$; or $\mathbb{R}^4$ if the temporal dimension is explicitly built in). This gesture vector (hence

[2] For an approach to coverbal gesture integration based on rhetorical relations see Lascarides and Stone (2009). However, since this account presupposes a grammatical affiliation, it probably rests on a (variant of an) approach sketched here.

type *gesture-vec*) *exemplifies* a predicate from the restrictor list of the affiliated verbal expression by unifying into its *conceptual vector meaning* (CVM), which in turn is an abstract, vector-based representation of shapes, movements, orientations, or object axis, spelled out within the vector space algebra of Zwarts (2003). A translation procedure from gesture representation onto vector representations (and a HPSG_{TTR} version) is given in Lücking (2016). But what is a gesture representation? Drawing on work in gesture annotation, gestures are represented in terms of their kinematic features, giving rise to a 'phonetic' gesture representation. Within the spatial reference system of the *gesture space* (McNeill, 1992), the movements of the wrists and orientations of palms and backs of each hand are coded. Additional information concern the handshape, spatio-temporal extents, and, in case of a bimanual gestures, the relation between both hands (cf. Martell et al., 2002; Lücking et al., 2010).

Gestures are functionally distinguished in terms of the *representation technique* they perform (Müller, 1998; Streeck, 2008). A representation technique is a culture-based practise such as *drawing*, *sculpturing*, *modelling*, or *placing*. In analogy to lexical resources from incremental speech processing (Sec. 3.3), we construct *gesture resources* from kinematic manual gesture representations. An example resource for a drawing gesture is given in (19a), which expands to the structure in (19b). Note that the gesture resource in (19a) is *not* an emblem (i.e., a lexicalized form–meaning pair like *thumbs-up*). A drawing gesture does not have a fixed form side. This is more obvious in the CONT feature in (19b), where the movement features of the gesture token gets 'vectorized' – the iconic aspect of a drawing gesture. The content of a drawing gesture then is that the vector sequence described by the gesture's wrist trajectory (Vec(carrier.wrist)) depicts the SHAPE attribute of the affiliated expression.

(19) a. $$\text{Gest}\left(\begin{bmatrix} \text{hand} & : \left\{\text{left, right}\right\} \\ \text{hs} & : \textit{hShape} \\ \text{carrier} & : \begin{bmatrix} \text{wrist} : \textit{Move} \\ \text{palm} : \textit{Dir} \\ \text{boh} : \textit{Dir} \\ \text{mov} : \textit{Conc} \end{bmatrix} \\ \text{sync} & : \begin{bmatrix} \text{s-loc} : \textit{gSpace} \\ \text{e-loc} : \textit{gSpace} \end{bmatrix} \\ \text{rel} & : \textit{gRel} \end{bmatrix}, \text{drawing}\right)$$

b. $$\begin{bmatrix} \text{g-event} : \begin{bmatrix} \text{e} : \begin{bmatrix} \text{hand} : \left\{\text{left, right}\right\} \\ \text{hs} : \textit{hShape} \\ \text{carrier} : \begin{bmatrix} \text{wrist} : \textit{Move} \\ \text{palm} : \textit{Dir} \\ \text{boh} : \textit{Dir} \\ \text{mov} : \textit{Conc} \end{bmatrix} \\ \text{sync} : \begin{bmatrix} \text{s-loc} : \textit{gSpace} \\ \text{e-loc} : \textit{gSpace} \end{bmatrix} \\ \text{rel} : \textit{gRel} \end{bmatrix} \\ \text{spkr} : \textit{Ind} \\ \text{addr} : \textit{Ind} \\ \text{c}_{\text{sp}} : \text{addressing(spkr,addr,e)} \end{bmatrix} \\ \text{mode} : \left[\text{act=draw} : \textit{gMode}\right] \\ \text{cont=Vec(carrier.wrist)} \wedge \\ \quad \text{aff.cont.cvm.shape} : \textit{Exemplification} \end{bmatrix}$$

For example, moving the wrist rightwards, back (i.e., towards the body of the gesturer), and leftwards in a rectangular manner ('line') – $\begin{bmatrix} \text{path} : \textit{line} \\ \text{wrist=mr}^\frown\text{mb}^\frown\text{ml} : \textit{Move} \end{bmatrix}$ – a U-shaped trajectory '⊔' is displayed. Note that an iconic gesture does not receive a direct translation into semantic predicates of the underlying formal semantics framework; it can *exemplify* such predicates, though. One could speculate that this difference underlies the non-at-issue status of most co-speech gestures observed by Ebert (2014).

A standard gesture movement is carried out in three *phases*, namely a *preparation phase* (where the hand is moved out of a rest position into gesture space), a *stroke* (where the gesture performs its actual meaningful trajectory), and a *retraction phase* (where the hand is moved from the stroke back into a rest position). The stroke may be 'frozen', amounting to a *post-stroke hold*. Such holds are particularly difficult for semantic modelling since they preserve the stroke's meaning for later uptake while speech already progresses (Rieser, 2015). However, part of such difficulties is a time-based notion of affiliation. We propose to replace Johnston's (1998) temporal constraint on speech-gesture ensembles by a relative one along the following lines:

(20) **Speech-gesture integration within multicharts**

a. *Stroke*: a gesture's stroke attaches to the closest verbal affiliate candidate;

b. *Post-stroke hold*: during a post-stroke hold, a gesture attaches to any closest affiliate candidate.

An affiliate candidate of course has to fulfil phonetic and CVM requirements in addition. Note that an affiliate which constitutes an ensemble with the stroke phase of a gesture is not available to the hold phase any more. Note further that the relative constraint has a testable consequence: if a possible affiliate occurs in between a gesture and its actually 'intended' affiliate, then the possible one becomes the actual one instead – the *no intervening communication event hypothesis*. By this measures any of the following patterns are captured (assuming that the shape-related predicate *rectangular* provides the CVM interface):

(21) a. A: ⌊⌋ *the house has a RECtangular shape*
(gesture before speech)

b. A: *the house has a RECtangular shape* ⌊⌋
(speech before gesture)

c. *A: the house has a RECtangular shape*
⌊⌋
(speech overlaps gesture)

d. *A: the house has a RECtangular shape*
B: ⌊⌋
(agent crossing speech–gesture integration)

However, the rectangular speech–gesture ensemble can be broken if another affiliate candidate intervenes, as in (22):

(22) *the house has a RECt- no circular shape*
⌊⌋

Since (22) involves a self-correction, a proper analysis would have much more to tell; however, corrections are not the topic of this paper and the examples hopefully already suffice to illustrate the relative patterning of speech and gesture integration within multimodal charts.

Holds (whose duration is indicated by asterisks) can affiliate to candidates not yet occupied by the stroke phase, which accounts, for instance, for verbal repetitions. In (23), the hold can be attached to the repeated attribution 'rectangular shape':

(23) *the house has a RECtangular*
⌊⌋*********
shape – rectangular shape
****

Given the rule of thumb that the preparation phase precedes the verbal affiliate of the stroke (McNeill, 1992), observing a preparation in multimodal parsing provides information for projecting an hypothesis of a multimodal integration rule – for instance, expecting a *sg-ensemble* headed by an NP in speech instead of a speech-only NP rule. Take, for instance, the chart parse of the example from Sec. 1: *the house [has a RECtangular]* ⌊⌋ *shape*. The string chart in (24) represents the state after having processed *the house has* and the gesture's preparation phase. Due to this input, a VP rule (e_9) and a gesture integration rule (e_{10}) have been triggered, but are still pending:

(24)

$$\begin{bmatrix}
e_1 & = \texttt{the} : \textit{Phon} \\
e_2 & : \text{Lex('the', DET)} \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e=e}_1 : \text{/the/}\end{bmatrix}\end{bmatrix} \\
e_3 & : (\begin{bmatrix}\text{rule=NP}\rightarrow\text{DET N} : \text{DET}^\frown\text{N} \\ \text{fnd=e}_2 : \textit{Sign}\end{bmatrix} \frown \begin{bmatrix}\text{fnd=e}_5 : \textit{Sign}\end{bmatrix}) \\
e_4 & = \texttt{house} : \textit{Phon} \\
e_5 & : \text{Lex('house', N)} \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e=e}_4 : \text{/house/}\end{bmatrix}\end{bmatrix} \\
e_6 & = \texttt{prep} : \textit{Phase} \\
e_7 & = \texttt{has} : \textit{Phon} \\
e_8 & : \text{Lex('have', V)} \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e=e}_7 : \text{/has/}\end{bmatrix}\end{bmatrix} \\
e_9 & : (\begin{bmatrix}\text{rule=VP}\rightarrow\text{V NP} \\ \text{fnd=e}_8 : \textit{Sign} \\ \text{req=NP} : \textit{Sign} \\ \text{e} : \text{required(req,rule)}\end{bmatrix}) \\
e_{10} & : (\begin{bmatrix}\text{rule=sg-ensemble}\rightarrow\text{X[accent,cvm] stroke} \\ \text{fnd=e}_6 : \textit{Phase} \\ \text{req1=stroke} : \textit{Phase} \\ \text{req2=X[accent,cvm]} : \textit{Sign} \\ \text{e} : \text{required(req1,req2,rule)}\end{bmatrix}) \\
e & : (\begin{bmatrix}e_1 : \text{start}(e_1) \\ e_2 : \text{start}(e_2)\end{bmatrix} \frown \begin{bmatrix}e_1 : \text{end}(e_1) \\ e_2 : \text{end}(e_2) \\ e_3 : \text{start}(e_3) \\ e_4 : \text{start}(e_4) \\ e_5 : \text{start}(e_5) \\ e_6 : \text{start}(e_6)\end{bmatrix} \frown \begin{bmatrix}e_3 : \text{end}(e_3) \\ e_4 : \text{end}(e_4) \\ e_5 : \text{end}(e_5) \\ e_6 : \text{end}(e_6) \\ e_7 : \text{start}(e_7) \\ e_8 : \text{start}(e_8) \\ e_9 : \text{start}(e_9) \\ e_{10} : \text{start}(e_{10})\end{bmatrix} \\
& \quad \frown \begin{bmatrix}e_7 : \text{end}(e_7) \\ e_8 : \text{end}(e_8)\end{bmatrix})
\end{bmatrix}$$

In the next steps, the indefinite article is processed and triggers an NP rule (we just consider a single NP hypothesis in e_{15} below). The gesture stroke is detected (e_{11}). The adjective *rectangular* is processed (e_{16}, e_{17}), which carries accent and CVM information required to complete the gesture integration rule (e_{10}). The string chart after these steps is as follows:

(25)

$$
\begin{bmatrix}
e_1 & = \texttt{the} : \mathit{Phon} \\
e_2 & : \text{Lex}(\text{‘the’, DET}) \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e}=e_1 : \text{/the/}\end{bmatrix}\end{bmatrix} \\
e_3 & : (\begin{bmatrix}\text{rule=NP}\rightarrow\text{DET N : DET}^\frown\text{N} \\ \text{fnd}=e_2 : \mathit{Sign}\end{bmatrix} \frown \begin{bmatrix}\text{fnd}=e_5 : \mathit{Sign}\end{bmatrix}) \\
e_4 & = \texttt{house} : \mathit{Phon} \\
e_5 & : \text{Lex}(\text{‘house’, N}) \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e}=e_4 : \text{/house/}\end{bmatrix}\end{bmatrix} \\
e_6 & = \texttt{prep} : \mathit{Phase} \\
e_7 & = \texttt{has} : \mathit{Phon} \\
e_8 & : \text{Lex}(\text{‘have’, V}) \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e}=e_7 : \text{/has/}\end{bmatrix}\end{bmatrix} \\
e_9 & : (\begin{bmatrix}\text{rule=VP}\rightarrow\text{V NP} \\ \text{fnd}=e_8 : \mathit{Sign} \\ \text{req=NP} : \mathit{Sign} \\ \text{e} : \text{required(req,rule)}\end{bmatrix}) \\
e_{10} & : (\begin{bmatrix}\text{rule=sg-ensemble}\rightarrow\text{X[accent,cvm] stroke} \\ \text{fnd}=e_6 : \mathit{Phase} \\ \text{req1=stroke} : \mathit{Phase} \\ \text{req2=X[accent,cvm]} : \mathit{Sign} \\ \text{e} : \text{required(req1,req2,rule)}\end{bmatrix} \\
& \quad \frown \begin{bmatrix}\text{fnd1}=e_{11} : \mathit{Phase} \\ \text{fnd2}=e_{17} : \mathit{Sign}\end{bmatrix}) \\
e_{11} & = \begin{bmatrix}\text{carrier}=\begin{bmatrix}\text{path} : \mathit{line} \\ \text{wrist=mr}^\frown\text{mb}^\frown\text{ml} : \mathit{Move}\end{bmatrix})\end{bmatrix} : \mathit{Stroke} \\
e_{12} & : \text{Gest(Vec(carrier),draw)} \wedge \begin{bmatrix}\text{g-ev} : \begin{bmatrix}\text{e}=e_{11} : \text{‘}\blacksquare\text{’}\end{bmatrix}\end{bmatrix} \\
e_{13} & = \texttt{a} : \mathit{Phon} \\
e_{14} & : \text{Lex}(\text{‘a’, DET}) \wedge \begin{bmatrix}\text{s-event} : \begin{bmatrix}\text{e}=e_{13} : \text{/a/}\end{bmatrix}\end{bmatrix} \\
e_{15} & : (\begin{bmatrix}\text{rule=NP}\rightarrow\text{DET A N:DET}^\frown\text{A}^\frown\text{N} \\ \text{fnd}=e_{14}\end{bmatrix} \\
& \quad \frown \begin{bmatrix}\text{fnd}=e_{17} : \mathit{Sign} \\ \text{req=N} : \mathit{Sign} \\ \text{e} : \text{required(req,rule)}\end{bmatrix}) \\
e_{16} & = \texttt{rectangular} : \mathit{Sign} \\
e_{17} & : \text{Lex}(\text{‘rectangular’, A}) \wedge \begin{bmatrix}\text{s-ev} : \begin{bmatrix}\text{e}=e_{16} : \text{/rectangular/}\end{bmatrix}\end{bmatrix} \\
e & : (\begin{bmatrix}e_1 : \text{start}(e_1) \\ e_2 : \text{start}(e_2)\end{bmatrix} \frown \begin{bmatrix}e_1 : \text{end}(e_1) \\ e_2 : \text{end}(e_2) \\ e_3 : \text{start}(e_3) \\ e_4 : \text{start}(e_4) \\ e_5 : \text{start}(e_5) \\ e_6 : \text{start}(e_6)\end{bmatrix} \frown \begin{bmatrix}e_3 : \text{end}(e_3) \\ e_4 : \text{end}(e_4) \\ e_5 : \text{end}(e_5) \\ e_6 : \text{end}(e_6) \\ e_7 : \text{start}(e_7) \\ e_8 : \text{start}(e_8) \\ e_9 : \text{start}(e_9) \\ e_{10} : \text{start}(e_{10}) \\ e_{11} : \text{start}(e_{11}) \\ e_{12} : \text{start}(e_{12})\end{bmatrix} \\
& \quad \frown \begin{bmatrix}e_7 : \text{end}(e_7) \\ e_8 : \text{end}(e_8) \\ e_{13} : \text{start}(ee_{13}) \\ e_{14} : \text{start}(e_{14})\end{bmatrix} \frown \begin{bmatrix}e_{13} : \text{end}(e_{13}) \\ e_{14} : \text{end}(e_{14}) \\ e_{15} : \text{start}(e_{15}) \\ e_{16} : \text{start}(e_{16}) \\ e_{17} : \text{start}(e_{17})\end{bmatrix} \frown \begin{bmatrix}e_{11} : \text{end}(e_{11}) \\ e_{12} : \text{end}(e_{12}) \\ e_{16} : \text{end}(e_{16}) \\ e_{17} : \text{end}(e_{17}) \\ e_{10} : \text{end}(e_{10})\end{bmatrix})
\end{bmatrix}
$$

The rules e_{15} and e_9 are still open and will be completed once the object noun is processed: this completes e_{15} and gives rise to an NP which in turn can take the ‘req’ slot of the VP rule in e_9. Along these lines, ultimately also the sentence hypothesis is confirmed (as are a couple of projective signs in between).

Along with the chart growing, the cont(ent) fields of the HPSG_{TTR} signs gets compositionally computed. To this end, a combination of functional application and unification (i.e. *merge* in TTR) is employed. An intransitive verb, for instance, is represented by means of a lambda abstract over an argument, as usual. The argument slot is eventually filled by a merge operation when the verb combines with the subject noun. By this means, ‘the monitoring and update/clarification cycle is modified to happen *at the end of each word utterance event*’ (emphasis in original, Ginzburg et al., 2018, 470). Such a move is necessary in order to analyse intra-sentential acknowledgements (Poesio and Traum, 1997):

(26) *A: The house has a rectangular shape*
B: *mhm*

The noun phrase utterance of A gives, via chart parsing, rise to the following NP semantics (where the sit-type value is of type *RecType*):

(27)

dgb.pending =

$$
\begin{bmatrix}
\text{s}=u_1 : \mathit{Sit} \\
\text{sit-type}=\begin{bmatrix}
\text{dgb-params} : \begin{bmatrix}\text{refind} : \mathit{Ind} \\ \text{refset} : \mathit{Set}(\mathit{Ind}) \\ \text{compset} : \mathit{Set}(\mathit{Ind}) \\ \text{maxset} : \mathit{Set}(\mathit{Ind}) \\ \text{c1} : \overrightarrow{\text{house}}(\text{maxset}) \\ \text{c2} : \text{union(refset,compset,maxset)}\end{bmatrix} \\
\text{cont} : \begin{bmatrix}\text{x=refind} : \mathit{Ind}\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

The lexical entry for ‘mhm’ (slightly modified from Ginzburg, 2012, 286) allows for utterance fragments – that is, the maximal element in pending need not be a full *IllocProp* but can be of its supertype *RecType*:

(28)
$$
\begin{bmatrix}
\text{phon} : \texttt{mhm} \\
\text{cat=interjection} : \mathit{SynCat} \\
\text{dgb-params} : \begin{bmatrix}\text{ag1} : \mathit{Ind} \\ \text{ag2} : \mathit{Ind} \\ \text{MaxPending} : \mathit{RecType} \\ \text{ad} : \text{address(ag1,ag2,MaxPending)}\end{bmatrix} \\
\text{cont}=\begin{bmatrix}\text{c0} : \text{Understand(ag2,MaxPending)}\end{bmatrix} : \mathit{IllocProp}
\end{bmatrix}
$$

'mhm' generates an acknowledgement move that shifts the acknowledged parts onto QUD and MOVES, 'spreading incrementality' over DGBs. An incremental backchannel account has also been developed for Dynamic Syntax (using TTR as incremental semantic framework, Eshghi et al., 2015): here, acknowledgement is modelled in terms of alignment of self- and other-pointers on parse trees. But why not just wait until the end of the turn? Apart from issues of anticipatory comprehension, we speculate that working memory burden also plays a role: once acknowledged, a sub-utterance need not to be maintained in the phonologic loop any more, releasing it from the need to be monitored. But we leave this for future research (Ginzburg and Lücking, 2020).

The so-called multimodal charts so far are strictly speaking *bi*modal charts. In order to integrate more than two modalities, a generalisation of bimodal charts, that is *proper* multimodal charts, has to be given. Technically, this can be done just by adding more layers to the multichart depicted in Fig. 1. Conceptually we suggest to use scores.

3.6 Scores

The string charts in Sec. 3.3 already exemplified a cascading structure in the sense that an acoustic event not only is an event in itself but also instantiates a 'sign event', whose components belong to linguistic knowledge and not to the directly observable world. Multimodal charts (Sec. 3.5) demonstrated that manual gesture events also can be processed. And there is no reason to stop here, given the variety of nonverbal signals (cf. the remarks at the beginning of Sec. 1). In this respect, communication events are like musical events, where a bunch of players (the articulators) jointly perform a piece (Duranti, 1997; this analogy has also been drawn in phonetics, where the vocal articulators (lips, tongue, velum, glottis) are conceived as being organised on a 'gestural score', Browman and Goldstein, 1990.[3]). For the coordinated action of playing the first four bars of the second movement of Beethoven's first Razumovsky quartet, op. 59 no. 1, Cooper (2013) provides the following string event analysis, where we abbreviate his representations of the musical material of the first bar with 'b1', and so on:

[3]We treat the complex interplay of 'sub-articulators' in phonetics as a single articulator on a tier. This indicates that a more detailed, hierarchical account is possible in this respect.

$$
(29)\quad \begin{bmatrix}
\mathrm{v1} & : \mathit{Ind} \\
\mathrm{c_{v1}} & : \mathrm{violin(v1)} \\
\mathrm{v2} & : \mathit{Ind} \\
\mathrm{c_{v2}} & : \mathrm{violin(v2)} \\
\mathrm{va} & : \mathit{Ind} \\
\mathrm{c_{va}} & : \mathrm{viola(va)} \\
\mathrm{co} & : \mathit{Ind} \\
\mathrm{c_{co}} & : \mathrm{cello(co)} \\
\mathrm{c_{player}} & : \mathrm{players(e,} \left\{\mathrm{v1,v2,va,co}\right\}\mathrm{)} \\
\mathrm{c_{play_{v1}}} & : \mathrm{play(v1,e.e_{v1})} \\
\mathrm{c_{play_{v2}}} & : \mathrm{play(v2,e.e_{v2})} \\
\mathrm{c_{play_{va}}} & : \mathrm{play(va,e.e_{va})} \\
\mathrm{c_{play_{co}}} & : \mathrm{play(co,e.e_{co})} \\
\mathrm{e} & : \left(\begin{bmatrix} \mathrm{e_{v1}} : \mathit{Silent} \\ \mathrm{e_{v2}} : \mathit{Silent} \\ \mathrm{e_{va}} : \mathit{Silent} \\ \mathrm{e_{co}} : \mathrm{([b1]^\frown[b2]^\frown[b3])} \end{bmatrix} \frown \begin{bmatrix} \mathrm{e_{v1}} : \mathit{Silent} \\ \mathrm{e_{v2}} : \mathrm{([b4])} \\ \mathrm{e_{va}} : \mathit{Silent} \\ \mathrm{e_{co}} : \mathrm{([b4])} \end{bmatrix}\right)
\end{bmatrix}
$$

We now adopt the notion of musical scores in order to model multimodal communication.

4 Communication scores

The communication event of an agent is partitioned into tiers, where each tier is tied to its articulator like the quartet is decomposed into the single string instruments. We focus on speech, gaze, head, and manual gesture here, further tiers can easily be added on that model. Note that the communication events on each tier include the empty event ε ('silence'), so there is no assumption that signalling goes on all the time on every channel.

$$
(30)\quad \mathit{Tiers} =_{\mathrm{def}} \begin{bmatrix}
\mathrm{ag} & : \mathit{Ind} \\
\mathrm{c_{signal}} & : \mathrm{produce(ag,e)} \\
\mathrm{mth} & : \mathit{Ind} \\
\mathrm{c_{mth}} & : \mathrm{mouth\text{-}of(mth,ag)} \\
\mathrm{c_{art_{mth}}} & : \mathrm{articulate(mth,e.e_{sp})} \\
\mathrm{eye} & : \mathit{Ind} \\
\mathrm{c_{eye}} & : \mathrm{eyes\text{-}of(eye,ag)} \\
\mathrm{c_{art_{eye}}} & : \mathrm{articulate(eye,e.e_{gz})} \\
\mathrm{l\text{-}hnd} & : \mathit{Ind} \\
\mathrm{c_{l\text{-}hnd}} & : \mathrm{left\text{-}hand\text{-}and\text{-}arm\text{-}of(hnd,ag)} \\
\mathrm{c_{art_{l\text{-}hnd}}} & : \mathrm{articulate(l\text{-}hnd,e.e_{gs})} \\
\mathrm{r\text{-}hnd} & : \mathit{Ind} \\
\mathrm{c_{r\text{-}hnd}} & : \mathrm{right\text{-}hand\text{-}and\text{-}arm\text{-}of(hnd,ag)} \\
\mathrm{c_{art_{r\text{-}hnd}}} & : \mathrm{articulate(r\text{-}hnd,e.e_{gs})} \\
\mathrm{c_{sync}} & : \mathrm{Rel(l\text{-}hnd,r\text{-}hnd)} \\
\mathrm{hd} & : \mathit{Ind} \\
\mathrm{c_{hd}} & : \mathrm{head\text{-}of(hd,ag)} \\
\mathrm{c_{art_{hd}}} & : \mathrm{articulate(hd,e.e_{hd})} \\
\mathrm{e} & : \left(\begin{bmatrix} \mathrm{e_{sp}} : \mathit{Phon} \\ \mathrm{e_{gz}} : \mathit{VisSit} \\ \mathrm{e_{gs}} : \mathit{Trajectory} \\ \mathrm{e_{hd}} : \mathit{headMove} \end{bmatrix}\right)^{+}
\end{bmatrix}
$$

Although (30) captures multimodal communication events, the speaker role is a distinguished one:

it is the speaker who produces verbal signals on the speech tier (i.e., *Phon* is not the empty string).

$$(31)\ \text{speaker} = \begin{bmatrix} \text{ag} & : \mathit{Ind} \\ \text{c}_{\text{signal}} & : \text{produce(ag,e)} \\ \text{e} & : (\begin{bmatrix} \text{e}_{\text{sp}} \neq \varepsilon : \mathit{Phon} \end{bmatrix}) \end{bmatrix}$$

We refer to the speaker in this narrower sense as 'c_{spkr} : spkr(ag)', or simply use the reserved label 'spkr' (conversely, 'addr' labels the agent not speaking according to (31)).

Multimodal communication events can be processed in terms of multicharts (Sec. 3.5). Communication scores combine multimodal scores to dialogue gameboards (Sec. 3.4). To this end, a DGB of the type in (15) is distributed over agent dimensions and a dialogue management dimension: utterance events (including the visual field, which is the content of the gaze tier) are tied to the respective contributors while the arguably more objective parts of a dialogue gameboard are tracked on the DGB dimension (for public vs. private partitions of the DGB see Ginzburg, 2012, Sec. 4.2). The blueprint of dyadic dialogue is that two agents jointly produce a discourse, a string of 'bars' (scaling up to multilogue is straightforward):

$$(32)\ \begin{bmatrix} \text{ag1} : \mathit{Ind} \\ \text{ag2} : \mathit{Ind} \\ \text{participants} : \{\text{ag1, ag2}\} \\ \text{c}_{\text{diag}} : \text{produce(ag1,ag2,e)} \\ \text{e} : \mathit{String(Bar)} \end{bmatrix}$$

A bar is an incrementally unfolding unit of discourse which captures (pieces of) contributions made by an agent in isolation as well as overlapping signalling events – thus, it replaces the traditional notion of *turn*, which is too sequential for covering the signalling complexities of dialogue.

$$(33)\ \mathit{Bar} =_{\text{def}} \begin{bmatrix} \text{e}_{\text{ag1}} : \begin{bmatrix} \text{c-utt} : \text{address(ag1,ag2,utt-time)} \\ \text{utt-time} : \mathit{Time} \end{bmatrix} \wedge \mathit{Tiers} \\ \text{e}_{\text{ag2}} : \begin{bmatrix} \text{c-utt} : \text{address(ag2,ag1,utt-time)} \\ \text{utt-time} : \mathit{Time} \end{bmatrix} \wedge \mathit{Tiers} \\ \text{e}_{\text{dgb}} : \begin{bmatrix} \text{facts} & : \text{set}(\mathit{Prop}) \\ \text{pending} & : \text{list}(\mathit{LocProp}) \\ \text{moves} & : \text{list}(\mathit{LocProp}) \\ \text{qud} & : \text{poset(Question)} \\ \text{convrule} & : \text{set}(\mathit{Rule}) \end{bmatrix} \end{bmatrix}$$

Examples are given in Figs. 2 and 3.

Dialogue progression works as follows: The possibly multimodal contributions of dialogue agents ag1 and ag2 are incrementally processed as shown in Sec. 3.5. The (multimodal) locutionary propositions manifest dialogue moves and are the link to DGB. Progress on the DGB dimension is then regimented by conversational rules. The progression of dialogues is now modelled in terms of scores of communication events, which not only integrates several modalities but also captures some overlap phenomena such as backchannelling. A lexical entry for backchannel nodding is as follows (that a couple of nonverbal signals constitute lexical resources has been worked out by Poggi, 2001 for various kinds of signals):

$$(34)\ \begin{bmatrix} \text{spkr} : \mathit{Ind} \\ \text{addr} : \mathit{Ind} \\ \text{e}_{\text{spkr}} : (\begin{bmatrix} \text{e}_{\text{sp}} : \mathit{Phon} \\ \text{e}_{\text{locprop}} : \text{Sign}(\text{e}_{\text{sp}}) \\ \text{p=e}_{\text{locprop}}\text{.cont} : \mathit{Prop} \end{bmatrix}) \\ \text{e}_{\text{addr}} : (\begin{bmatrix} \text{e}_{\text{hd}}\texttt{=nod} : \mathit{headMove} \\ \text{cont=Accept(addr,sprk,p)} : \mathit{IllocProp} \end{bmatrix}) \end{bmatrix}$$

That is, nodding has a dialogue move-based meaning, contributing an illocutionary proposition. Being a lexical resource on their own, a nonverbal signal event *e* can also be the argument of a multimodal extension of the Sign(*e*) function from (13). Take, for example, the short exchange between a speaker (S) and an addressee (A) reported by Bavelas and Gerwing (2011):

(35) S: *And I got a light for Christmas*
A: [slight nod]

An analysis of the backchanelling nodding from (35) is given in Fig. 2, where an empty initial dialogue context is assumed (and DGB typings are omitted for brevity). Research on backchannelling (Bavelas et al., 2000; Bavelas and Gerwing, 2011; Tolins and Fox Tree, 2014) agrees that backchannel signals influence the development of dialogue. We can make these findings more precise in comparing backchannelling from Fig. 2 with its (hypothetical) counterpart in Fig. 3 where nodding is a turn in itself and follows the speaker's utterance. The crucial difference pertains to the DGB dimension: while without backchannelling (Fig. 3) it takes two steps to construct the final dialogue gameboard, backchannelling (Fig. 2) allows to get there in one swoop. Other examples from Sec. 2 can be modelled in basically the same way.

5 Head shake

In Sec. 2 we briefly pointed at utterance coherence, which (presumably not very surprisingly) is also

$$\begin{bmatrix} \text{participants} : \{\text{A, B}\} \\ \text{c}_{\text{diag}} : \text{produce(A,B,e)} \\ \text{e} : \left(\begin{bmatrix} \text{e}_{\text{dgb}} : \begin{bmatrix} \text{facts}=\{\}\text{: set}(\textit{Prop}) \\ \text{pending}=\langle\,\rangle : \text{list}(\textit{LocProp}) \\ \text{moves}=\langle\,\rangle : \text{list}(\textit{LocProp}) \\ \text{qud}=\{\}\text{: poset}(\textit{Question}) \\ \text{convrule}=\{\}\text{: set}(\textit{ConvRule}) \end{bmatrix} \end{bmatrix} \frown \begin{bmatrix} \text{e}_A : \begin{bmatrix} \text{spkr=A} : \textit{Ind} \\ \text{e}_1 : \begin{bmatrix} \text{e}_{\text{sp}}=\texttt{AIL} : \textit{Phon} \end{bmatrix} \end{bmatrix} \end{bmatrix} \frown \begin{bmatrix} \text{e}_A : \begin{bmatrix} \text{spkr=A} : \textit{Ind} \\ \text{e}_2 : \begin{bmatrix} \text{e}_{\text{sp}}=\texttt{FC} : \textit{Phon} \end{bmatrix} \\ \text{e}_{\text{locprop}} : \text{Sign}(\text{e}_1\text{e}_2) \\ \text{p1}=\text{e}_{\text{locprop}}\text{.cont: } \textit{Prop} \end{bmatrix} \\ \text{e}_B : \begin{bmatrix} \text{ag2=B} : \textit{Ind} \\ \text{e}_3 : \begin{bmatrix} \text{e}_{\text{hd}}=\texttt{nod} : \textit{Move} \end{bmatrix} \\ \text{e}_{\text{locprop}} : \text{Sign}(\text{e}_3) \\ \text{accept}=\text{e}_{\text{locprop}}\text{.cont} : \textit{IllocProp} \end{bmatrix} \\ \text{e}_{\text{dgb}} : \begin{bmatrix} \text{facts}=\{\text{p1}\} \\ \text{pending}=\langle\,\rangle \\ \text{moves}=\left\langle \begin{matrix} \text{Accept(B,A,p1),} \\ \text{Assert(A,B,p1)} \end{matrix} \right\rangle \\ \text{qud}=\{\} \\ \text{convrule}=\{\text{FreeSpeech,Accept,Fact update}\} \end{bmatrix} \end{bmatrix} \right) \end{bmatrix}$$

Figure 2: Backchannelling. `AIL` abbreviates 'And I got a light', `FC` abbreviates 'for Christmas'.

obeyed in tier-crossing discourse. We briefly want to illustrate coherence and the kinds of tier-crossing phenomena we want to deal with by means of three simple examples of head shake. For *intra*-speaker head shakes seem to be sensitive to first vs. third person subjects:

(36) a. Peter believes you
head shaking

b. Peter doesn't believe you
head shake

c.?I believe you
head shaking

d. I don't believe you
head shaking

In (36a), the head shake signals the speaker's attitude to the proposition expressed (≈ 'I can't believe that Peter believes you'). (36b) is ambiguous between an attitude and negation interpretation: the head shake can either be as in (36a) ((≈ 'I can't believe that Peter doesn't believe you')), or be a nonverbal expression of the negative particle *n't*.[4] Both readings – emphasising the *multi*-modal, score-based approach – can be dissociated if an emblematic index-finger shaking (meaning roughly "no") is produced simultaneously (but see Subsec-3.5 on refinements with respect to timing) in addition to the head shake. In this case, index finger and head shake are both interpreted as denials – that is, contributing to the illocutionary role –, not as attitude signals, leading to an inconsistency with (36a) but not (36b), since (b) but not (a) expresses a denial.

Curiously, interpretations change when the proposition expressed is a first person report (speaker and subject are identical), as in (36c) and (d). Now, we argue, head shake only contributes to the illocutionary role of the turn, namely signalling a *denying* move. However, denying needs to be witnessed on the propositional level: if there is no negative particle, there is nothing denied, leading to the inconsistency observed in (36c).[5]

6 Conclusion

The formal outline developed in this paper is an attempt to connect incremental natural language processing to multimodality and dialogue gameboards by following the model of music scores. These communication scores may pave the way for formal conversation analysis (i.e., bridging between CA and formal dialogue semantics). Part of future work is to spell out dialogical constraints and generalisations that hold across tiers and between contributions, including a multimodal extension of lexical resources. Notwithstanding multimodality, there

[4]For further uses of the head shake see Kendon (2002).

[5]Head shake, illocutionary roles and attitudes are, among others, further discussed in Ginzburg & Lücking, manuscript in prep.

$$\begin{bmatrix} \text{participants} : \{\text{A, B}\} \\ \text{c}_{\text{diag}} : \text{produce(A,B,e)} \\ \text{e} : \left(\begin{bmatrix} \text{e}_{\text{dgb}} : \begin{bmatrix} \text{facts}=\{\}\text{: set}(\textit{Prop}) \\ \text{pending}=\langle\,\rangle : \text{list}(\textit{LocProp}) \\ \text{moves}=\langle\,\rangle : \text{list}(\textit{LocProp}) \\ \text{qud}=\{\}\text{: poset}(\textit{Question}) \\ \text{convrule}=\{\}\text{: set}(\textit{ConvRule}) \end{bmatrix} \end{bmatrix} \frown \begin{bmatrix} \text{e}_A : \begin{bmatrix} \text{spkr=A} : \textit{Ind} \\ \text{e}_1 : \begin{bmatrix} \text{e}_{\text{sp}}=\texttt{AIC} : \textit{Phon} \end{bmatrix} \\ \text{e}_{\text{locprop}} : \text{Sign}(\text{e}_1) \\ \text{p1=e}_{\text{locprop}}\text{.cont: } \textit{Prop} \end{bmatrix} \\ \text{e}_{\text{dgb}} : \begin{bmatrix} \text{facts}=\{\} \\ \text{pending}=\langle\,\rangle \\ \text{moves}=\langle \text{Assert(A,B,p1)} \rangle \\ \text{qud}=\{\text{p1?}\} \\ \text{convrule}=\{\text{FreeSpeech}\} \end{bmatrix} \end{bmatrix} \frown \begin{bmatrix} \text{e}_B : \begin{bmatrix} \text{ag2=B} : \textit{Ind} \\ \text{e}_2 : \begin{bmatrix} \text{e}_{\text{hd}}=\texttt{nod} : \textit{Move} \end{bmatrix} \\ \text{e}_{\text{locprop}} : \text{Sign}(\text{e}_2) \\ \text{accept=e}_{\text{locprop}}\text{.cont} : \textit{IllocProp} \end{bmatrix} \\ \text{e}_{\text{dgb}} : \begin{bmatrix} \text{facts}=\{\text{p1}\} \\ \text{pending}=\langle\,\rangle \\ \text{moves}=\left\langle \begin{matrix} \text{Accept(B,A,p1),} \\ \text{Assert(A,B,p1)} \end{matrix} \right\rangle \\ \text{qud}=\{\} \\ \text{convrule}=\{\text{Accept,Fact update}\} \end{bmatrix} \end{bmatrix} \right) \end{bmatrix}$$

Figure 3: Without backchannelling. `AIC` abbreviates 'And I got a light for Christmas'.

still seems to be a 'leading voice' among signals, usually speech.

Acknowledgments

We would like to thank three anonymous reviewers not only for their detailed and valuable comments, but also for their benevolent appreciation of the nascent nature of the work presented here. This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the program *Investissements d'Avenir* (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001.

References

Katya Alahverdzhieva, Alex Lascarides, and Dan Flickinger. 2017. Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*, 5(3):421–464.

John L. Austin. 1950. Truth. In *Proceedings of the Aristotelian Society. Supplementary*, volume xxiv, pages 111–128. Reprinted in John L. Austin: *Philosophical Papers*. 2. ed. Oxford: Clarendon Press, 1970.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

Janet B. Bavelas and Jennifer Gerwing. 2011. The listener as addressee in face-to-face dialogue. *International Journal of Listening*, 25(3):178–198.

Rens Bod and Remko Scha. 1996. Data-oriented language processing: An overview. Technical report, Department of Computational Linguistics, Institute for Logic, Language and Computation, University of Amsterdam.

Catherine P. Browman and Louis Goldstein. 1990. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18:299–320.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2-3):333–362.

Robin Cooper. 2008. Type theory with records and unification-based grammar. In Fritz Hamm and Stephan Kepser, editors, *Logics for Linguistic Structures*, number 201 in Trends in Linguistics: Studies and Monographs, pages 9–33. Mouton de Gruyter, Berlin and New York.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Tim Fernando, and Nicholas Asher, editors, *Philosophy of Linguistics*, number 6 in Handbook of Philosophy of Science, pages 271–323. Elsevier, Oxford and Amsterdam.

Robin Cooper. 2013. Type theory, interaction and the perception of linguistic and musical events. In Martin Orwin, Christine Howes, and Ruth Kempson, editors, *Language, Music and Interaction*, pages 67–90. College Publications.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology – LiLT*, 10(4):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.

Thierry Coquand, Randy Pollack, and Makoto Takeyama. 2003. A logical framework with dependently typed records. In *Typed Lambda Calculi and Applications. Proceedings of the 6th International Conference*, TLCA 2003, pages 105–119.

Alessandro Duranti. 1997. Polyphonic discourse: Overlapping in Samoan ceremonial greetings. *Text – Interdisciplinary Journal for the Study of Discourse*, 17(3):349–382.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102.

Cornelia Ebert. 2014. The non-at-issue contributions of gestures. Workshop on Demonstration and Demonstratives, April 11-12 2014, Stuttgart.

Nick J. Enfield. 2009. *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*. Number 13 in Language, Culture and Cognition. Cambridge University Press, Cambridge, UK.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 261–271.

Jerome Feldman. 2012. The neural binding problem(s). *Cognitive neurodynamics*, 7(1):1–11.

Tim Fernando. 2007. Observing events and situations in time. *Linguistics and Philosophy*, 30:527–550.

Tim Fernando. 2011. Constructing situations and time. *Journal of Philosophical Logic*, 40(3):371–396.

Gottlob Frege. 1918. Der Gedanke. *Beiträge zur Philosophie des deutschen Idealismus*, 1(2):58–77.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the first International Workshop on Computational Semantics*, Tilburg, The Netherlands. Katholieke Universiteit Brabant.

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

Jonathan Ginzburg, Robin Cooper, Julian Hough, and Schlangen David. 2018. Incrementality and clarification/sluicing potential. In *Proceedings of Sinn und Bedeutung 21*, volume 1, pages 463–480.

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.

Jonathan Ginzburg and Andy Lücking. 2020. Context as emotionally charged memory. In *Proceedings of The 24th Workshop on the Semantics and Pragmatics of Dialogue*, SemDial/WatchDial.

Jonathan Ginzburg and Massimo Poesio. 2016. Grammar is a system that characterizes talk in interaction. *Frontiers in Psychology*, 7:1938.

Alvin I. Goldman. 1970. *Theory of Human Action*, legacy ed. 2015 edition. Number 1830 in Princeton Legacy Library. Princeton University Press, Princeton, NJ.

Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, Gregory J. Mills, Ronnie Cann R., Wilfried Meyer-Viol, and Patrick G.T. Healey. 2011. Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1):199–233.

Michael Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics – Volume I*, pages 624–630, Montreal, Quebec, Canada.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–276.

Adam Kendon. 2002. Some uses of the head shake. *Gesture*, 2(2):147–182.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, MA.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731).

Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter, Berlin and Boston. Zugl. Diss. Univ. Bielefeld (2011).

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*, volume 8 of *Annals of Computer Science and Information Systems*, pages 383–392. IEEE.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The Bielefeld speech and gesture alignment corpus (SaGA). In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010, pages 92–98, Malta. 7th International Conference for Language Resources and Evaluation.

Andy Lücking and Jonathan Ginzburg. 2019. Not few but all quantifiers can be negated: towards a referentially transparent semantics of quantified noun phrases. In *Proceedings of the Amsterdam Colloquium 2019*, AC'19, pages 269–278.

Craig Martell, Chris Osborn, Jesse Friedman, and Paul Howard. 2002. FORM: A kinematic annotation

scheme and tool for gesture annotation. In *Proceedings of Multimodal Resources and Multimodal Systems Evaluation*, pages 15–22, Las Palmas, Spain. MITRE.

David McNeill. 1992. *Hand and Mind – What Gestures Reveal about Thought*. Chicago University Press, Chicago.

Cornelia Müller. 1998. *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*. Number 1 in Körper – Kultur – Kommunikation. Berlin Verlag, Berlin. Zugl. Diss. FU Berlin (1996).

Massimo Poesio and David Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

Isabella Poggi. 2001. The lexicon and the alphabet of gesture, gaze, and touch. In Angélica de Antonio, Ruth Aylett, and Daniel Ballin, editors, *Intelligent Virtual Agents*, number 2190 in Lecture Notes in Computer Science, pages 235–236. Springer, Berlin and Heidelberg.

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

Kristina Poncin and Hannes Rieser. 2006. Multi-speaker utterances and co-ordination in task-oriented dialogue. *Journal of Pragmatics*, 38(5):718–744. Focus-on Issue: Linguistic Theory and Pragmatics.

Hannes Rieser. 2015. When hands talk to mouth. Gesture and speech as autonomous communicating processes. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, SEMDIAL 2015: goDIAL, pages 122–130, Gothenburg, Sweden.

Emanuel A. Schegloff. 1984. On some gestures' relation to talk. In J. Maxwell Atkinson and John Heritage, editors, *Structures of Social Action. Studies in Conversational Analysis*, Studies in Emotion and Social Interaction, chapter 12, pages 266–296. Cambridge University Press, Cambridge, MA.

Jürgen Streeck. 2008. Depicting by gesture. *Gesture*, 8(3):285–301.

Jackson Tolins and Jean E. Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164.

Joost Zwarts. 2003. Vectors across spatial domains: From place to size, orientation, shape, and parts. In *Representing Direction in Language and Space*, number 1 in Explorations in Language and Space, chapter 3, pages 39–68. Oxford University Press, Oxford, NY.

Classification of low-agreement Pronouns through Collaborative Dialogue: A Proof of Concept

Janosch Haber and **Massimo Poesio**
Queen Mary University of London
{j.haber|m.poesio}@qmul.ac.uk

Abstract

When evaluating model performance on automated annotation tasks such as anaphora resolution and specifically pronoun resolution, the gold standards often postulate a single correct referent for each referring expression. Previous research on annotator disagreement however found that in some cases there might not actually be a single correct referent, but rather multiple acceptable alternatives - or no specific solution at all. In this paper we aim to extend the study on pronouns with low annotator agreement by investigating how speakers react to various occurrences of *it* pronouns in a collaborative dialogue setting. We propose that different types of ambiguity and under-specification can be distinguished based on whether or not speakers discuss the resolution of a given pronoun during the task, and whether or not they agree on a referent for it afterwards. Applied to corpus samples which previously received low annotator agreement, we propose that this methodology provides a comparatively cheap means to aid distinguishing annotation noise from genuine classification difficulty.

1 Introduction

For decades now it has been practice to train or at least evaluate Natural Language Processing (NLP) systems on annotated datasets. Linguistic annotations are traditionally compiled using either a number of domain experts or trained annotators, or, if annotations are required on a larger scale, in recent years are increasingly crowd-sourced from layman judgements through Games-with-a-purpose (GWAPs, see von Ahn et al., 2006; von Ahn and Dabbish, 2008), Citizen Science Projects and other paid and unpaid online applications. As especially the latter approach tends to lead to noisy data, it is customary to collect multiple labels for every item and establish the final gold label by aggregating or adjudicating the various collected annotations. This means that the resulting gold standard - based on which models are trained and against which systems will be tested - postulates a single correct label for each item only. Focusing on anaphora resolution, the task of resolving the dependencies of expressions that rely on some form of linguistic context (Poesio et al., 2016, p. 24), empiric research on annotator disagreement however revealed that not in all cases is data collection noise to blame for obtaining a range of alternative labels, but rather that sometimes the different labels are due to genuine disagreement among the annotators (see e.g. Poesio and Artstein, 2005a; Versley, 2008).

1.1 Causes of Annotator Disagreement

Genuine annotator disagreement on pronoun resolution can be the result of a number of factors that play a role in the processing of a anaphoric expressions. Here, we will focus on three types of pronouns that have been found to make up a substantial part of low-agreement samples: ambiguous, under-specified and expletive pronouns. The term *ambiguity* is often used to to indicate expressions with unclear meaning, but in a more strict sense only refers to expressions whose interpretation can be selected from a discrete set of alternatives, separating the phenomenon from instances of vagueness (Pinkal, 1985; Poesio, Forthcoming). If an anaphoric expression is ambiguous in the narrow sense, we expect annotations to resemble a distribution of labels according to the different interpretations' dominance, and the collected labels should therefore be spread over a few alternative candidates. *Under-specification* is used to classify instances of dialogue or written text for which a listener or reader does not seem to build a complete interpretation (Poesio and Reyle, 2001). First evidence of under-specified expressions had been

reported in Poesio et al. (2006) who observed that participants in a collaborative dialogue task used pronouns *it* and *that* to refer to complex objects that were not explicitly introduced to the discourse before. While often syntactically infelicitous or semantically ambiguous, under-specified expressions have usually been found not to interfere with task success, originating the notion of *justified sloppiness* to account for infelicitous expressions that do not negatively affect the interaction and in some cases might even increase task efficiency (Piantadosi et al., 2012). If the referent of a pronoun is under-specified in this way, we expect that a traditional annotation setup will yield a wide spread of candidate referents as annotators are likely to select different combinations of available anchors to proxy the under-specified referent not explicitly available in the linguistic context. *Expletive*, or *non-referring* pronouns lastly do not have a linguistic referent at all, but rather function as dummy pronouns catering to grammatical requirements of syntactic composition. If identified correctly, expletive pronouns should not be assigned a referent, but annotation noise may lead to a small set of spuriously linked anchors.

1.2 Discussing Disagreements

To further investigate the notion of genuine annotator disagreement caused by ambiguity, under-specification and expletives as opposed to annotation noise, we propose to extend the analysis of items obtaining low annotator agreement in crowd-sourced corpora like Phrase Detectives (Poesio et al., 2019) by testing them in a dedicated dialogue task. The reasoning behind this proposal is as follows: In crowd-sourced annotation tasks, annotators usually are not aware of their disagreement as it emerges only as a result of their individual label assignments. This means that, for example, disagreements on difficult items can arise because i) annotators are not aware of the ambiguity or under-specification of a given expression and label it in different ways, or ii) they might be unable to express ambiguity or under-specification in a canonical way corresponding to those of other annotators. If annotators however get an opportunity to discuss items, they could be made aware of alternative readings, either because those are directly proposed by other annotators, or because of disagreements in interpretation when comparing individually assigned labels, and referents can be assigned collaboratively. Observing these referent discussions in the dialogue task thus could be used as an indicator of item difficulty and samples causing them marked accordingly, whereas low-agreement items not discussed in this setting are likely to be a result of annotation noise.

Explicit discussion of annotation labels is usually used in the adjudication of expert annotators' labels (see for example Finlayson and Erjavec, 2017), but here we propose to apply this paradigm implicitly in a fully crowd-sourcing setting by having crowd workers discuss text segments containing pronouns that received low agreements on their crowd-sourced referent labels. To do so, we developed a novel two-staged dialogue task. In the first part of the task, a short text passage containing a low-agreement pronoun is split into segments, which are presented to a participant pair in random order. The participants are then asked to collaboratively re-produce the original passage by re-ordering the segments. This requires the participants to establish the referents for at least those pronouns central to the understanding of the passage, and is likely to prompt a discussion if participants are in doubt whether they share the same interpretation of these pronouns. Once the participants agree on an ordering of the text snippets, they enter the second part of the task where - now individually - they are asked to assign referents to the pronouns occurring in the text. We propose that combining observations on whether or not participants discussed a pronoun in the first part and whether or not they agree on the individually assigned referents in the second part will yield insightful information for a classification of low-agreement items:

1. If low agreement was caused by annotation noise in the original data collection, we expect participants in our setup to agree on the referent in the second stage, but not necessarily discuss the pronoun during the initial collaborative re-ordering phase, because it likely that the pronoun is actually straightforward to resolve and doesn't require deliberation.

2. If low agreement is due to referent ambiguity, we expect participants to discuss the pronoun and its resolution during the first part of the task, and determine how they will interpret it for the re-ordering of the passage. This means

that they should agree on the assigned referent in the second part of the task.

3. If an under-specified item caused low annotator agreement, we expect that participants in the dialogue task will not extensively discuss the pronoun during the first stage, as the *justified sloppiness* hypothesis predicts that the pronoun's referent does not need to be fully specified to complete the interaction objective (here re-ordering the text segment) - but we also expect that participants are unlikely to agree on a referent in the second part of the task, as they never made a referent explicit.

4. Lastly, non-referring expletive *it*s are expected to produce high agreement (on not assigning a referent) during the second part of the task, but to be unlikely to trigger any discussion of referents during the first part.

In this paper we present a pilot study intended as a proof of concept for the classification of low-agreement items through collaborative dialogue. We conducted a small-scale pilot run in the lab rather than applying actual crowd-sourcing techniques, and used a longer input text that specifically contains instances of all three types of problematic *it* pronouns in order to quickly collect dedicated data for all cases. We show that the collaborative task allows us to classify the three types of *it* pronouns occurring in the text using the output of the two stages of the task, validating the task setup, and outline the further development of the task into a proper crowd-sourcing tool.

2 Related Work

The dialogue task presented in this paper draws from previous work surrounding the discovery of under-specified expressions, especially observed in dialogue settings, and evidence collected to support the idea that annotator disagreement is not necessarily data collection noise, but rather can sometimes be used as valuable data for improving computational attempts to anaphoric resolution.

2.1 Under-specification

Investigating the transcripts of the TRAINS corpus (Gross et al., 1993; Allen and Heeman, 1995), a set of dialogues collected in an wizard-of-oz setup for planning alongside an automated system, Poesio et al. (2006) observed that in some cases participants - as well as the instructor posing as the automated planning system - used anaphoric expressions that seemed to not have a clear referent in the preceding discourse. Consider for example excerpt (1) where participant A is tasked to send a boxcar to a train station in Corning. B is the instructor posing as an automated planning system:

(1) A: Can we kindly hook up, uh, engine E2 to the boxcar at Elmira and send <u>it</u> to Corning as soon as possible please?
B: Okay.
A: Do let me know when <u>it</u> gets there.
B: Okay, <u>it</u> should get there at 2AM.

Resolving the highlighted *it* pronouns traditionally would require the selection of one of the two available anchors *boxcar* or *engine*, rendering them ambiguous in this context. Speaker B however does not seem to raise a complaint over this ambiguity but continues the interaction as if the referent was identified without any issues. Poesio and Reyle (2001) propose a number of alternative ways to explain speaker B's processing of the under-specified reference, but independent of how exactly this processing occurs, the interaction can be presumed successful, as the plan formulated by speaker A is understood and correctly executed by speaker B. This led the authors to introduce the concept of *justified sloppiness*, arguing that expressions can be left under-specified when their realisation is irrelevant to the conversation goal. Versley (2008) later adopted this notion of under-specification and formulated the *Generalised Sloppiness Hypothesis* based on Asher and Pustejovsky (2006)'s dot-object formulation for complex types. According to this model, anchor *engine E2* is represented as an under-specified complex object with the two combined aspects *engine•train*, and *boxcar* as *boxcar•train*, respectively. The instruction *Send ? to Corning* then selects for the *train* aspect of both anchors only, allowing them to be combined into a singular mereologic structure that can be referred to with *it*.

Taking a different approach, Recasens et al. (2011) propose that co-reference does not require identity of reference, but rather that items which are merely "near-identical" can be perceived as co-referent. This hypothesis builds on previous work by Swets et al. (2008) who observed lingering effects of syntactic garden-path sentences and

proposed that the interpretation of those structures in online processing is not fully disambiguated but rather "good enough" to continue reading.

2.2 Annotator Disagreement as Data

Concerning the use of annotator disagreement to enrich model input istead of disregarding it as data collection noise, Plank et al. (2014b) for example observed that annotator disagreements in part-of-speech (POS) tagging were systematic across domains and to a certain extend also across languages, and that a majority of high-disagreement items actually stemmed from linguistically debatable instances. Utilising expected annotator disagreement for hard cases as a weight in the loss function of an automatic POS tagger, Plank et al. (2014a) subsequently recorded gains in model accuracy and downstream task performance. Others, too, have suggested to specifically identify and adapt the treatment of ambiguous and subjective cases in corpora from a range of domains and tasks, including for example image segmentation (Firman et al., 2018), judgements about visual scenes (Sharmanska et al., 2016), dialogue addressing annotations (Reidsma and op den Akker, 2008) or classifying expressions as semantically old or new (Klebanov and Beigman, 2014).

Battling the noise in crowd-sourced annotations, a growing body of recent work presents approaches to estimate annotator reliability and use it as a metric to weigh annotations when aggregating gold labels from annotations, as for example in Aroyo and Welty (2013, 2015); Schüller (2018) and Paun et al. (2018).

3 Method

The pilot study presented in this paper was developed to establish a proof of concept before applying the dialogue task to a large-scale crowd-sourced validation of previously collected annotations. This means we tested the task with a small number of participants in a lab setting rather than online, and instead of using actual text segments containing low agreement items from corpora such as ARRAU (Poesio and Artstein, 2005b) or Phrase Detectives (Poesio et al., 2019), we created a longer sample text that contains instances of all three of the *it* pronoun types mentioned earlier.

The pilot text segment takes the form of an interview with a fictional band member. The interviewee reports about a failure in guitarist *Sophie*'s equipment during a live show and how she and the sound engineer *Megan* attempt to fix it. The story predominantly uses *she* pronouns to refer to either of these two characters, and reference needs to be established in order to determine the actors for most part of the story. In context, the *she* pronouns are unambiguous, but due to the initially random ordering of snippets in the task, they remain ambiguous until placed in order. Besides these personal pronouns, the story includes *it* references that in the same way are initially ambiguous between a number of objects but can largely be resolved once context is established, non-referring *it* pronouns (which remain expletive in established context), and *it* references to mereologic constructions that are not explicitly mentioned in the text. The text segment is divided into an introduction paragraph, ten snippets containing mostly a single sentence each, and a short ending paragraph.[1] The ten snippets are indicated with a random letter to aid the transcription of deictic references and show pronouns underlined and numbered with a small sub-script.

3.1 Task Setup

Two participants with university-level, self-reported native English language skills were instructed that they were to participate in a two-staged task. The first part would require them to collaboratively re-construct a story which was cut into a number of snippets, and the second part to answer some questions about the re-ordered story. The participants were then seated next to one another at a table in a small conference room. Paper prints of text snippets containing the start and ending paragraph of the story were laid in the middle of the table, as well as a simple diagram showing the story's setting with colour-coded circles for characters and icons for objects. Each participant was then given five random snippets. The participants were asked to first read the story's start and their own snippets and then collaboratively order the ten snippets below the story start. Participant interaction was audio recorded during the task and the instructor took notes of deictic references to task elements to aid the transcription of the interaction.

Once the participants signalled that they agreed on a snippet ordering, both were handed mention sheets listing all of the text's underlined pronouns. The participants were then asked to individually

[1] See Appendix A.1 for the full pilot text.

assign referents to the pronouns by either writing down the referent or assigning the same colour as indicated in the diagram. This second stage was completed without participant interaction. After participants finished filling in the mention sheets, the instructor compared the sheets and asked questions about some of the pronouns, especially if they were not assigned a referent or if the participants disagreed on referents, in order to make sure that the mention sheets represented the resolution of pronouns as intended by the participants. The mention sheets were then compared for differences in the resolution of pronouns to establish agreement and disagreement on referents by the end of the interaction.

3.2 Annotation Scheme

Besides the participants' agreements and disagreements on the particular referents of the pronouns we are interested in whether or not a pronoun's referent was discussed during the interaction in the first stage. Since this is less straightforward to assess than participant agreement, we instructed trained annotators to go through the transcripts and mark all instances of pronouns being discussed based on a simplified account of core speech acts *propose*, *discuss* and *accept* (e.g. Poesio and Traum, 1997) combined into two different cases:

1. **Pronoun resolved.** A pronoun is considered to be resolved during the interaction if i) its referent was either explicitly discussed by the participants or ii) a referent was indirectly established by snippet ordering and accepted by both participants.

   A referent is considered to be discussed explicitly if either i) a participant refers to the pronoun and then proposes a referent for it, ii) a participant re-formulates the snippet to contain a referent instead of the pronoun originally used in the snippet, or iii) a participant refers to a snippet that contains a pronoun and proposes that it "is" or "is done by" or "is about" a certain referent, and the referent is accepted by the other participant. A referent is established indirectly if either i) a participant reads a block of snippets containing a given pronoun, indicating that they form a coherent unit, and the pronoun is ambiguously resolved in the newly established context, or ii) a participant proposes that a snippet should precede or succeed another snippet or block of snippets which form a context that unambiguously resolves the pronoun, and the proposal is accepted by the other participant. A proposed referent finally is considered as accepted if the other participant either i) explicitly agrees with the proposal or ii) changes the topic of the conversation, indicating that the proposal does not require further discussion.

2. **Pronoun unresolved** A pronoun is considered to remain unresolved after the interaction if a proposed referent was rejected by the other participant or the pronoun and its referent were not discussed during the interaction.

The annotators were not made aware of the proposed distinction between ambiguous, non-referring and under-specified *it* pronouns.

4 Results

During the pilot run we collected and transcribed the conversations of 11 participant pairs discussing the re-ordering of the pilot story. Conversations took between 5 and 20 minutes (mean = 11:04 min, std = 5:58 min), with transcripts containing between 38 and 151 utterances (mean = 109.09, std = 42.49). The mention sheets individually filled in by the participants after their interaction show an overall agreement on 60.80 % for the referents of the pronouns used in the story's snippets, with agreement rates for the different pairs of participants ranging between 31.25% and 87.50%. Noticing these high fluctuations, we decided to exclude transcripts from further analysis if participants did not agree on the referents of at least 7 of the 8 personal pronouns in the story. Since the personal pronouns need to be resolved collaboratively during the interaction in order to establish a common ground about the events in the story, we argue that in interactions where participants fail to agree on the referents of these mentions, it is unlikely that they both correctly understood the task or the story. Only 7 out of the 11 participant pairs met this requirement, resulting in a filtered dataset with conversations lasting between 6 and 18 minutes (mean = 10:14 min, std = 5:08 min), transcripts containing between 64 and 151 utterances (mean = 108.29, std = 37.51) and an overall agreement on referents in 71.43% of the cases, with a lowest individual score of exactly 50%.

Pronoun Type	1	2	3	4	5	6	7	Total
Personal Pronoun	100.00%	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%	98.21%
Ambiguous *it*	100.00%	0.00%	100.00%	33.33%	0.00%	66.67%	100.00%	57.14%
Non-referring *it*	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%	57.14%
Under-specified *it*	33.33%	0.00%	33.33%	0.00%	33.33%	0.00%	66.67%	23.81%
Total	87.50%	50.00%	81.25%	68.75%	56.25%	75.00%	81.25%	71.43%

Table 1: Referent agreements per participant pair and pronoun type in the data collection pilot run.

Pronoun Type	1	2	3	4	5	6	7	Total
Personal Pronoun	84.38%	87.50%	87.50%	84.38%	78.13%	71.88%	75.00%	81.25%
Ambiguous *it*	100.00%	91.67%	91.67%	83.33%	75.00%	75.00%	75.00%	84.52%
Non-referring *it*	87.50%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	98.21%
Under-specified *it*	91.67%	100.00%	91.67%	91.67%	66.67%	58.33%	91.67%	84.52%
Total	89.06%	92.19%	90.63%	87.50%	78.13%	73.44%	81.25%	84.60%

Table 2: Annotator agreement on their assessment whether or not a participant pair resolved a given pronoun during their interaction. Results grouped per transcript and pronoun type.

4.1 Evaluation of Agreements

A detailed overview over the agreements of the different participant pairs and pronoun types is shown in Table 1. In the filtered dataset, all but one participant pair agree on the referents of all eight personal pronouns in the story, with the remaining pair disagreeing only on one item, resulting in an average agreement rate of 98.21% for the personal pronouns. This is significantly higher than for any of the *it* pronoun types (z-score for population proportions p-value < 0.01). The total agreement on both ambiguous and non-referring *it* pronouns in turn is 57.14%, significantly higher than the total agreement rate of 23.81% for the under-specified *it* pronouns (p-values = 0.028 and 0.046, respectively). Note that agreement is regarded independent of whether participants assigned the intended referent, and agreement on non-referring pronouns was assumed if neither participant assigned a referent as well.

4.2 Evaluation of Annotations

We collected annotations from four trained annotators instructed to follow the annotation guidelines as stipulated above. The average observed agreement concerning the resolution of pronouns by participants reached 0.73, with inter-annotator agreement however only scoring a Cohen's κ (Cohen, 1960) of 0.47, and pairwise κ scores ranging between 0.31 and 0.75. Table 2 shows in more detail the observed annotator agreement concerning whether or not a pronoun's referent was resolved by the participants in the interaction, differentiated by participant pair and pronoun type. Annotator agreement varies between different participant pairs, ranging from 73.44% to 92.19%, but is almost identical among the different pronoun types, with only non-referring pronouns showing a significantly higher annotator agreement rate than any of the other types.

4.3 Evaluation of Referent Discussion

In order to assess whether a pronoun's referent should be considered as resolved during the interaction, we aggregated annotator labels through majority vote. Considering the low annotator agreement observed above, we did so for all four annotator labels and the labels obtained from the two annotators with highest agreement only and compared results. This filtering however only had a minimal effect, and consequently it was decided to use the full set of annotations and keep its expressiveness. Table 3 shows the ratios of resolved pronouns per transcript and pronoun type based on the aggregated labels of all four annotators. According to the annotators' judgement, ambiguous *it* pronouns are collaboratively resolved most often during the interactive part of the dialogue task, although not significantly more often than the ambiguous personal pronouns (z-score p-value = 0.1436). Non-referring *it* are annotated to have never been collaboratively resolved during the interaction, a finding that directly corresponds to the nature of the pronoun type. Under-specified pronouns lastly are discussed significantly less than the personal and ambiguous *it* pronouns (p-values < 0.01), but not significantly more often than the non-referring pronouns (p-value = 0.1391). Note that no observed

Pronoun Type	1	2	3	4	5	6	7	Total
Personal Pronoun	100.00%	100.00%	75.00%	75.00%	75.00%	87.50%	62.50%	82.14%
Ambiguous *it*	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	100.00%	95.24%
Non-referring *it*	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Under-specified *it*	66.67%	0.00%	33.33%	0.00%	0.00%	0.00%	0.00%	14.29%
Total	81.25%	68.75%	56.25%	56.25%	56.25%	62.50%	50.00%	61.61%

Table 3: Pronoun referent resolution per participant pair and pronoun type as judged by independent external annotators.

discussion of a referent does not entail that participants did not assign a referent at all. This becomes especially clear when keeping in mind that referents for ambiguous and most under-specified pronouns actually were assigned by the individual participants during the second part of the task.

4.4 Combination of Factors for Classification

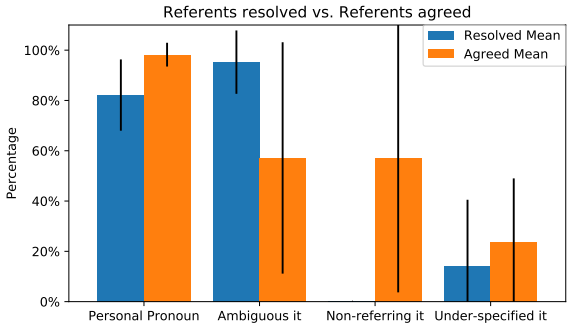


Figure 1: Distribution of pronouns resolved during the interaction (blue) and referents agreed afterwards (orange), grouped per pronoun type.

Figure 1 shows the overall distribution of pronouns collaboratively resolved by participants during their interaction in the first part of the task (blue) compared to the ratio of referent agreement in the second part (orange). The figure suggests that there are structural differences in participants' reactions to the four different types of pronouns tested, with (ambiguous) *she* pronouns showing almost perfect agreement with a relatively high amount of discussions, ambiguous *it* pronouns showing more discussion but less agreement, non-referring *it* pronouns exhibiting mediocre agreement but no discussion whatsoever, and under-specified *it* pronouns displaying low levels of both, discussions and agreement. When tested with χ^2, the resulting test statistic is improbably large compared to the χ^2 distribution given six degrees of freedom ($\chi^2 = 48.23$, p-value < 0.01), indicating that the distribution of the four combinations of metrics significantly differ between the three pronoun types. Combining information on whether pronouns were discussed in the first part and whether they were agreed upon in the second part thus seems to uniquely identify the three different types of pronouns, as supported by the contingency table in Table 4.

Pronoun Type	¬R ¬A	R ¬A	¬R A	R A
Ambiguous *it*	0	9	1	11
Non-referring *it*	6	0	8	0
Under-specified *it*	14	2	4	1

Table 4: Contingency table containing counts of instances where pronoun referents were collaboratively resolved (R) during the first part of the task and agreed upon (A) in the second part of the task.

5 Discussion

While the obtained results seem promising prima facie, there are some factors to be taken into account when interpreting them with regard to a classification of low agreement items, and in perspective of transforming the pilot layout to a crowd-sourcing setting.

5.1 Pilot Shortcomings

Ambiguous pronouns. The ambiguous personal and *it* pronouns in the pilot text were not ambiguous in context but rather ambiguous as to their referent prior to putting them in order. When using previously annotated texts as input data, referents can remain (and in many of the genuine low-agreement items will remain) ambiguous after assigning a snippet order. We however argue that this does not have a major impact on the task setup, as the important factor is whether or not the pronoun and its referent were discussed during the first part of the task. Additionally, it is likely that participants will collaboratively decide on a specific interpretation even for genuinely ambiguous pronouns in order to ground their interpretation in the discussion of snippet ordering.

Snippet ordering. In the pilot setup, we analysed

transcripts independent of the fact whether participants re-created the intended ordering of the text segment or not. Considering however that the pilot text was created so that it could generate a number of sensible orderings, and that a future data collection using crowd-sourcing will have much shorter input texts (see section 5.2), we believe that this does not discount our proof of concept.

Annotation of implicit referent proposals. The annotation guidelines for marking the discussion and resolution of referents during the interaction allows for *implicit* proposals through snippet ordering, i.e. when a snippet containing a pronoun is proposed to succeed a snippet containing an anaphoric referent. We realised that these cases were difficult to spot even for trained annotators, especially considering that under-specified and non-referring pronouns do not actually refer to a preceding anchor. When applied in a crowd-sourced manner, implicit proposals therefore will need to be re-defined or disregarded entirely.

Rejection rate. A last issue we cannot readily explain is the high number of participant pairs who did not agree on referents for the *she* pronouns crucial for an understanding of the pilot text. It appears that in some cases the task of re-ordering snippets overshadowed the processing of the text itself, in some cases participants had questionable text comprehension skills, and in some cases the the task might have been misunderstood. We however expect that a simplified setup with shorter input texts will mitigate this effect.

5.2 Transformation to Crowd-Sourcing

Considering the pilot to provide proof of concept, we believe that the developed dialogue task could be applied in a crowd-sourcing setting to classify low-agreement items in previously annotated corpora. A number of modifications however needs to be made to transform the pilot setup into a data collection tool suitable for crowd-souring. Specifically, we need to automate the generation of input segments from previously annotated corpora, and transform to an online setting i) the collaborative dialogue phase, ii) the individual referent assignment phase, and iii) the annotation of whether or not a pronoun is discussed in the dialogue stage. Regarding the generation of input texts, further work is needed to investigate whether the contexts of low-agreement items derived from corpora like ARRAU or Phrase Detectives can be automatically reduced to a length feasible for the collaborative re-ordering task. Considering the conversion of the first stage of the task, we already started developing a first digital version of the pilot using crowd-sourcing tool SLURK (Schlangen et al., 2018). The tasks' user interface was re-designed to contain a private section holding the snippets and a shared section to collaboratively arrange them, as well as a chat box, moving from spoken to written dialogue to eliminate the need for transcription and making the task more accessible. We are currently working on updating the instructions to further gamify the task to increase worker commitment and internal motivation. We see few issues with implementing the individual second part of the task, but consider obtaining independent annotations concerning the discussion of referents through crowd-sourcing to be the most daunting part of converting the task setup. As we already observed relatively low annotator agreement even among linguistically trained annotators, the best strategy for simplifying the annotation task for crowd-sourcing remains a central open issue to be addressed by future work.

6 Conclusion

The data collected through the pilot setup presented in this paper indicate that three different types of pronouns that are likely to cause low annotator agreement in classic annotation tasks can be differentiated based on observations obtained from a dedicated dialogue task which implicitly requires participants to collaboratively resolve their referents. Ambiguous expressions are found to be discussed and resolved more often than other types of pronouns, with the referents of non-referring pronouns usually being agreed upon without explicit discussion and under-specified pronouns triggering relatively little discussion but also leading to low levels of agreement. In future work, this combination of observations could be used to augment current gold standard annotations by classifying low-agreement items as either caused by genuine item difficulty or annotation noise, and provide richer data for training and testing NLP systems. Our findings also provide some support for the hypotheses of *justified sloppiness* and *good enough representations*, suggesting that referents of under-specified pronouns do not need to be fully resolved in order to be interpreted similarly enough to not hinder understanding and ultimately task success.

Acknowledgements

The work presented in this paper was supported by the DALI project, ERC Grant 695662. The authors would like to thank Derya Çokal and Andrea Bruera for their input, and the anonymous reviewers for their feedback.

References

Luis von Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Commun. ACM*, 51(8):58–67.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: A Game for Collecting Common-sense Facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 75–78, New York, NY, USA. ACM.

J Allen and P A Heeman. 1995. TRAINS Spoken Dialog Corpus [CD-ROM]. *Philadelphia, PA: Linguistic Data Consortium*.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Nicholas Asher and James Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6(1):38Asher, N., & Pustejovsky, J. (2006). A type comp.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mark A Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer.

Michael Firman, Neill DF Campbell, Lourdes Agapito, and Gabriel J Brostow. 2018. Diversenet: When one right answer is not enough. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5598–5607.

Derek Gross, James F. Allen, and David R. Traum. 1993. The trains 91 dialogues. Technical report, University of Rochester, USA.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396.

Silviu Paun, Bob Carpenter, J D Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280 – 291.

Manfred Pinkal. 1985. *Logik Und Lexikon: Die Semantik des Unbestimmten*. De Gruyter.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio. Forthcoming. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Blackwell Companion to Semantics*. Wiley.

Massimo Poesio and Ron Artstein. 2005a. Annotating (anaphoric) ambiguity. In *In Proc. of the Corpus Linguistics Conference*.

Massimo Poesio and Ron Artstein. 2005b. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Massimo Poesio and Uwe Reyle. 2001. Underspecification in anaphoric reference. *Fourth International Workshop on Computational Semantics (IWCS-4)*.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution*. Springer.

Massimo Poesio, Patrick Sturt, Ron Artstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*.

Massimo Poesio and David R Traum. 1997. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. slurk–a lightweight interaction server for dialogue experiments and data collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (Aix-Dial/semdial 2018)*.

Peter Schüller. 2018. Adjudication of coreference annotations via answer set optimisation. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(4):525–546.

Viktoriia Sharmanska, Daniel Hernández-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2194–2202.

Benjamin Swets, Timothy Desmet, Charles Clifton, and Fernanda Ferreira. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3-4):333–353.

A Appendices

A.1 Pilot Text Segment

The text segment used in the pilot study contained the following snippets (presented in intended order):

Introduction
MusicManiac Magazine interviewed art-rock trio Backyard Billionaires during the SonicBang! festival last week. The band was founded in 2002 and consists of guitarist Sophie Hood, keyboard player James Flores and drummer Tony Marino. Since 2005, the trio is accompanied by Tony's sister Megan who creates live visualisations and doubles as the band's sound-engineer. James told us the following story about the first time that Megan was put in charge of mixing their sound during a concert: We arrived late a bit and had little time for setting up, but we got everything working and Megan managed to create a decent mix for the venue. We got right into our set and played two or three songs, but then we suddenly couldn't hear Sophie anymore.

P From the corner of my eye I saw her1 double-checking her$_2$ gear while we$_3$ continued playing.

Y At the same time I saw Megan panicking at the back of the room.

A She$_1$ was pushing all kinds of buttons trying to get the guitar back, but instead she2 only managed to mute me$_3$ and Tony as well.

T It$_1$ got pretty quiet in the room and soon some people in the audience started laughing and others started to boo.

S It$_1$ was rather awkward.

V Sophie finally figured out that her$_1$ amplifier was the problem.

C It$_1$ had stopped working due to a blown fuse.

I Having found the source of the problem, she$_1$ picked up her guitar again.

L She1 unplugged it$_2$ and hooked it$_3$ up to a second input of my$_4$ speaker.

U It$_1$ didn't work directly, but she$_2$ turned some knobs and managed to get it$_3$ back into the mix.

X (End)
In the end it$_1$ didn't sound as good as before, but at least we$_2$ could finish the gig and prevent a total disaster.

The underlined pronouns were classified as follows:

Ambiguous Personal Pronouns
P1, P2, A1, A2, V1, I1, L1, U2
Ambiguous *it* Pronouns
C1, L2, L3
Under-specified *it* Pronouns
U1, U3, X1
Non-referring *it* Pronouns
T1, S1

Communicable reasons: How children learn topoi through dialogue

Ellen Breitholtz and Christine Howes
Centre for Linguistic Theory and Studies in Probability
Department for Philosophy, Linguistics and Theory of Science
University of Gothenburg
ellen.breitholtz@ling.gu.se; christine.howes@gu.se

Abstract

Children's acquisition of language requires their learning of not just words/concepts and linguistic structure but how these interact in dialogue with knowledge about the world, our interlocutors, the shared environment and social norms. In this paper we explore how children acquire the rhetorical resources that they need in dialogue. These *topoi* are the underpinning warrants for incomplete (*enthymematic*) arguments in dialogue. We illustrate our account with examples from dialogues with children that demonstrate the topoi which they have learned – particularly in cases where these topoi are unexpected from the adult language user's perspective, and sketch a formal model using Type Theory with Records.

1 Introduction

As shown in (1), participating in any dialogic exchange requires a wealth of knowledge, not just about the linguistic items used, but about the world, our interlocutor(s), the shared environment and social norms. These factors are usually considered to be outside the remit of linguistics proper, consigned either to the pragmatics wastebasket (Bar-Hillel, 1971) or to sociolinguistics. However, the distinctions between linguistics, pragmatics and social factors are hard to justify when we look at language as it is used in everyday interaction, as noted by Labov (1972): "I have resisted the term sociolinguistics for many years, since it implies that there can be a successful linguistic theory or practice which is not social." (p13).

(1) Dave: ...you're gonna be home from football until four, you gonna have your dinner, want a bath.
Lee: Yeah, but I might not go to school tomorrow.
Dave: Why?
Lee: Cos of my cough.
Dave: How can you play football and not go to school then?
Lee: Cos I was going out in the fresh air, I'm alright when I'm out in the fresh air.
Dave: So why aren't you going to school then?
Lee: I'm in the class room all day dad. [BNC KBE 10554-10561]

Here, in (1), it is evident to any competent language user that Lee is conveying that he is well enough to play football but not well enough to go to school because football takes place outdoors. Conversely, his father Dave infers that if Lee is well enough to play football then he is well enough to go to school. But how do we learn that being ill restricts certain activities; or the possible exceptions to this 'rule', such as where the activity occurs? In this paper we will explore how inference plays a role in building up the rhetorical resources required to be active participants in dialogue through interaction. We perceive the rhetorical resources available to a conversational participant as an open set of *topoi*. "Topos" is a term used by Aristotle to refer to a principle of reasoning warranting an argument in discourse.

The structure of the paper is as follows: First we will give some background to reasoning in dialogue, the role of interaction in language acquisition, and the building up of rhetorical resources. In Section 3 we will look at a number of examples from interactions between children and adults which demonstrate how children draw on and acquire rhetorical resources in interaction. In Section 4 we will suggest a way of modelling how the acquisition of rhetorical resources works and finally, in Section 5, we will discuss how the insights of this paper may be developed.

2 Background

2.1 Reasoning in interaction

Reasoning is essential in communication since interacting with others frequently involves making non-logical common-sense inferences linking context, background knowledge and beliefs to utterances in the dialogue in order to understand one another. Following Breitholtz and Cooper (2011); Breitholtz (2011, 2014a), we will use the Aristotelian term *enthymeme* in connection with such inferences. An enthymeme is an argument which appeals to what is in the listener's mind, i.e. an interlocutor must draw on background knowledge or contextual information to correctly interpret the argument. If a dialogue participant presents the argument P therefore Q, an interlocutor must supply a warrant that P is a valid reason for Q in order for the argument to be successful. These warrants are often referred to as *topoi* (Aristotle, 2007; Ducrot, 1988).

In (1), the enthymeme from Dave's perspective can be depicted as (2).[1] This could be underpinned by a more generally applicable topos such as the ones shown in (3) and (4).

(2) $$\frac{\text{ill}(Lee)}{\text{stay_home}(Lee)}$$

(3) $$\frac{\text{ill}(x) \quad \text{need_rest}(x)}{\text{stay_home}(x)}$$

(4) $$\frac{\text{ill}(x) \quad \text{contagious}(x)}{\text{stay_home}(x)}$$

When we interact we expect topoi to be common ground, or – if they are not – to be *accommodated* (adopted by dialogue participants Karttunen, 1974; Stalnaker, 1974) during the course of the interaction. In many contexts there might be several acceptable topoi, and misunderstandings and disagreement can arise if interlocutors assume different topoi (Jackson and Jacobs, 1980).

2.2 Interaction in acquisition

Although traditional linguistics and much developmental psychology still assumes that there must be an innate language learning facility due to the presumed 'poverty of the stimulus' of a child's linguistic input (Berwick et al., 2011), there is a large body of evidence that refutes this position, from both a computational (Clark and Lappin, 2010) and a more social perspective. This work (exemplified by Halliday, 1975, 1994 and Tomasello, 1992 a.o.) emphasises the nature of language as action, and makes explicit the role of interaction in language acquisition.

Specifically, research on child language acquisition underscores the importance of the social environment for the language learning child (Stephens and Matthews, 2014). Children are active in interactions with their caregivers long before they produce language and evidence suggests that it is this learning to interact (e.g. through gaze Gredebäck et al., 2010 and turn-taking Hilbrink et al., 2015; Casillas, 2014) which bootstraps language acquisition (Levinson's (2006) 'interaction engine').

Research on children in the early stages of verbal language acquisition also shows that children learn new words and concepts through interaction (Clark, 2015). The input from adults may be explicit or implicit. An example of explicit input would be an adult saying to a child "look at that, it's an elephant", while implicit input would be to just use the word "elephant" to refer to an object without introducing it. Brown and Hanlon (1970) show that child directed explicit disapprovals of grammar mistakes are very unusual. Rather, more recent research suggests that implicit input, such as corrective feedback, occurs more frequently and has a significant effect on language learning (Hiller and Fernández, 2016).

How children's language acquisition in interaction can be formalised has not been much explored, with one notable exception being the sketch using an information state update approach and dialogue gameboards (see section 4, below) developed in Ginzburg and Moradlou (2013). Additionally, Clark (2015) notes that a Type Theory with Records account offers a principled way to formalise how language use develops through interaction. Our formal sketch of topoi acquisition (section 4, below) using the same mechanisms fits in with such approaches.

2.3 Building rhetorical resources

As with lexical concepts, which are learned through repeated encounters with words in interaction,

[1] In this semi-formal notation, the premises are shown above the line and the conclusion below, as is standard. The wiggly line denotes a not strictly logical chain of reasoning. These are not intended to be complete formal representations, rather as a convenient and clear way of representing our intuitions about topoi and enthymemes. More complete formal representations are shown in section 4.

topoi can be learned through repeated encounters with enthymematic arguments in dialogue. And, also analogously to the acquisition of concepts, this can be more or less explicit "You can't go to school today because you're poorly. You have to stay home and rest", compared to "I might not go to school tomorrow ... cos of my cough" (1).

However, topoi may also be acquired through inference. In order to abstract principles of reasoning from co-occurring situations, a child must have a notion of situation type, that is it needs to be able to make the judgement that a situation of a particular type, such as when they let go of an object, is followed by a situation of another type, namely one where the object falls to the ground. In order for a child to acquire the topos "If x is dropped, x falls to the ground", the child must abstract away from particular situations to establish that this topos of gravity generally holds and can be used to make predictions about objects in general, not just about the specific situations about e.g. "my spoon" in which they have encountered it. Children may then use the topos in a situation that the child judges to be of the same type as that of the original situation. Often the child manages to apply a topos which is acceptable in the situation, and the dialogical reasoning works seamlessly. However, as with children's well-documented overextention of lexical items (Clark, 2009; Gelman et al., 1998), in some cases the child overextends the domain to which the topos applies. In Section 3 we will look at some examples of how children apply topoi successfully and unsuccessfully in interaction with adults.

There are also specific dialogue strategies available for acquiring new knowledge, including topoi – such as asking questions (Chouinard et al., 2007). Bova (2011); Bova and Arcidiacono (2013) show that the main functions of why-questions asked by children in the home is argumentative and explanatory. Schlöder et al. (2016) suggest an analysis where why-questions are taken to be focused on eliciting acceptable reasons. Correspondingly, in order to be functional, responses to why-questions must be warranted by an acceptable principle of reasoning. On this account, the reason why illness is an acceptable answer to Dave's question "Why?" in (1) is that there are topoi regarding illness and rest (when you are sick you should rest in order to get better, as depicted in (3)) and illness and contagion (if you are ill you should not mix with others so as to not pass on the infection

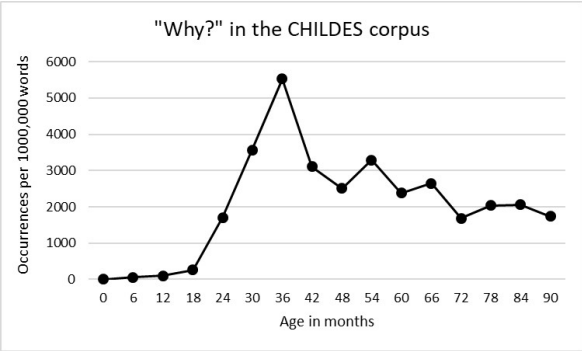


Figure 1: Frequency of 'why' in child language by age

(4)). These underpin Lee's utterance "Cos of my cough" in response to Dave's why-question regarding Dave's statement that he might not go to school tomorrow. Since why-questions specifically relate to pragmatic reasoning in dialogue, we can expect language users who are less familiar with the topoi that pertain to a given domain and are acceptable in a particular socio-cultural and situational context to ask more why-questions. A search of the CHILDES-corpus (MacWhinney, 2000) shows that children have a period around the age of 3 where the number of why-questions they ask peak (see figure 1, obtained using the ChildFreq tool Bååth 2010). This asking of why-questions is an efficient way for a child to expand their rhetorical resources through a recently acquired conversational strategy (Frazier et al., 2009).

3 Childrens' use of enthymemes

Young children are also clearly able to draw enthymematic connections in interaction, as shown in (5).

(5) *4;1 year old Greta is playing with two dolls "Loria" and "Masha", January 2020*

Greta:	Loria's sleeping *[snoring noises]*
Mother:	Loria's really noisy when she's sleeping. Has she got a cold?
Greta:	Yes
(as Loria):	I have got a cold so I can't go to school ...
	Later in the same game:
(as Masha):	But why can't we go to school?
(as Loria):	Because we've got a cold and we snore bad

In (5) Greta uses an enthymematic utterance "I have got a cold so I can't go to school" demonstrating that she has previously acquired an appropriate topos (perhaps the one shown in (3), above), which licenses the enthymeme. She also adds a new topos, depicted in (6), that has been supplied enthymematically by her mother in the form of a question early on in the dialogue and combines these two topoi in a play dialogue between the two dolls, thus demonstrating that she can apply the new topos appropriately.

(6) $$\frac{\text{snores}(x)}{\text{has_cold}(x)}$$

New situations, such as the current coronavirus pandemic, lead to new topoi becoming available, or more specific instantiations of existing topoi. In the case of coronavirus, this offers a unique opportunity to investigate the emergence of new topoi which became part of the standard rhetorical resources of a very large number of people over a very short period of time. This is in contrast to the more usual situation where new topoi emerge in restricted domains, activities or groups.

Even for young children, this may also lead to more specification of existing topoi as shown in (7). In this example, Greta is off nursery with a very slight cough which would not normally be sufficient to stay at home (so the topos in (3) does not apply). The official advice in Sweden at the time was: "If you feel ill with symptoms including a runny or blocked nose, cough, or fever you should avoid contact with other people. This also applies if you only feel a little bit unwell. Do not go to work or to school. It is very important not to risk passing the illness to anyone else."[2]

(7) *Greta (4;3) is off nursery with a minor cold playing with a doll "Lily", March 2020*

(as Lily): There's this illness going around that's why we're not going to dagis (=nursery)

In this case, the enthymeme used by Greta does not make reference to either herself or Lily being even slightly ill – it is rather the presence of the illness (coronavirus) in the community at large which is responsible for their staying at home.

[2]Source: https://www.folkhalsomyndigheten.se/the-public-health-agency-of-sweden/communicable-disease-control/covid-19/

3.1 Unexpected topoi

As discussed in section 2.3, cases where children acquire rhetorical resources which do not completely correspond with those of an adult are particularly illustrative of how topoi are learned. These show how children generalise from small amounts of data to reason about novel situations, as in (8).

(8) Child: *Lever din mormor?*
Lives your grandmother?
'Is your grandmother alive?'

Mother: *Nej, hon är död.*
No, she is dead.

Child: *[med anklagande ton]*
[with accusatory tone]

Puttade du henne?
Pushed you her?
'Did you push her?'

Mother commenting on twitter:

Ja, treåringen kan ha tittat
Yes, three-year-old$_{DET}$ can have watched
LITE för mycket på Lejonkungen.
little too much Lion-King$_{DET}$
'Yes, the three year old might have watched the Lion King a little too often.'

In (8), the three year old child's understanding of death is generalised from their limited experience, which, according to the mother, comes exclusively from the film 'The Lion King'. In this film, the main character Simba's father dies after he is pushed from a cliff edge by his brother Scar. The child here has correctly generalised the concept of death to other living things than lions, but has stored a 'death topos' with a causal relationship, namely that if someone is dead then someone related pushed them (9).

(9) $$\frac{\text{dead}(x) \quad \text{related}(x,y)}{\text{pushed}(x,y)}$$

This topos is applied constructively by the child in the dialogue shown in (8). When she learns that her mother's grandmother is dead, she concludes that she must have been pushed and infers that it was her mother who did the pushing.

In a similar example ((10) taken from Breitholtz 2015), the child has extrapolated a topos (11) from their previous experience of the concept 'widow', which comes exclusively from fairy tales.

(10) *Reading a bedtime story to 4 year old child*

Mother: *Snövits mor dog kort efter födseln och en tid senare gifte hennes far, Kungen, om sig. Hans nya hustru var vacker men fåfäng och elak.*

Snow White's mother died shortly after the birth, and after some time her father, the king, remarried. His new wife was beautiful but vain and wicked.

Annie: *Ja mamma – en änka!*
Yes mum – a widow!

(11) $\underbrace{\text{beautiful}(x) \quad \text{vain}(x) \quad \text{wicked}(x)}_{\text{widow}(x)}$

To anyone familiar with the conventional meaning of the word *widow*, it seems obvious that the child has got it wrong. However, as discussed in Breitholtz (2015), this gives us an insight into how most of us usually get it right, and suggests that reasoning using *enthymemes* can be a means not only of lexical disambiguation (as suggested by Pustejovsky, 1998), but also a means of acquiring new concepts.

While clearly linked to reasoning and interaction, it could be argued that these two examples of unexpected topoi are only about the lexical concepts that the child has acquired. However, although we believe that this is because lexical concepts can also be conceived as underpinned by topoi (and therefore use the same processes – see also Rehder 2003), there are also examples of children using unexpected topoi which cannot be reduced to a lexical concept, which also require explanation.

Figure 2, taken from a graphic explanation of coronavirus aimed at children,[3] and the statement from the BBC news website shown in 12.[4] show that the topos (13) that older people (and those with underlying medical conditions) are more likely to become more ill, or die if they contract coronavirus is now a generally accepted topos, that obviously did not exist in 2019.

(12) The elderly and the unwell are more likely to die, if they contract coronavirus.

(13) $\underbrace{\text{has_corona}(x) \quad \text{old}(x)}_{\text{die}(x)}$

[3]Source: https://www.npr.org/sections/goatsandsoda/2020/02/28/809580453/just-for-kids-a-comic-exploring-the-new-coronavirus

[4]Source: https://www.bbc.com/news/health-51674743, 24th March 2020

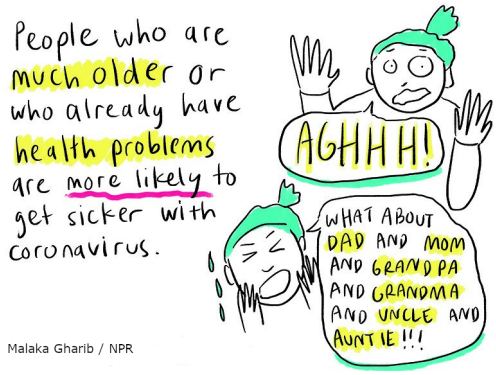


Figure 2: Underlying topos: Old people are more likely to get more ill with coronavirus

Although a more general version that older people (and those with underlying medical conditions) are more likely to become more ill, or die if they contract other potentially fatal infectious diseases probably did exist for most adults prior to the coronavirus, Greta has not previously encountered such a topos. Nevertheless, based on her corona-specific new topos, she is able to apply a more general (in this case incorrect) version of this topos (15).

(14) *Conversation with 4;3 year old Greta in March 2020 (coronavirus times)*

Greta: What would happen if you drank the sea water?

Mother: It would make you poorly.

Greta: Really poorly?

Mother: Yes.

Greta: Old people would die. I don't know about us though.

(15) $\underbrace{\text{is_poorly}(x) \quad \text{old}(x)}_{\text{die}(x)}$

4 Updating enthymematic dialogues

The dialogue in (14) is an example where a child overextends the domain for when a particular topos applies. In this section we will sketch a formal account of such overextension.

To account for the reasoning involved in the building up of rhetorical resources, we will use an information state update approach using dialogue gameboards cast in TTR, a type theory with records (Larsson, 2002; Ginzburg, 2012). The basic idea of this approach is that agents involved in interaction

need to coordinate, and we use gameboards to represent how the agents keep track of where they are in the creation of particular dialogue events (e.g. a project like finding out something or a move like asking, responding etc). Each agent has their own view of the state of the game, and thus we have separate gameboards for each of the participants in an interaction. Being able to account for separate views of the state of the dialogue plays an essential role in coordination, especially with regards to miscommunication, where there is a mismatch between the participants' dialogue gameboards.

4.1 Dialogue gameboards as types in TTR

Following Ginzburg (2012); Cooper and Ginzburg (2015); Cooper (in progress) we model dialogue gameboards in TTR, a type theory with records (Cooper, 2005, 2012). The basis of TTR is our ability to perceive and classify the world, i.e. to perceive objects and situations in the world as being of *types* such as *Ind*, the type of entities such as humans, animals, things, and *ptypes*, consisting of a predicate and its arguments, for example *see(a,b)*, "a sees b". In order to represent complex situations which potentially involve many ptypes and individuals, as well as other more general types, we use *record types*. A record type is a structure of pairs of labels and types, where labels may represent things like individuals, predicates and events.

$$(16)\quad \begin{bmatrix} \mathrm{x}{:}\mathit{Ind} \\ \mathrm{c}_{dog}{:}\mathrm{dog(x)} \\ \mathrm{c}_{run}{:}\mathrm{run(x)} \end{bmatrix}$$

The object to which the label x points in (16) is of type *Ind*. There are two constraints on the type of situation – that this individual is a dog (c_{dog}:dog(x)) and that it runs (c_{run}:run(x)). In addition to record types we also want to be able to talk about situations that are witnesses of record types, actual situations that are of certain types. We represent such objects as *records*. A record is a structure where the labels are associated with *values* rather than types. In (17) we see a record representing one particular situation. This situation is of the type in (16) if all the values are of the appropriate types (fido: *Ind*, s_1: dog(fido) and s_2: run(fido)). If these conditions are fulfilled, the record in (17) is a witness of the type of situation in (16). For an in-depth discussion and formal definitions, see Cooper (in progress).

$$(17)\quad \begin{bmatrix} \mathrm{x} & = \mathrm{fido} \\ \mathrm{c}_{dog} & = s_1 \\ \mathrm{c}_{run} & = s_2 \end{bmatrix}$$

In 18 we see an example of a (minimal) dialogue gameboard. The field "shared" holds information that the agent takes to be shared, either as it has been explicitly referred to in the dialogue, or because the agent expects it to have been accommodated. The label "rhet_resources." is associated with the set of topoi that the agent has access to.

$$(18)\quad \begin{bmatrix} \mathrm{private}{:}\begin{bmatrix}\mathrm{rhet_resources{:}set}(\mathit{Topos})\end{bmatrix} \\ \mathrm{shared}{:}\begin{bmatrix}\mathrm{eud{:}list}(\mathit{Enthymeme})\end{bmatrix} \end{bmatrix}$$

4.2 Drinking seawater

Following Breitholtz and Cooper (2011); Breitholtz (2014b,a); Ginzburg et al. (2015); Breitholtz et al. (2017) we model topoi and enthymemes as functions from records to record types[5].

Intuitively this means that if we have a situation of a particular type, we can predict a certain type of situation. For example, the enthymeme conveyed by the mother in (14) says that if someone drinks sea water, it will make them poorly, that is, if you perceive a situation where someone drinks sea water, you can predict that it will make them poorly, as seen in (19).

$$(19)\quad \epsilon_1 = \lambda r{:}\begin{bmatrix} \mathrm{x}{:}\mathit{Ind} \\ \mathrm{c{:}person(x)} \\ \mathrm{e{:}drink_sea_water(x)} \end{bmatrix} . \begin{bmatrix}\mathrm{e{:}make_poorly}(r.\mathrm{e},\ r.\mathrm{x})\end{bmatrix}$$

[5]There are several reasons why we want to model enthymemes and topoi as the same kind of formal object. First, there is no principled difference between the specificity of enthymemes in general and topoi in general – what in one context is an enthymeme may in another context be a topos warranting another enthymeme. Secondly, it is problematic to model enthymemes and topoi as different kinds of formal objects parallel to types of situations as record types and situations as records. We are treating enthymemes and topoi as functions from records to record types (of type *Rec* → *RecType*). If we were to treat topoi as types of enthymemes we would need to introduces new types of types (subtypes of *RecType*) in order to make the distinctions we want. Alternatively, we could model an enthymeme as a function from a record to another record (of type *Rec* → *Rec*). This is problematic, however, since we want the enthymeme to represent a function from an actual (observed) situation to a type of situation to which there might or might not be a witness. For example, enthymemes can be used as instructions for action, that is, to create a situation of the type returned by the function by applying the enthymeme to the observed situation. In such cases there are no witness to the type returned by the function, unless the instructions are carried out.

The enthymeme conveyed by the child, on the other hand, says that if old people drink sea water, they will die, as seen in (20).

$$(20)\quad \epsilon_2 = \lambda r: \begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{c:person(x)} \\ \text{c}_1\text{:old(x)} \\ \text{e:drink_sea_water(x)} \end{bmatrix} . \begin{bmatrix} \text{e:make_die}(r.\text{e},r.\text{x}) \end{bmatrix}$$

The enthymeme produced by the child, ϵ_2, is not likely to be acceptable to most adults, and it is unlikely that the child has received input saying explicitly that old people would die from drinking sea water. Still, there must be some topos warranting it. So how did the child acquire this topos? We argue that is is through overextension and accommodation.

Sometimes we encounter enthymemes which we cannot make sense of, either since the topos is unfamiliar or because we fail to recognise the enthymeme as a specification of a particular topos which is already in our resources. The unfamiliar enthymeme could then be tentatively incorporated into the rhetorical resources of the language user. When the agent encounters similar enthymemes, they may eventually extend the domain of these related enthymemes and construe a topos that warrants all of them. However, children's tendency to overextension (Barrett, 1978) combined with our general ability of accommodation cause children to sometimes integrate topoi in their resources that are not necessarily warranted by the input. In this case, the child has presumably encountered much input of the kind in figure 2 and news stories like the one in (12), which convey the enthymeme in (21).

$$(21)\quad \epsilon_{input} = \lambda r: \begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{y:}\mathit{Ind} \\ \text{z:}\mathit{Type} \\ \text{c:person(x)} \\ c_1\text{:person(y)} \\ \text{c}_2\text{:young(x)} \\ \text{c}_3\text{:old(y)} \\ \text{c}_4\text{:cause_of_harm(z)} \\ \text{c}_5\text{:corona(z)} \\ \text{c}_4\text{:(x)} \\ \text{e:make_poorly(z,x)} \end{bmatrix} . \begin{bmatrix} \text{e:make_die}(r.\text{z}, r.\text{y}) \end{bmatrix}$$

The mother might already have access to a topos warranting ϵ_{input}. However, it would probably be based on various sources of input and not as general as the topos adopted by the child, but rather one where the domain of the topos is delimited to situations involving infectious diseases or similar, as seen below in (22).

$$(22)\quad \tau_{adult} = \lambda r: \begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{y:}\mathit{Ind} \\ \text{z:}\mathit{Type} \\ \text{c:person(x)} \\ \text{c}_1\text{:person(y)} \\ \text{c}_2\text{:young(x)} \\ \text{c}_3\text{:old(y)} \\ \text{c}_4\text{:cause_of_harm(z)} \\ \text{c}_4\text{:disease(x)} \\ \text{e:make_poorly(z,x)} \end{bmatrix} . \begin{bmatrix} \text{e:make_die}(r.\text{z}, r.\text{y}) \end{bmatrix}$$

The child, on the other hand, having overextended ϵ_{input}, has adopted a topos like τ_{child} below in 23.

$$(23)\quad \tau_{child} = \lambda r: \begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{y:}\mathit{Ind} \\ \text{z:}\mathit{Type} \\ \text{c:person(x)} \\ \text{c}_1\text{:person(y)} \\ \text{c}_2\text{:young(x)} \\ \text{c}_3\text{:old(y)} \\ \text{c}_4\text{:cause_of_harm(z)} \\ \text{e:make_poorly(z,x)} \end{bmatrix} . \begin{bmatrix} \text{e:make_die}(r.\text{z}, r.\text{y}) \end{bmatrix}$$

τ_{child} says that if there is a situation where a young person is affected by some cause of harm which makes them poorly, and there is an old person who is affected by the same cause, we are licensed to predict a type of situation where the old person dies. Such topos would warrant Greta's enthymeme ϵ_2 conveyed in (20).

So, how are these topoi integrated into our rhetorical resources? For standard adult topos acquisition, the update rule we want to use is one that says that if there is no accessible topos in the agent's rhetorical resources, they are allowed to add the encountered enthymeme to their resources, thus making it a tentative topos. However, in order to account for the adoption of generalised versions of enthymemes – which is a highly efficient strategy for building up rhetorical resources, although it sometimes leads to overextension – we must adjust the update rule to allow for accommodation of more general versions of encountered enthymemes.

$$f_accommodate_topos =$$
$$\lambda r\colon \begin{bmatrix} \text{private:} \begin{bmatrix} \text{rhet_resources:set}(\textit{Topos}) \end{bmatrix} \\ \text{shared:} \begin{bmatrix} \text{eud:list}(\textit{Enthymeme}) \end{bmatrix} \end{bmatrix} .$$
$$\lambda e\colon \neg \begin{bmatrix} \text{t:}\textit{Topos} \\ \text{c}_1\text{:in}(r.\text{private.rhet_resources(t)}) \\ \text{c}_2\text{:spec(t, fst}(r.\text{shared.eud})) \end{bmatrix} .$$
$$\begin{bmatrix} \text{private:} \begin{bmatrix} \text{rhet_resources=[fst.}(r.\text{shared.eud})\text{]:set}(\textit{Topos}) \end{bmatrix} \end{bmatrix}$$

Figure 3: $f_accommodate_topos$

The update rule in Figure 3 says that if there is an enthymeme on the an agent's dialogue gameboard, for example ϵ_{input}, and it is not the case that there is a topos in the agent's rhetorical resources in relation to which ϵ_{input} is a specification (spec), then the agent is licensed to add the enthymeme to their rhetorical resources. By an enthymeme being a specification of a topos (or another enthymeme) we mean that if there is an enthymeme ϵ = $\lambda r : T_1 \cdot T_2(r)$, and a topos τ = $\lambda r : T_3 \cdot T_4(r)$, ϵ is a specification of τ iff $T_1 \sqsubseteq T_3$ and $\epsilon(r) \sqsubseteq$ of $\tau(r)$ for any r. So, this update rule would account for an agent encountering ϵ_{input} and simply adding that to their resources. However, the process leading to overextension of topoi cannot be accounted for through this update rule. Instead, we need an alternative rule like $f_accommodate_topos'$ in Figure 4, which is like $f_accommodate_topos$, except that the result of applying it is that the agent adopts a topos which is a *generalisation* of the enthymeme currently under discussion rather than the version conveyed in the actual discourse.

5 Conclusions and future work

In this paper we have seen how examples of errors in children's acquisition of topoi show how they generalise from exposure to repeated enthymematic arguments in dialogue. We have provided a formal sketch for how this process operates, but of course at this stage it is just that – a sketch – and many of the details need fleshing out if our intuitions are to be able to provide a full model of how children build up their rhetorical resources.

We believe that our approach – providing a formal analysis of a limited number of genuine examples is complementary to rigorous empirical analysis (experimental and corpus studies) and that looking at uses of language in interactions with children also offers insights into adult language use and the dynamic nature of all of our rhetorical resources. This is particularly apparent in a situation like the current coronavirus pandemic, in which new topoi have quickly become shared across populations, and are now taken largely for granted.

One question that arises is how to delimit the adopting of topoi which are generalisations of encountered enthymemes. For example, it seems quite acceptable that someone who has no previous experience of cats on encountering a particular purring cat construes a topos saying that cats purr. However, it is not obvious that we would like a theory that allows any generalisation, for instance to all four legged animals. However, as we have seen, children do indeed make these kinds of overextensions, and we would like to be able to account for that.

On our view it is clear that the use of topoi is not, like much else in language, all-or-nothing, but must be couched probabilistically. This is true for both when a particular topoi is applied or accommodated in a dialogue, and for the content of the topoi themselves (as in (12), (13) where the conclusion should not be that old people will definitely die if they contract coronavirus, just that they are more likely to than younger people). Interestingly, our examples suggest that young children may not extract such nuanced information – though whether they are able to do so is of course an empirical question. Thus, one of our current research goals is to introduce a probabilistic component into our model, for example extend the probabilistic account of TTR (Cooper et al., 2015) to topoi.

As a basis for this, we would need considerably more data than what we have presented in this paper. However, quantitative corpus studies of enthymemes and topoi are problematic since they are not easy to find automatically, and manual searches that would produce enough data for quantitative studies are not feasible. One option that might be worth pursuing in this context is enthymeme mining (Razuvayevskaya and Teufel, 2017; Maraev et al., 2020).

6 The last word

(24) Greta: Coronavirus I want you to go away soon. I'm bored of it. Are you, too?

$$
\begin{gathered}
f_accommodate_topos' = \\
\lambda r\colon \begin{bmatrix} \text{private:} \begin{bmatrix} \text{rhet_resources:set}(\textit{Topos}) \end{bmatrix} \\ \text{shared:} \begin{bmatrix} \text{eud:list}(\textit{Enthymeme}) \end{bmatrix} \end{bmatrix} . \\
\lambda e\colon \neg \begin{bmatrix} \text{t:}\textit{Topos} \\ \text{c}_1\text{:in}(r.\text{private.rhet_resources(t)}) \\ \text{c}_2\text{:spec(t, fst}(r.\text{shared.eud})) \end{bmatrix} . \\
\begin{bmatrix} \text{private:} \begin{bmatrix} \text{rhet_resources=[gen(fst.}(r.\text{shared.eud}))\text{]:set}(\textit{Topos}) \end{bmatrix} \end{bmatrix}
\end{gathered}
$$

Figure 4: $f_accommodate_topos'$

Acknowledgments

We thank the Swedish Research Council for funding our project "Incremental Reasoning in Dialogue" (2016-01162), and our third projcet member, Robin Cooper, for valuable input and discussion.

References

George A. Kennedy Aristotle. 2007. *On Rhetoric, a theory of civic discourse*. Oxford University Press.

Rasmus Bååth. 2010. Childfreq: An online tool to explore word frequencies in child language. *Lucs Minor*, 16:1–6.

Yehoshua Bar-Hillel. 1971. Out of the pragmatic wastebasket. *Linguistic inquiry*, 2(3):401–407.

Martyn D. Barrett. 1978. Lexical development and overextension in child language. *Journal of child language*, 5(2):205–219.

Robert C. Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science*, 35(7):1207–1242.

Antonio Bova. 2011. Functions of "why" questions asked by children in family conversations. *Procedia – Social and Behavioral Sciences*, 30:776–782.

Antonio Bova and Francesco Arcidiacono. 2013. Investigating children's why-questions: A study comparing argumentative and explanatory function. *Discourse Studies*, 15(6):713–734.

Ellen Breitholtz. 2011. Enthymemes under Discussion: Towards a micro-rhetorical approach to dialogue modelling. In *Proceedings of SPR-11 ILCLI International Workshop on Semantics, Pragmatics, and Rhetoric*.

Ellen Breitholtz. 2014a. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, University of Gothenburg.

Ellen Breitholtz. 2014b. Reasoning with topoi - towards a rhetorical approach to non-monotonicity. In *Proceedings of the 50th anniversery convention of the AISB*, pages 190–198. AISB.

Ellen Breitholtz. 2015. Are widows always wicked? learning concepts through enthymematic reasoning. In *Type Theory and Lexical Semantics: ESSLLI 2015*.

Ellen Breitholtz and Robin Cooper. 2011. Enthymemes as rhetorical resources. In *SemDial 2011 - Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (LosAngelogue)*, pages 149–157.

Ellen Breitholtz, Christine Howes, and Robin Cooper. 2017. Incrementality all the way up. In *Proceedings of the Computing Natural Language Inference Workshop*.

Roger Brown and C. Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In J.R. Hayes, editor, *Cognition and the development of language*. Wiley, New York.

Marisa Casillas. 2014. Turn-taking. In Danielle Matthews, editor, *Pragmatic development in first language acquisition*, pages 53–70. John Benjamins Amsterdam, The Netherlands.

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.

Alexander Clark and Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Eve V. Clark. 2009. *First language acquisition*. Cambridge University Press.

Eve V. Clark. 2015. Semantics and language acquisition. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, pages 714 – 733. Wiley-Blackwell.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 12:99–112.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*. Elsevier BV.

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

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *Linguistic issues in language technology*, 10(4):1–45.

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

Oswald Ducrot. 1988. Topoï et formes topique. *Bulletin d'études de la linguistique française*, 22:1–14.

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.

Susan A. Gelman, William Croft, Panfang Fu, Timothy Clausner, and Gail Gottfried. 1998. Why is a pomegranate an apple? the role of shape, taxonomic relatedness, and prior lexical knowledge in children's overextensions of apple and dog. *Journal of Child language*, 25(2):267–291.

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

Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*.

Jonathan Ginzburg and Sara Moradlou. 2013. The earliest utterances in dialogue: towards a formal theory of parent/child talk in interaction. In *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Amsterdam, The Netherlands. SEMDIAL.

Gustaf Gredebäck, Linn Fikke, and Annika Melinder. 2010. The development of joint visual attention: a longitudinal study of gaze following during interactions with mothers and strangers. *Developmental science*, 13(6):839–848.

Michael Alexander Kirkwood Halliday. 1975. Learning how to mean. In *Foundations of language development*, pages 239–265. Elsevier.

Michael Alexander Kirkwood Halliday. 1994. The place of dialogue in children's construction of meaning. In R.B. Ruddel, M. Rapp Ruddell, and Singer H., editors, *Theoretical models and processes of reading*. International Reading Association, Newark, DE.

Elma E. Hilbrink, Merideth Gattis, and Stephen C. Levinson. 2015. Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction. *Frontiers in psychology*, 6:1492.

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.

Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*, 66(3):251–265.

Lauri Karttunen. 1974. Presupposition and linguistic context. *Theoretical linguistics*, 1(1-3):181–194.

William Labov. 1972. *Sociolinguistic patterns*. 4. University of Pennsylvania Press.

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

Stephen C. Levinson. 2006. On the human "interaction engine'. In N. J. Enfield and Levinson S. C., editors, *Roots of Human Sociality: Culture, Cognition and Interaction*. The Wenner-Gren Foundation, Oxford.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Vladislav Maraev, Ellen Breitholtz, and Christine Howes. 2020. How do you make an AI get the joke? Here's what I found on the web. In *First AISB Symposium on Conversational AI (SoCAI)*.

James Pustejovsky. 1998. *The Generative Lexicon*. MIT Press.

Olesya Razuvayevskaya and Simone Teufel. 2017. Finding enthymemes in real-world texts: A feasibility study. *Argument & Computation*, 8(2):113–129.

Bob Rehder. 2003. Categorization as causal reasoning. *Cognitive Science*, 27(5):709–748.

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

Robert Stalnaker. 1974. Pragmatic presuppositions. In Milton K. Munitz and Peter Unger, editors, *Semantics and Philosophy*. New York University Press.

Gemma Stephens and Danielle Matthews. 2014. The communicative infant from 0-18 months: The social-cognitive foundations of pragmatic development. In Danielle Matthews, editor, *Pragmatic development in first language acquisition*, pages 13–36. John Benjamins Publishing Company.

Michael Tomasello. 1992. The social bases of language acquisition. *Social development*, 1(1):67–87.

Extensions are Indeterminate if Intensions are Classifiers

Staffan Larsson
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
sl@ling.gu.se

Abstract

In this paper, we explore some consequences of the idea of using classifiers to model intensions of natural language expressions for the notion of extensional meaning, i.e. the idea that the meaning of a word can be modelled as the set of referents in the world (or in a possible world). The upshot is that at least for words referring to observable situations, extensional meaning is derivative of intensional meaning, and modelling meanings as sets of referents is of limited applicability.

1 Introduction

In formal semantics in the Montague tradition (Montague, 1974), which we might refer to as Possible Worlds Semantics (PWS), the meaning of a word such as "dog" is taken to be its extension, i.e. the set of all dogs in the world (or in a possible world). As a research program, PWS has in many ways been quite successful. Within the research area of formal semantics, PWS is still the dominant theoretical framework, and forms the theoretical foundation for a majority of the research published in the major journals and conferences in the field.

Over the last decade or so, work in several neighbouring areas have, in various ways and more or less explicitly, used *classifiers* to model *intensional* meanings. Putting it more precisely, in this work *perceptual meaning* has been modelled intensionally using classifiers. Perceptual meaning is an important aspect of the meaning of linguistic expressions referring to physical objects (such as concrete nouns or noun phrases).

Knowing the perceptual meaning of an expression allows an agent to identify perceived objects and situations falling under the meaning of the expression. For example, knowing the perceptual meaning of "blue" would allow an agent to correctly identify blue objects. Similarly, an agent which is able to compute the perceptual meaning of "a boy hugs a dog" will be able to correctly classify situations where a boy hugs a dog.

In this paper, we will explore some consequences of the idea of using classifiers to model intensions of natural language expressions for the notion of extensional meaning. We first discuss the role of classifiers in natural language processing, in formal semantics, and in TTR, a Type Theory with Records (Cooper, 2012, in progress). We then briefly discuss semantic coordination before moving on to a comparison with related ideas put forward in the framework of possible world semantics. The next section comprises the central part of this paper, namely a three-pronged argument as to why extensional semantics are not well suited for modelling perceptual meanings, especially if the latter are regarded as classifiers, and hence why extensional semantics does not (and indeed cannot) constitute a general account of natural language meaning. We then discuss cases where nevertheless extensional semantics may be adequate, also throwing some light on why extensional semantics may initially appear to be an intuitive theory of natural language meaning. Finally, we briefly discuss the applicability of the argument to non-concrete words, before providing a summary and conclusions.

2 Classifiers and natural language

The usefulness of *classifiers* (Rosenblatt, 1958; Harnad, 1990) in modeling natural language meaning has in recent years found renewed interest, fuelled by progress in image recognition, image captioning and Visual Question Answering using deep learning methods (Antol et al., 2015; You et al., 2016; Monroe et al., 2016). Many models for tackling these tasks rely explicitly or implicitly on the idea of modelling meanings of perceptual words (i.e. words describing perceivable properties of ob-

jects and situations) as classifiers. In the abstract, one can define classifiers as functions that correspond to a natural language phrase or sentence, and that take perceptual input (e.g. from a camera) and decide to what degree the phrase or sentence describes the perceptual input.

$$\textit{Classifier} : \textit{PerceptualData} \rightarrow [0, 1]$$

If the possible degrees are limited to 0 and 1, we have a binary classifier; if it can be any real number between 0 and 1, we have a continuous (e.g. probabilistic) classifier. Following Schlangen et al. (2016), we refer to the idea of modelling linguistic meanings using classifiers as the *words-as-classifiers* approach.

Several approaches have been proposed for modelling meanings of (some) words as classifiers of perceptual (often visual) data (Dobnik, 2009; Schlangen et al., 2016; Monroe et al., 2016; McMahan and Stone, 2015; Larsson, 2015; Fernández and Larsson, 2014; Schlangen et al., 2016; Larsson, 2017). Various types of classifiers have been used, and the output of the classifiers has been rendered and connected to language in various ways (see Larsson (2017) for an overview). This approach has also been used in work on visual question answering (Andreas et al., 2016).

3 Classifiers and formal semantics

Marconi (1997) distinguishes inferential and referential meaning. *Inferential* word meanings enable inferences from uses of the word. Such meanings are sometimes referred to as "high level" or "symbolic", and are typically modelled in formal semantics. *Referential* meaning, on the other hand, allows speakers to identify objects and situations referred to. Referential meaning is sometimes referred to as "low-level" or "subsymbolic". Our working hypothesis is that referential meaning can be modelled using classifiers that output formal representations, thus connecting "high level" formal representations to "low level" perceptual information, and in this we follow Larsson (2011) and Larsson (2015). This is a way of addressing the *symbol grounding problem* put forward by Harnad (1990) in a way that is compatible with formal semantics.

The crucial step in making use of classifiers in formal semantics is to regard them as (parts of) *representations of intensions* of linguistic expressions (Larsson, 2015). Traditionally, the intension of an expression helps determine whether some item belongs to the extension of the expression. Here, this translates to using a classifier to help determine whether some perceptual data derived from some item can be used to classify that item as falling under the expression, i.e., to be included in its extension.

The idea of regarding classifiers as representing intensions is related to proposals by Muskens (2005) and Lappin (2012) to identify the intensions of an expression with an algorithm (implemented using logic programming or a functional programming language, respectively) for determining its extension. The idea of representing referential meanings as classifiers can perhaps be regarded as an application of this general idea to (although we are here using general notation for functions rather than any specific programming language).

4 Classifiers and TTR

Larsson (2020), following Cooper (2019), presents a version of Type Theory with Records which places classifiers at the core of formal semantics, and shows the role of classifiers in deciding whether (or to what extent) an utterance content (an utterance meaning interpreted in context) correctly describes a perceptually available situation (such as a visual scene). See Appendix A for a brief introduction to TTR. The technical details of Larsson (2020) are not important for our present purposes, but are included to make more concrete what we mean when we talk about modelling intensions as classifiers.

The core definition linking meaning to classifiers is

(1) $s : T$ iff $\textit{Clfr}(T)(s) = T$

where s is a situation being classified, T is a ptype (representing the content of an utterance), and *Clfr* is a classifier function associated with T.

As an illustration, Larsson (2011, 2015, 2020) uses a simple dialogue game called the left-or-right (LoR) game. In this game, one agent (A) places objects on a square surface, and the other agent (B) classifies these objects as being to the right or not. In first language acquisition, training of perceptual meanings typically takes place in situations where the referent is in the shared focus of attention and thus perceivable to the dialogue participants. It is assumed that the dialogue participants are able to establish a shared focus of

attention. A (simple) sensor collects some information (sensor input) from the environment and emits a real-valued vector. The sensor is assumed to be oriented towards the object in shared focus of attention.

Larsson formalises the notion of a simple perceptron classifier and provide its TTR type. Whereas a (non probabilistic) classifier normally gives a Boolean output (corresponding to whether the neuron triggers or not), we want as output a ptype (or the negation thereof):

(2) $\pi_{right} : \begin{bmatrix} \text{w}:\mathbb{R}^+ \\ \text{t} :\mathbb{R} \end{bmatrix} \rightarrow \begin{bmatrix} \text{foo}:\mathit{Ind} \\ \text{sr} :\mathbb{R}^+ \end{bmatrix} \rightarrow \mathit{Type}$

such that if

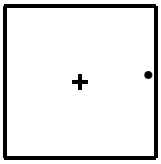
- par: $\begin{bmatrix} \text{w} & : & \mathbb{R}^+ \\ \text{t} & : & \mathbb{R} \end{bmatrix}$ and
- $r : \begin{bmatrix} \text{foo} & : & \mathit{Ind} \\ \text{sr} & : & \mathbb{R}^+ \end{bmatrix}$,

then $\pi_{right}(par, r) =$

$$\begin{cases} \text{right}(r.\text{foo}) & \text{if } r.sr \cdot par.\text{w} > par.\text{t} \\ \neg\ \text{right}(r.\text{foo}) & \text{otherwise} \end{cases}$$

Here, par is a record containing classifier parameters; for a perceptron, a weight vector and a threshold. The second argument to the classifier is the situation to classify, which needs to be of a type specifying a sensor reading (sr) and an object in the focus of attention (foo). Note that the function itself is defined outside TTR. This allows any classifier to used with TTR, no matter how it is implemented.

Assume that an agent A places an object on the surface and says "That one is to the right", or just "Right".

(3) 

Agent B watches and gets a position sensor reading $\begin{bmatrix} 0.900 & 0.100 \end{bmatrix}$ which is part of B's take on the current situation (s_1):

B now interprets A's utterance in the context the situation s_1 by computing $[\![\text{right}]\!](s_1)$, which gives the result $[\![\text{right}]\!](s_1) = \text{right}(obj_{45})$.

(4) $s_1 = \begin{bmatrix} \text{sr}_{\text{pos}} & = & \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \\ \text{foo} & = & obj_{45} \end{bmatrix}$

Next, B decides if A's utterance correctly describes (her take on) the situation, i.e. if

(5) $s_1 : [\![\text{right}]\!](s_1)$, i.e., if $s_1 : \text{right}(obj_{45})$

For

- T=right(obj_{45}),
- Clfr(right(obj_{45}))=π_{right}, and
- par=$\begin{bmatrix} \text{w} = \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} \\ \text{t} = 0.090 \end{bmatrix}$,

we get

(6) $s_1 : \text{right}(obj_{45})$ iff $\pi_{\text{right}}(par)(s)=$ right(obj_{45})

As shown in the partial derivation in Figure 1 (and in more detail in Larsson (2020)), the RH of this equation holds, and hence B can conclude that s_1 is indeed appropriately described by A's utterance.

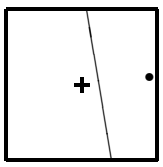
5 Semantic coordination and classifiers

In Larsson (2011) and Larsson (2015), the idea of intensions as classifiers is combined with a notion of semantic coordination — the process of interactively agreeing on the meanings of words and expressions, and (simultaneously) agreeing on which words are appropriate in a given context. Shared meanings (modelled as intensions) are achieved by agents interactively coordinating their respective takes on those meanings, which involves training classifiers based on input from dialogue interaction.

The process of semantic coordination displays a fundamental dialectic between extensional and intensional meaning, in the following way: An individual or situation s is claimed (explicitly or implicitly) by a dialogue participant to be in the extension of (i.e., to be adequately referred to by) an expression e. As a result, the other dialogue participant updates her intensional meaning (classifier) of e based on s. Later, the learner will apply this intensional meaning (classifier) to new situations to determine whether they are in the extension of e.

As an example, we may imagine two rounds of the left-or-right game playing out as in (7).

(7) A: "(the object is to the) right"

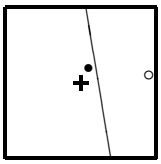


$$\pi_{\text{right}}(par)(s) = \text{right}(obj_{45}) = \pi_{\text{right}}(\begin{bmatrix} \text{w} = \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} \\ \text{t} = 0.090 \end{bmatrix})(\begin{bmatrix} \text{sr}_{\text{pos}} = \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \\ \text{foo} = obj_{45} \end{bmatrix}) =$$

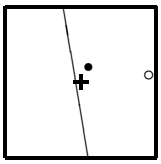
$$(\begin{cases} \text{right}(obj_{45}) & \text{if } \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \cdot \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} > 0.090 \\ \neg\,\text{right}(obj_{45}) & \text{otherwise} \end{cases}) = \text{right}(obj_{45})$$

Figure 1: Example classification derivation

B: "okay"
A: "(the object is to the) right"



B: "aha"



In round 1, B comes to the conclusion that A's utterance matches B's own classification (with the threshold between "left" and "right" shown as a diagonal line). However, in the second round, B's initial classification does not match A's utterance. We imagine that as a result of this, B retrains his classifier that represents the perceptual meaning of "(the object is to the) right" for B. Concretely, this amounts to applying the perceptron training rule to obtain an updated parameterisation (see Larsson (2015) for details) of the classifier, thus updating B's take on the meaning of "right" in the game.

We have now seen a simple example of how perceptual meaning, modelled as a classifier, can be learnt from conversational interaction in a shared perceptual environment. When humans interact they *reciprocally* (Fernández et al., 2011) adapt to each others' language use on multiple levels. Although the left-or-right game models one-sided learning rather than coordination, it could quite easily be altered to illustrate coordination directly. For example, by letting A and B switch roles after each round. In this symmetric left-or-right game, the agents would converge on a meaning of "right" that neither of them may subscribe to initially.

6 Classifiers and Possible Worlds

This view of the nature of linguistic meaning differs in several important respects from the standard view of possible worlds semantics (PWS, as put forward in Montague (1974) and many others):

- Intensions are represented independently from extensions (in the form of classifiers).
- Extensions are *derived* indirectly from intensions by applying classifiers to (takes on) situations, rather than being directly *given*.
- Intensions are *dynamic* and can be revised as a result of interaction (semantic coordination).

We also take it that as a result of semantic coordination, meanings are *intersubjective*, i.e. shared between communities or dyads of speakers, rather than objective in the sense of being independent of individual agents (similar to mathematical concepts).

In Barker (2002) and Lassiter (2011), dynamic extensions are achieved by a notion of possible languages (parallel to the notion of possible worlds), each specifying the extensions of linguistic expressions. By excluding possible languages, one modifies the range of possible extensions and gets successively closer to the actual language. As many others, Barker takes extensions to be at least in part subjective (or else it would not be possible to learn new things about extensions from other agents).

We can perhaps briefly map out a couple of different positions one may take with respect to the nature of linguistic intensions and extensions (see Figure 2. Firstly, extensions may either be considered as static or as dynamic, i.e. up for revision. Second, extensions can be considered as directly given, or as (somehow) indirectly derived from intensions, e.g. by applying intensions (modelled as classifiers) to (takes on) situations or the world. Third, extensions can be regarded as objective (agent-independent) or subjective (agent-dependent). In the subjective category we include intersubjective extensions, i.e. extensions shared among a collective of agents (but not independent of those agents).

7 Indeterminate extensions

We may now ask, if the extensions of at least some natural language expressions are regarded as dy-

Extensions	static/dynamic	given/derived	objective/(inter)subjective
PWS (Montague)	static	given	objective
PWS (Barker)	dynamic	given	(inter)subjective(?)
TTR	dynamic	derived	(inter)subjective

Figure 2: How three different approaches to meaning treat extensions

namic and derived from intensions modelled as classifiers, are they still *determinate* in the sense that *there is at any given time a single extension for each expression* in a language? This, after all, is one of the core assumptions of standard possible worlds semantics, where meanings are modelled as extensions in a model. We would argue that they are not, based on three related arguments. First, however, we want to establish a few more or less trivial facts about classification and coordination.

7.1 Some observations about classification and coordination

First of all, classifiers **generalise over examples**. They are trained on examples and their job is to generalise over these examples so that they can classify previously unseen individuals and/or situations. Normally, the different speakers of a language are exposed to different sets of examples that they train their classifiers on. This means that even if speakers generally agree how to classify a majority of cases, there may be (actual or potential) borderline cases where they (would) make different judgements. In addition, even if two agents are exposed to the same training data, they not make identical generalisations from that data. Another way of putting this is that the classifier is underdetermined by the training data.

Second, classification is **sensitive to noise**. For borderline cases especially, noise coming e.g. from perception may affect the result of classification. In the left-or-right game, for example, noise may affect the perception of the object on the surface and the detection of its precise location, which in borderline cases may determine the outcome of classification.

Third, classification is not abstract, but a concrete process that is physically manifest and **extended in time**.

Finally, since meanings (modelled as classifiers) are coordinated in a community, meanings (and their associated classifiers) are always **provisional and potentially up for revision**. Note that this is a fact about *potential*, so that even a word whose meaning is unchanged for a very long time still has the potential to have its meaning be called into question and become subject to revision.

7.2 The time argument

With this in mind, let us proceed to our first argument, the *time* argument. This relies on two of our assumptions above: classification is a process that takes time, and meanings are always provisional. Given that an agent has classified some situation as being correctly described by an expression e, who is to say whether this is correct or not? If we accept that linguistic meanings are conventional (De Saussure, 1989), the answer is: the speech community that she is part of. The task of coordinating our judgements, i.e. agreeing on how to talk about the world, is therefore vital for the very existence of linguistic meaning, and relies crucially on interactively coordination on meanings (semantic coordination).

Could we then perhaps talk about a determinate extension once a community has come to a final agreement about how to use a word? Well, how could we (or they) tell that they have? Only by successively going through all potentially relevant items (situations or entities) and seeing if they agree. If the notion of extensions in possible worlds (including, presumably, the actual world) is to be taken seriously (which it possibly should not, see Lappin (2012)), this means going through all objects in the universe (which first of all requires individuating all physical matter into discrete objects). Whether the extension is universal or situation-specific, this is a process that will take some time, and what guarantee is there that no single member of the community during this time will modify their take on the meaning of e?

What this argument purports to show is that there is no method that *guarantees* shared extensions, and hence that the idea of meaning as a determinate (objective or intersubjective) extension does not reflect the reality of language use. This does not mean communication is impossible, since humans are quite good at clearing up misunderstandings

and interactively coordinating on meanings that are sufficiently shared for the purposes of their interaction. The argument also does not show that humans can never achieve intersubjectively shared extensions (see also Section 9), only that this is not the general case.

7.3 The generalisation argument

The second argument (based on *generalisation*) goes as follows: given some dataset of entities referred to using an expression e, there are several generalisations that adequately describe it (Quine, 1960). Different agents may agree on all data in training set but have subtly different generalisations. New situations or entities may be encountered which are different from anything in the training set, and in such cases different agents may make make different judgements, even if they agree on everything in the (previous) dataset. This means that we cannot be sure that agents are coordinated on the extension of e unless the dataset contains all relevant entities/situations in the world, which in the worst case is all entities/situations in the world. So the procedure for arriving at a generalisation that ensures a determinate shared extension cannot realistically be executed.

It may be argued that this is just the familiar problem of concept learning, which must be faced by any theory of meaning. We have no problem with that; what we try to show is that *if* one decides (wisely, on our view) to model concrete meanings using classifiers (to account for semantic learning and coordination), then one cannot also keep the notion of determinate extensions of natural language words and expressions.

7.4 The noise argument

The outcome of acts of classification are not decided in advance – if they were, there would be no point in classifying anything. However, the idea of actually classifying everything of a certain category does not seem practically feasible. But if it is not possible to actually classify everything, perhaps we can rely on a counterfactual definition of extensions? Something along the lines of "if all members of the community were to review all entities/situations, they would all agree that the extension is . . .".

This brings us to our third argument, the *noise* argument. Extensions are derived from intensions modelled as classifiers, which take noisy real-world perceptual data as input, which means that noise is an inherent feature of real-world classification procedure. Any counterfactual definition would therefore have to rely on a notion of noise-free classification, which in turn would relies on some capability of deciding what counts as noise, and what does not. However, this leads to an infinite regress since the classification of noise is just another classification problem, sensitive to the same arguments the original classification problem.

The noisiness of classification is related to vagueness, but we would argue vagueness is a separate problem. Even if vagueness is dealt with in an extensional semantics, e.g. by introducing some notion of probabilistic set membership (e.g. fuzzy sets (Hersh and Caramazza, 1976)), there is still a question of whether a certain outcome of observation and classification can even in principle be reliably repeated on repeated classifications, since this requires excluding noise from classification, which in turn requires classifying noise and so on *ad infinitum*.

Another possible counter-argument is that even in physics, many foundational concepts cannot be measured without noise, but this does not disqualify them. However, this misses the point: we are not disqualifying object-level concepts as such (on the grounds of being sensitive to noise, or indeed on any grounds), but the idea that they – in the general case – can be modelled as determinate extensions.

8 What about non-concrete words?

So far, we have talked only about concrete words (i.e. words referring to observable reality). What about non-concrete words? We take it that (following, among others, Carnap (1998)) at least many abstract words are ultimately grounded in experience. Take for example *democracy*, a paradigm abstract term which nevertheless is indirectly observable – as shown by the practice of sending international observers to an election with the task of deciding if and to what degree an election is democratic.

9 When are extensional semantics applicable?

Despite the problems noted above for a purely extensional semantics,from the perspective of a general theory of language use, there may be certain types of activities in which it is nevertheless applicable in the sense that extensions can be treated as given and determinate. These are activities where classification procedures are explicitly oper-

ationalised and agreed upon, such as in the natural sciences when they are in their "normal" (intra-paradigmatic) state of accumulating knowledge according to a more or less fixed conceptualisation. Another example is activities that are sufficiently routinised to be susceptible to formalisation, corresponding to what Dreyfus (1992) refers to as *systematic domains* – a set formal representations that can be used in a e.g. a dialogue system, and that embodies a static, intersubjective and uncontroversial interpretation of the situation in which the system will function.

It may be speculated that it is these and similar special cases of language use that underlie the intuitive attraction of extensional semantics. However, insofar as one is interested in a *general* theory of natural language meaning, it is not sufficient to handle only some particular types of language use. Indeed, even within the natural sciences, semantic change and coordination frequently occurs, as exemplified by Ludlow (2014) using the case of Pluto which was re-classified from being a planet to no longer being a planet. Also in everyday activities, semantic coordination is commonplace (Myrendal, 2015, 2019).

In addition to the cases mentioned, in any specific situation, agents may succeed in jointly classifying all relevant objects (and situations) into shared extensions using their individual classifiers together with protocols for semantic coordination (Larsson, 2018) . As long as these classifications are done sufficiently explicitly to become grounded, the extensions are, at least for all practical purposes, determinate and shared. Regardless of how this is achieved (whether through routinisation, scientific operationalisation or semantic coordination in a specific situation), these determinate extensions could in principle theoretically be treated using a purely extensional semantics.

In TTR such extensions are modelled as *witness caches* that can replace classifiers once (situation-specific) determinate extensions have been achieved (Larsson, 2020). This is in line with the idea of "natural language as a toolkit for building formal languages" (Cooper and Ranta, 2008). Reflecting this, the actual core definition in (Larsson, 2020) is not (1) but (8):

(8) $s : T$ iff $\mathit{Clfr}(T)(s) = T$ or $s \in F(T)$

where $F(T)$ is the witness cache, for type T – a set of situations (in the case of ptypes) previously judged to be of type T. On our present account, the witness cache for a type and an agent could represent mutually (within a community) agreed-upon classifications in a systematic domain. In the event of semantic coordination and updates to classifier parameters, the witness cache would most likely need to be cleared to avoid inconsistencies. To build a new witness cache, the community would again need to go through the process of jointly agreeing on a revised systematic domain, including extensions of central concepts.

10 Misclassification

If the correctness of any classification cannot be determined with reference to some absolute truth, then what does it mean to misclassify something? This question touches on the long-standing problem of rule-following and semantic normativity in the philosophy of language (Wittgenstein, 1953; Kripke, 1982; Boghossian, 1989; Glüer and Pagin, 1998).

In short, our answer is that it depends on what is being classified – specifically, whether there is an agreed-upon systematic domain (providing extensions) or operationalisation agreed upon by experts and deferred to by those not competent or willing to carry out the operationised classification procedure (as in the natural sciences). If there is, misclassification means classifying differently than the generally accepted classification. (However, note that to the extent that counterfactuality is involved here, in the form of assumptions of how something *would* be classified by e.g. an expert, the noise argument may still be applicable but to a lesser extent; operationalisation is no complete guarantee for agreement and noise may still creep in.)

If there is no operationalisation or systematic domain, misclassification (e.g. in the case of many artifacts or social constructions such as democracy) can mean simply classifying differently than others in the community (whose classification prevails). In the case of artifacts, the creator(s) of the artifact (if such a person or group can be singled out) may have more of a say than others concerning the correct classification of the artifact, similar to the case of experts in science. Another possibility is that an individual at time t classifies something x as a but later classifies the same thing as b and comes to regard the latter classification as correct; in this case they can say "I was wrong in classifying x as a, it is actually b. An there are probably more

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

Figure 3: Schema of record and record type

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

Figure 4: Sample record and record type

variants of misclassification.

11 Summary and conclusion

We explored some consequences of idea of using classifiers to model intensions of natural language expressions for the notion of extensional meaning. Three arguments (the time argument, the generalisation argument, and the noise argument) are offered, on the basis of some observations about classifiers, why extensional semantics is not well suited for modelling perceptual meanings, especially if the latter are regarded as classifiers, and hence why extensional semantics does not (and indeed cannot) provide a general account of natural language meaning. We also discussed cases where nevertheless extensional semantics may be adequate, and the applicability of the argument to non-concrete words.

In future work, we would like to connect the arguments made here more explicitly to the wealth of work on rule-following and semantic normativity (Wittgenstein, 1953; Kripke, 1982; Boghossian, 1989; Glüer and Pagin, 1998). However, the best argument for any theory of natural language semantic is not to be found in polemic debates but in empirical coverage, computational tractability, and overall usefulness in scientific and practical matters. Hence, developing the idea of classifiers in natural language semantics further and showcasing its full potential is and remains our main area of future work.

Acknowledgements

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

A Appendix: TTR

We can here only give a brief and partial introduction to TTR; see also Cooper (2005) and Cooper (2012). To begin with, $s : T$ is a judgment that some s is of type T. To make explicit who is making this judgment, the of-type relation may be subscripted with an agent A, as in $:_A T$. One *basic type* in TTR is Ind, the type of an individual; another basic type is $\mathbb{R}$, the type of real numbers. Given that T_1 and T_2 are types, $T_1 \rightarrow T_2$ is a *functional type* whose domain is objects of type T_1 and whose range is objects of type T_2.

Next, we introduce *records* and *record types*. If $a_1 : T_1, a_2 : T_2(a_1), \dots, a_n : T_n(a_1, a_2, \dots, a_{n-1})$, where $T(a_1, \dots, a_n)$ represents a type T which depends on the objects $a_1, \dots, a_n$, the record to the left in Figure 3 is of the record type to the right.

In Figure 3, $\ell_1, \dots \ell_n$ are *labels* which can be used elsewhere to refer to the values associated with them. A sample record and record type is shown in Figure 4.

Types constructed with predicates may be *dependent*. This is represented by the fact that arguments to the predicate may be represented by labels used on the left of the ':' elsewhere in the record type. In Figure 4, the type of c_{man} is dependent on ref (as is c_{run}).

If r is a record and ℓ is a label in r, we can use a *path* $r.\ell$ to refer to the value of ℓ in r. Similarly, if T is a record type and ℓ is a label in T, $T.\ell$ refers to the type of ℓ in T. Records (and record types) can be nested, so that the value of a label is itself a record (or record type). As can be seen in Figure 4, types can be constructed from predicates, e.g.,

"run" or "man". Such types are called *ptypes* and correspond roughly to propositions in first order logic. Given a set of predicates and a set of possible arguments, the set of possible ptypes is $\mathbf{PType}$, thus allowing for polymorphic predicates. The arity of a ptype P is a set of tuple of types $Arity(P)$. For example $Arity(\text{run}) = \{\langle Ind \rangle\}$.

A fundamental type-theoretical intuition is that something of a ptype T is whatever it is that counts as a proof of T. One way of putting this is that "propositions are types of proofs". In Figure 4, we simply use prf(T) as a placeholder for proofs of T; below, we will show how low-level perceptual input can be included in proofs.

Semantic phenomena which have been described using TTR include intensionality and mental attitudes (Cooper, 2005, in progress), dynamic generalised quantifiers (Cooper, 2004), modality (Cooper, in progress), vagueness (Fernández and Larsson, 2014), co-predication and dot types in lexical innovation, frame semantics for temporal reasoning, reasoning in hypothetical contexts (Cooper, 2011), enthymematic reasoning (Breitholtz and Cooper, 2011), clarification requests (Cooper, 2010), various kinds of negation (Cooper and Ginzburg, 2011), and information states in dialogue (Cooper, 1998; Ginzburg, 2012; Cooper, in progress).

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*.

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

C. Barker. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy*, 25(1):1–36.

Paul A Boghossian. 1989. The rule-following considerations. *Mind*, 98(392):507–549.

Ellen Breitholtz and Robin Cooper. 2011. Enthymemes as rhetorical resources. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011)*, pages 149–157, Los Angeles (USA).

Rudolf Carnap. 1998. *Der logische aufbau der welt*, volume 514. Felix Meiner Verlag.

Robin Cooper. 1998. Information states, attitudes and dependent record types. In *ITALLC98*, pages 85–106.

Robin Cooper. 2004. Dynamic generalised quantifiers and hypothetical contexts. In *Ursus Philosophicus, a festschrift for Björn Haglund*. Department of Philosophy, University of Gothenburg.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3:333–362.

Robin Cooper. 2010. Generalized quantifiers and clarification content. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, Poznań. Polish Society for Cognitive Science.

Robin Cooper. 2011. Copredication, quantification and frames. In *LACL*, volume 6736 of *Lecture Notes in Computer Science*, pages 64–79. Springer.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Robin Cooper. 2019. Types as learnable cognitive resources in pyttr. In Cleo Condoravdi and Tracy Holloway King, editors, *Tokens of Meaning: Papers in Honor of Lauri Karttunen*, pages 569–586. CSLI Publications.

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

Robin Cooper and Jonathan Ginzburg. 2011. Negation in dialogue. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011)*, Los Angeles (USA).

Robin Cooper and Aarne Ranta. 2008. Natural languages as collections of resources. In Robin Cooper and Ruth Kempson, editors, *Language in flux: relating dialogue coordination to language variation, change and evolution*. College Publications, London.

Ferdinand De Saussure. 1989. *Cours de linguistique générale*, volume 1. Otto Harrassowitz Verlag.

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.

Hubert Dreyfus. 1992. *What computers still can't do*. The MIT Press.

Raquel Fernández and Staffan Larsson. 2014. Vagueness and learning: A type-theoretic approach. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (∗SEM 2014)*.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, Barcelona, Catalonia, Spain.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford University Press, New York.

Kathrin Glüer and Peter Pagin. 1998. Rules of meaning and practical reasoning. *Synthese*, 117(2):207–227.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1990):335–346.

Harry M Hersh and Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 105(3):254.

Saul A. Kripke. 1982. *Wittgenstein on Rules and Private Language : An Elementary Exposition*. Harvard University Press.

Shalom Lappin. 2012. An Operational Approach to Fine-Grained Intensionality. *UCLA Working Papers in Linguistics, Theories of Everything*, 17(2):180–186.

Staffan Larsson. 2011. The ttr perceptron: Dynamic perceptual meanings and semantic coordination. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011)*, Los Angeles (USA).

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369. Published online 2013-12-18.

Staffan Larsson. 2017. Compositionality for perceptual classification. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in cognitive science*, 10(2):389–408.

Staffan Larsson. 2020. Discrete and probabilistic classifier-based semantics. In *Proceedings of PaM, Probability and Meaning*.

Dan Lassiter. 2011. Vagueness as probabilistic linguistic knowledge. In R. Nowen, R. van Rooij, U. Sauerland, and H. C. Schmitz, editors, *Vagueness in Communication*. Springer.

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press.

Diego Marconi. 1997. *Lexical competence*. MIT press.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. *arXiv preprint arXiv:1606.03821*.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven. Ed. and with an introduction by Richmond H. Thomason.

Reinhard Muskens. 2005. Sense and the computation of reference. *Linguistics and Philosophy*, 28(4):473–504.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. Ph.D. thesis, University of Gothenburg.

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

Willard Van Orman Quine. 1960. *Word and Object*. MIT Press.

F Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408.

David Schlangen, Sina Zarrie, 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 (ACL 2016)*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell Ltd.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

THERE IS AS YET INSUFFICIENT DATA FOR A MEANINGFUL ANSWER*

Maria Boritchev
LORIA, UMR 7503,
Université de Lorraine, CNRS, Inria,
54000 Nancy, France
maria.boritchev@loria.fr

Maxime Amblard
LORIA, UMR 7503,
Université de Lorraine, CNRS, Inria,
54000 Nancy, France
maxime.amblard@univ-lorraine.fr

Abstract

Formal studies of discourse raise numerous interrogations on the nature and the definition of the way consecutive sentences combine with one another. The shift from discourse to dialogue brings forward even more specific issues. Dialogue acts are more intrinsically connected because of the dynamicity of the interaction. This article introduces a proof of concept of a formal compositional treatment of the relationship between consecutive utterances. Starting from neo-Davidsonian event semantics, we propose to use the *relative response set* as an intermediate set tool that allows us to define notions of question-answer correspondence, model the effect of clarification requests on previous utterances and compute semantic representations of dialogue interactions.

1 Introduction

Dialogue is governed by both implicit and explicit, quite complex rules of interaction. It allows us, dynamically, to share an important informational content by using relatively simple utterances. While we do participate in dialogues, we do not fully understand the underlying mechanisms allowing the combination of dialogue acts: what makes dialogue acts compatible, coherent. In this article, we present an analysis of dialogue interactions. First, we focus on questions and answers which are at the root of the difference between dialogue and discourse. What is a question? What is an answer? What makes an answer satisfactory? How to define question/answer relationship in a precise, computational way?

It is relatively easy to define *questions*, as dialogue acts, syntactically, at least as utterances marked with one of the question marks/a rising intonation/a question particle/a question word. Semantically, a question is a sentence or phrase used to find out information[1]. Defining what constitutes an *answer* and how it fulfils its answering role is way more complex. Sometimes, the answer to a question is an entity, one element of information. Sometimes, the answer defines or shapes a set of entities or elements of information, but is not sufficient to choose an element of this set. Still, we answer questions daily, and there is an extensive literature about identifying answers to questions (Traum, 1994; Ginzburg and Sag, 2000; Rao et al., 2016; Lai et al., 2019; Yao et al., 2013). After a long period of expert systems, recent developments are largely rooted in neural networks and word embeddings (Mikolov et al., 2013; Österlund et al., 2015).

The later approaches integrate little or no linguistic descriptions of dialogue itself. One of the main reasons for this is that formal or computational models of dialogue are scarce. We went through the literature on questions and answers throughout the different existing formalisms based on formal semantics (AMR (Banarescu et al., 2013), MRS (Copestake et al., 2005), *etc.*), see section 2. The developments of *Type Theory with Records* (Cooper and Ginzburg, 2015) give us one of the most advanced current models for dialogue. This formal approach roots its representation in complex type theory.

Linguistic studies of dialogue based on real-life data show evidence of a frequently appearing type of utterance: *clarification request* (CR). They are designed to help the dialogue continue even when there are lapses in the process of exchanging conversational content. (Purver, 2004) defines clarification requests as "device[s] allowing a user to ask

* https://en.wikipedia.org/wiki/The_Last_Question

[1] Definition based on https://dictionary.cambridge.org/dictionary/english/question, visited on May 8th 2020.

about some feature (e.g. the meaning or form) of an utterance, or part thereof". Clarification requests can be used for different metacommunicational reasons (Ginzburg, 2012):

- To ask for a repetition or to request clausal confirmation: a confirmation question concerning the semantic contribution of a particular constituent within the entire clausal content of the utterance;
- To confirm intended content: query on a content a speaker intends to associate with a given (sub-)utterance, independently of the remaining content of the clause being resolved or not;
- To correct the other speaker.

We want to account for clarification request phenomena in dialogue in a computational (logic, compositional, dynamic) way. In particular, we are interested in the action of clarification requests on the previous utterance. Corpus observations conducted on the British National Corpus (BNC, see (Burnard, 2000)) list several types of clarification requests (Purver et al., 2003), see table 2. Computationally speaking, how different is the action of those different types of clarification requests?

Computational approaches of dialogue as opposed to ones for discourse run into very specific difficulties. We find particular phenomena such as sequences of questions, simultaneous presence of an information and its opposite, along with usage of Non-Sentential Utterances. Beyond mechanical aspects of dialogue, one needs to consider meta-levels: links with argumentation (argument-mining, topoi studies), see for example (Breitholtz, 2014), along with simultaneous dialogues management (Asher et al., 2016a).

Following the neo-Davidsonian event semantics representation of declarative sentences introduced by (Champollion, 2011) and the semantic representation of questions introduced by (Boritchev and Amblard, 2019a), we define the *relevant response set* with respect to a question Q, $\mathscr{R}_Q$. This set constitutes an intermediate computational step for dialogue representation.

We use the taxonomy of answer types introduced in (Ginzburg, 2012) and $\mathscr{R}_Q$ to formally characterize different types of answers and interpret the capacity of answerhood of a given utterance to a given question, see section 5. To do so, we adapt the modern type theoretical definitions given in (Ginzburg, 2012) to our montagovian-like approach. We then extend our model using the taxonomy of clarification requests introduced in (Purver, 2004) in order to model articulation of utterances in dialogue. We use examples from the DinG (*Dialogue in Games*) corpus[2], under construction (Boritchev and Amblard, 2018), to test our formal approach. This corpus is composed of transcriptions of recordings of people playing the *Catan* board game, in French. The participants are designated by the colour of their tokens: **Red**, **White**, **Yellow**, **Blue**. This game is also used in the STAC (*Strategic Conversation*) corpus (Asher et al., 2016b), which gathers logs of virtual games in English. (Hunter et al., 2018) proposes a detailed study of the written dialogue interactions based on (Asher et al., 2003). One of the future developments of DinG lies in testing the proposed logical underpinnings at a multilingual and multi-modal level.

2 Dialogue Models

There is an extensive litterature studying dialogue, (Grice, 1975), (Austin, 1975), (Freed, 1994), just to cite a few examples. These approaches mostly come from linguistics and philosophy. The computational aspect of dialogue is largely discussed in applications such as chatbot programming or information retrieval, where the main focus is semantic content research. The computational linguistics point of view, that we defend, lacks of an operational, computational description of dialogue through the study of dialogue acts and the way they combine, rooted in real-life data. The work we present in this article follows the tradition of formal semantics account of interrogatives as set by (Groenendijk et al., 1997), (Ginzburg, 1995a,b) and the structured meaning approach as presented in (Krifka, 2001). We work on real-life data that comes from the DinG corpus because it gives us both straightforward and complex examples, while being quite restrained in topic and vocabulary, which simplifies the computations.

Among formal, computational linguistics approaches, one can find Abstract Minimal Representation (AMR). In this formalism, questions are marked with a label: `amr-unknown`. Polar questions are represented as declarative sentences with an `amr-unknown` in the `:polarity` field, corresponding to the idea that one doesn't know the

[2] `https://gitlab.inria.fr/amblard/ding` for data. Frequent updates to be expected.

truthfulness of the sentence[3]. AMR doesn't provide deep means of representation for quantifiers, which is not a problem if one wants mainly to represent what is going on in a dialogue, but constitutes one for representation based logical reasoning. On the linguistics side, there is no native way to represent clarification requests (CRs) because AMR wasn't specifically designed for dialogue.

Some works from Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) also approach dialogue from a perspective similar to the one for discourse semantics. (Amblard et al., 2015) presents an extension of SDRT for dialogue analysis that introduces a horizontal relation for question-answer relationship. (Xuereb and Caelen, 2004) uses SDRT for Human-Machine interaction tasks. These are simplifications of a theory that is not dedicated to dialogue. Though particular properties can be put forward in this way, it doesn't present a native way to model questions and answers.

Minimal Recursion Semantics (MRS) is a framework with complex descriptions for representation and reasoning (Egg, 1998; Yao and Zhang, 2010). It contains numerous rules with elaborate articulations. MRS allows question representation but does not natively account for the partial aspects of answers, that are prevalent in dialogue. Non-Sentential Utterances (NSU) constitute around 10% in English speaking corpora such as BNC (Fernández et al., 2007). NSUs can appear under various forms and for different types of illocutionary use. (Schlangen and Lascarides, 2003) presents an MRS and SDRT based treatment of NSUs. The approach presented in our article takes care of two types of NSUs: short answers (*"yes"*, *"no"*, *"stone"*, etc.) and CRs. We plan on developing our approach by adapting ideas from (Schlangen, 2003).

DIT++ (Bunt, 2009) is a semantics-inspired framework for human dialogue annotation and analysis, that integrates relations coming from SDRT. This framework provides a very detailed and precise classification of questions and answers. Clarification requests can be analysed in this framework, though they are dissolved among the other dialogue acts (not a category on their own). DIT++ is not a computational model, it is not designed for reasoning on the representations. Yet, any computational model can thrive on annotations provided by this framework. In particular, the representations presented in our article can and should be built using DIT++ annotation schema.

(Cooper and Ginzburg, 2015) presents KoS, an approach deeply rooted in linguistics and based on Type Theory with Records. It is oriented towards dialogue management modeling and revolves around a context-centered view of the dialogue. Each speech turn is viewed as a function modifying the interlocutor's context, as in dynamic semantics. Representations of dialogues in KoS mostly describe situations, in order to produce a big picture of the interaction. However, there is no use of syntax-semantics interface in KoS, thus it isn't easy to set a compositional construction of a model for utterances starting from representations of individual components of the sentences. Still, KoS accounts for different types of dialogue acts and for NSUs.

3 The Relevant Response Set

Sentences can be represented in a logical, compositional and dynamic way (Cresswell, 1976). We rely on these notions to grow and define the representation of utterances and the concept of *relevant response set*. We assume that the latter helps to materialize the implicit mechanisms going on in dialogue, in particular for notions of compatibility between two successive utterances.

First, we introduce the representations that we use. As we do not have the space here to broadly present and explain all the theories, we will use examples in addition to pointers to extensive references.

Our goal is to produce semantic representation (SR) of utterances. Several strategies can be used to achieve this; our approach is part of a compositional vision of treatments. As we want to keep the process generic, we follow a neo-Davidsonian representation of semantics (Maienborn, 2011). (Boritchev and Amblard, 2019a) offers a process to compute compositional representations of questions by using neo-Davidsonian event semantics formalisms (Champollion, 2011).

To each declarative sentence S we associate the first order logic formula $[\![S]\!]$, following (Champollion, 2015). This gives us the possibility to compute the semantic representation (SR) of utterances which syntactically behave as declarative sentences: some of the assertions, some of the answers.

[3]Source: https://github.com/amrisi/amr-guidelines/blob/master/amr.md#questions at 07.06.20.

A polar, or *yes/no*-question requests a confirmation or a denial of its informational content. Therefore, one can see it as a declarative sentence (with the same content) that will be either directly accepted or negated and then accepted, depending of whether the answer is *"Yes"* or *"No"*. We denote the declarative content of a polar question Q_p as $decl(Q_p)$.

Definition 1 (SR of a polar question) *Consider Q_p, a polar question. The* semantic representation *of Q_p, following (Boritchev and Amblard, 2019a), is the semantic representation of Q_p's declarative content:*

$$[\![Q_p]\!] = [\![decl(Q_p)]\!].$$

Example 1 *Polar question, DinG, SR of a polar question*

Red$_1$ *Quelqu'un veut m'échanger de l'argile ?*
Red$_1$ *"Does someone want to trade clay with me?"*

$$\begin{aligned}
&[\![\mathbf{Red}_1]\!] \\
&= [\![\textit{Does someone want to trade clay with me?}]\!] \\
&= [\![\textit{Someone wants to trade clay with me}]\!] \\
&= [\![\textit{Someone wants to trade clay with } \textbf{\textit{Red}}]\!] \\
&= \exists e.\exists x.\exists y.\ trade(e) \wedge Ag_1(e,x) \wedge Th(e,y) \\
&\quad \wedge clay(y) \wedge Ag_2(e, \mathbf{Red})
\end{aligned}$$

In this example, we identify an event `trade` that is built upon two agents and a theme. It is represented with a predicate[4].

The computation of the semantic representation is based on the syntax-semantic interface. Following (Champollion, 2011), we assume that the syntax-semantic interface we use is underpinned by the generative theory (Chomsky, 1999, 1995). Thus the parsing of a *wh*-question is derived from a first phase, the deep syntax, where the complements/arguments are in canonical positions.

Example 2 *Wh-question, DinG*
In this example, ***White*** *is talking specifically to* ***Red****.*

White$_1$ *Tu as combien de moutons ?*
White$_1$ *"How many sheep do you have?"*

For **White**$_1$, the considered intermediate state is *"You have how many sheep"*. This corresponds exactly to the structure needed to compute the semantic representation of the question. Then, an abstract variable is introduced, corresponding to the thematic role that is being interrogated. In **White**$_1$, this role is `Amount`. We denote $DeepS(S)$ the deep syntactic state of a sentence S.

Definition 2 (SR of a *wh*-question) *Consider Q_w a wh-question, where w is the thematic role corresponding to the wh-word appearing in the question. Let x be a free first order logic variable. Then the semantic representation of Q_w is the semantic representation of Q_w's deep syntactic view, where the variable corresponding to the thematic role w is x, and it is bound by λ. $[\![DeepS(Q_w)]\!]$ starts with λx and x appears in $DeepS(Q_w)$.*

$$[\![Q_w]\!] = [\![DeepS(Q_w)]\!].$$

See (Boritchev, 2017) for a correspondence between wh-words and thematic roles.

Following these definitions, example 2 gives us the following representation (see section 4 for details on the computation):

$$\begin{aligned}
&[\![\mathbf{White}_1]\!] \\
&= [\![\text{How many sheep do you have?}]\!] \\
&= [\![\text{Red has how many sheep}]\!] \\
&= \lambda x.\exists e.\exists z.have(e) \wedge Ag(e, \mathbf{Red}) \wedge player(\mathbf{Red}) \\
&\quad \wedge Th(e,z) \wedge sheep(z) \wedge Amount(z,x)
\end{aligned}$$

Remark 1. The fact that the representation of a question still contains a λ-abstraction handles the querying part of the question. For discourse-oriented representations computations, this would be the sign of a malfunction of the syntax-semantics interface.

Remark 2. We consider that the question is about the variables in the order of their abstraction. The representation of $Q_{Ag,Th,Time,Location}$ = *"Who killed whom, when and where?"* starts with an abstraction on the variable to which the predicate representing the thematic role `Agent` is applied.

(Ginzburg, 2012) distinguishes 5 types of notions for the answers: (1) *simple answerhood*, (2) *aboutness*, (3) *strong exhaustive answer*, (4) *partially resolving answer*, and (5) *question/question relations* (answer to a question by a question). We propose to introduce an intermediate level of description between questions and answers. To be able to formally define what constitutes an

[4]See (Babonnaud, 2019) for a discussion about choices of predicate ontologies.

answer to a question Q, we introduce the *relevant response set* $\mathscr{R}_Q$. It is the set of entities that fulfill the constraints introduced in the semantic representation of the question Q.

Definition 3 (Co-predicated variables) *Consider a first order logic formula F, a predicate p and two variables x and y. x and y are* co-predicated *in P if $p(x, y)$ or $p(y, x)$ occur in F.*

Definition 4 (A question's relevant response set) *Consider Q a question. The* relevant response set *w.r.t. Q, denoted $\mathcal{R}_Q$, is syntactically built from $[\![Q]\!]$ by restricting it to the quantifiers and predicates of Q applied to x, corresponding to the thematic role that is being interrogated, and its co-predicated variables in Q.*

Example 3 *Computation of $\mathscr{R}_{\mathbf{Red_1}}$ from $[\![\mathbf{Red_1}]\!]$ (example 1)*

$$\begin{aligned}
&\mathscr{R}_{\mathbf{Red}_1}\\
&= \textit{set of all entities satisfying } [\![\mathbf{Red}_1]\!]\textit{'s constraints}\\
&= \textit{set of all players that can take part in the}\\
&\quad \textit{trade event}\\
&=\{x|\exists e.\ trade(e) \wedge Ag_1(e, x) \wedge Ag_2(e, \mathbf{Red})\}\\
&=\lambda x.\exists e.\ trade(e) \wedge Ag_1(e, x) \wedge Ag_2(e, \mathbf{Red})
\end{aligned}$$

Remark 3. For the sake of simplicity, we do not take the modality conveyed by using the verb *to want* into account in this computation. There are several ways to process modality in discourse, see for example (Bybee and Fleischman, 1995; Capone, 1997; Van Ditmarsch et al., 2007).

The goal here is not to define an algorithm to build the representation, but to be able to infer $\mathscr{R}_Q$ from the representation of the question Q itself. $\mathscr{R}_{\mathbf{Red_1}}$ is built from the representation $[\![\mathbf{Red_1}]\!]$ using all the predicates concerning the queried variable (x) and all the predicates concerning the non-event variables in those predicates. The question-answer relationship is built around $\mathscr{R}$. Intuitively, this approach is close to a montagovian type-raising on the question-answer relationship. Now that we have defined how to build $\mathscr{R}_Q$ from Q, we can formally define the notion of *answer*. If we interpret $\mathscr{R}_Q$ as the representation of the answer's informational content, we also need to account for the answer's possible polarity, that varies.

Definition 5 (Answer) *An answer A to a question Q is a couple composed of a subset of $\mathscr{R}_Q$ and a polarity:*

$$\mathcal{A} = (\mathscr{R}', p), \mathscr{R}' \subseteq \mathscr{R}_Q, p \in \{+, -\}$$

Where $+$ can be represented by $\lambda P.P$ and $-$ by $\lambda P.\neg P$, see (Groenendijk et al., 1997).

Remark 3. The case of interro-negative questions uses several levels of negation, thus an adaptation of our model is necessary, see section 5.1 for further details.

Example 4 *A possible answer to Red_1*

Consider the answer "Blue does!", uttered by ***Yellow****. Then:*

$$A = (\{\mathbf{Blue}\}, +).$$

Our observations of the effect clarification requests have on preceding utterances (see section 6) bring us to the definition of two functions. These functions account for the effects of CRs on $\mathscr{R}_Q$.

Definition 6 (Clarification function) *The clarification function* CF *takes as arguments a λ-term L of neo-Davidsonian event semantics and two predicates p_1 and p_2, such that p_1 appears in L and the arity of p_1 and p_2 is the same.* CF(L, p_1, p_2) *is L where p_1 has been substituted by p_2.*

CF accounts for modifications of predicates inside an existing $\mathscr{R}_Q$. The next function, CF+, adds information inside an existing $\mathscr{R}_Q$.

Definition 7 (Additive clarification function) *The additive clarification function* CF+ *takes as arguments a λ-term L of neo-Davidsonian event semantics and one predicate, and the result of its application is the addition of the predicate in the L, under the scope of the λ.*

We use $\mathscr{R}_Q$, A, CF and CF+ to model the articulation of utterances in dialogue. First, we present the way our model works with an extensive example.

4 The Model: an Example

The discussion in section 3 gives an overview of the semantic representation we want to compute. We focus on the compositional treatments that produce the representation. In the following, we compute the semantic representation of example 2.

We focus on a slightly modified version of the example with an explicit player, **Red**, as instead of the deictic 'you' we prefer to use a constant

in the semantic representation. Another solution would be to use a free variable, but we prefer not to use them as we want to produce well-formed logical formulae. We claim that the representation of the question is equivalent to the one of the deep syntax. We consider both the syntax and thematic role: here, **Red** is the `Agent`, *"sheep"* the `Theme` and *"how_many"* the `Amount`.

$$
\begin{aligned}
&[\![\text{How many sheep do you have?}]\!] \\
=\ &[\![\text{you have how many sheep}]\!] \\
=\ &[\![\textbf{Red}\text{ has how many sheep}]\!] \\
=\ &[\![\textbf{Red}\text{[ag] has how many[amount] sheep[th]}]\!]
\end{aligned}
$$

Remark 3. 'How many' is treated as a multiword expression, following a simplified version of (Asher and Lascarides, 1998). The *wh*-word is considered as the syntactic determiner, and semantically expresses the quantity/amount.

The syntactic relations in the sentence can be represented with the tree of the figure 1. The nodes of the tree account for the semantic derivation and present the types of the intermediate computation steps. Leaves are either words of the sentence or thematic roles.

Here, we follow a simplification of the syntax-semantic interface for generative theory, (Chomsky, 1999), in the same perspective as (Champollion, 2015). For a developed presentation of the syntax-semantic interface for generative theory in montagovian view, see (Amblard, 2007).

Following the application order given by the syntactic tree, we compute the semantic representation of example 2.

$$
\begin{aligned}
=\ &[\![\textbf{Red}\text{[ag] has how many[am] sheep}[th]]\!] \\
=\ &[\![(\textbf{Red}\ [agent])\ (\text{has}\ (((\text{how_many}\ [amount]) \\
&\text{sheep}\ )[theme]))]\!]
\end{aligned}
$$

Once we have the structure of the syntax-semantic interface, we explicitly have the functional application for our λ-terms. First, we need to discuss the semantic types.

We start from Montague's semantics based on simple type theory (Church, 1940), from which we inherit the elementary types e for *entities* and t for *truth values*, and the application $\langle _, _ \rangle$ (Cresswell, 1976).

Following (Champollion, 2011), we use v as the elementary type for *events*. Finally, we introduce a list of elementary types that are specific to semantic roles; here, we will consider the type n for *numerals*, corresponding to the semantic role $[amount]$.

In a montagovian tradition, a sentence is of type t: it corresponds to its truth value, as a sentence is either true or false. In neo-Davidsonian event semantics, a sentence is a quantification on an event, it is of type $\langle vt, t \rangle$, that we denote $\mathbb{S}_e$.

Consider the computation of the node marked ①. **Red** is a noun phrase (NP), that is classically of type $\langle et, t \rangle$, that we denote $\mathbb{NP}$. $[agent]$ is a thematic role, and following a neo-Davidsonian event semantics interpretation, it changes a noun phrase into a neo-Davidsonian event semantics noun phrase (NP_e).

Note that in that work the semantic of a $\mathbb{NP}$ does not include the event. We claim that the NP is semantically defined for itself and it is its inclusion in the sentence which need a variable which make the link in the formula. Here, $[\![\textbf{Red}]\!] = \textbf{Red}$ but the semantic representation of the $\mathbb{NP}$ contains $agent(e, \textbf{Red})$ where e is the agent variable of the main verb. Note that in that case, we need to switch the functor/argument relation between the $\mathbb{VP}$ and the $\mathbb{NP}$.

The $\mathbb{NP}$ is of type $\langle \langle et, t \rangle, \langle \langle vt, t \rangle, \langle vt, t \rangle \rangle \rangle$, that we denote $\langle \mathbb{NP}, \mathbb{NP}_e \rangle$. Following (Champollion, 2015). The corresponding λ-terms are:

$$
\begin{aligned}
[\![\textbf{Red}]\!] &= \lambda P.P(\textbf{Red}) \\
[\![[agent]]\!] &= \lambda Q.\lambda V.\lambda f.Q(\lambda x.V(\lambda e.[f(e) \\
&\qquad \wedge Ag(e, x)]))
\end{aligned}
$$

It is important to notice that the latter term composes a previous term of the derivation with a thematic predicate. The challenge here is the unification of the event variable.

Then, the type of the verb is adapted to the fact that we are in a framework with events. Therefore, *has* is of type $\langle \mathbb{NP}_e, \langle \mathbb{NP}_e, \mathbb{S}_e \rangle \rangle$. Using these notations, we give the types of the constituents in table 1.

In a purely computational perspective, asking a question amounts to invert the functor/argument relationship to rise the questioned element to the surface. This idea gives us the type of $[\![how_many]\!]$.

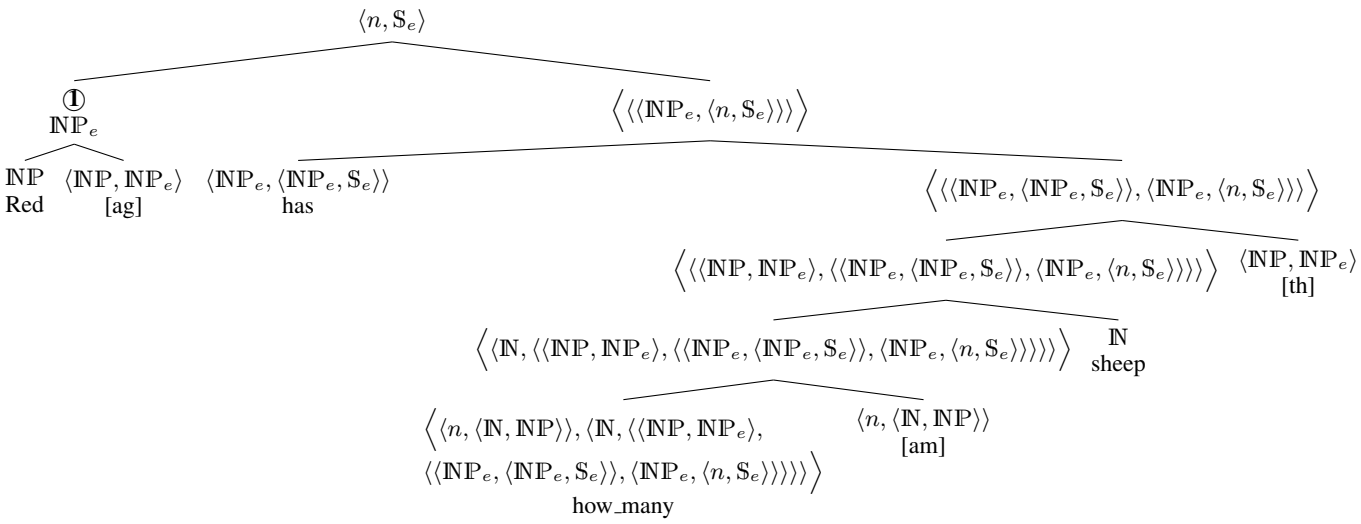


Figure 1: Representation of the syntactic relations

$[\![Red]\!]$	$\mathbb{NP}$
$[\![[agent]]\!]$	$\langle \mathbb{NP}, \mathbb{NP}_e \rangle$
$[\![has]\!]$	$\langle \mathbb{NP}_e, \langle \mathbb{NP}_e, \mathbb{S}_e \rangle \rangle$
$[\![how_many]\!]$	$\Big\langle \langle n, \langle \mathbb{N}, \mathbb{NP} \rangle \rangle, \langle \mathbb{N}, \langle \langle \mathbb{NP}, \mathbb{NP}_e \rangle, \langle \langle \mathbb{NP}_e, \langle \mathbb{NP}_e, \mathbb{S}_e \rangle \rangle, \langle \mathbb{NP}_e, \langle n, \mathbb{S}_e \rangle \rangle \rangle \rangle \rangle \Big\rangle$
$[\![[amount]]\!]$	$\langle n, \langle \mathbb{N}, \mathbb{NP} \rangle \rangle$
$[\![sheep]\!]$	$\mathbb{N}$
$[\![[theme]]\!]$	$\langle \mathbb{NP}, \mathbb{NP}_e \rangle$

Table 1: Semantic types, simplified notation

Then, we have:

$$[\![[amount]]\!] = \lambda n.\lambda R.\lambda P.\exists x.(R(x) \wedge P(x) \wedge Amount(x, n))$$

We note that here the type of $[\![how_many]\!]$ is quite complex. The double type raising leads to a unique term around which the whole semantic representation of the sentence is build. An important remark is the fact that the term must be first combined with a term of type n, pushed at the end of the type with $\langle n, \mathbb{S}_e \rangle$, which is the type of a sentence where an n is missing.

This exactly corresponds to our idea to combine the different pieces of semantic representations into a single formula.

We propose a treatment of the example 2. The same treatment can be applied if the question is on the $\mathbb{NP}$ subject, but in that case, we have to rewrite the λ-term.

Another solution is to shift the focus from *"how_many"* to the thematic role itself, by deconstructing *"how_many"* and writing questioning versions of the thematic roles, that would be different from the declarative ones. We leave this discussion open for future works.

5 The Model: Case Studies

In this section, we illustrate how $\mathscr{R}_Q$ is used as an intermediate level of representation between questions and answers and how answers can be related to a question.(Ginzburg, 2012) introduces five answer types. We distinguish *simple answers* from the four others, that are more complex and are addressed in section 5.2.

5.1 Simple answers

A *simple answer* is a single-item utterance in a dialogue concerning either an instantiation of a question or the negation of such an instantiation.

Example 5 *Simple answer, DinG*[5]

Blue$_1$ *C'est quoi qui rapporte le plus de thune ?*
Blue$_1$ *"What brings in the most cash?"*

Yellow$_2$ *Le 6*
Yellow$_2$ *"The 6"*

The answer is unique (6). It is an instantiation of the problem *"Among the possible numbers on the dice, what is the one that brings in the most cash?"*.

Definition 8 (Simple answer)
A simple answer is an answer $\mathcal{A} = (\mathscr{R}', p)$

[5] In Catan, profit depends on the number on which the die falls.

where $\mathscr{R}'$ is a singleton.

$$\mathcal{A}_{simple} = (\{r\}, p).$$

Example 6 *Semantic representation of* ***Yellow***$_2$

$$[\![Yellow_2]\!] = (\{6\}, +).$$

Polar questions constitute a special case of interest, as for a polar question Q_p, $\mathscr{R}_{Q_p} = \emptyset$. A simple answer to a polar question is a confirmation or a negation of its declarative content. Thus, it corresponds (only) to the polarity: $(\emptyset, p), p \in \{+, -\}$.

Therefore, there is a difference between a simple answer to a (*wh*- or disjunctive) question, where $\mathscr{R}'$ is a singleton, and an answer to a polar question, where $\mathscr{R}'$ is empty. This difference allows us to argument for the pair model, as the first part of the pair models the informational content, needed to complete the dialogue, while the polarity is the way this element is included in the dialogue.

Example 5 shows that it is possible to find straightforward examples in real-life data that fit the simple answers definition. Yet, several problems frequently occur. First, while classifying answers by types is a difficult problem, finding what segment of data is an answer is a whole other problem of its own.

Example 7 *Simple answer, DinG*
Yellow$_1$ *Tu as fait combien ?*
Yellow$_1$ *"How many did you make?"*

Red$_2$ *oh oui 11 d'accord*
Red$_2$ *"Oh yes 11 OK"*

In example 7, the simple answer is *"11"*, but it is surrounded by phatic elements (*"Oh yes"* and *"OK"*). If **Yellow**$_1$ was a polar question, these phatic elements could have been simple answers. Therefore, any operationalization of our approach first involves identifying the span of the answers.

Example 8 *Simple answer, DinG*
Yellow$_1$ *Tu veux pas faire d'échange ?*
Yellow$_1$ *"Don't you want to make a trade?"*

Red$_2$ *Non*
Red$_2$ *"No"*

Another usual difficulty lies in the treatment of interro-negatives and their answers. **Yellow**$_1$ is a negative polar question, its declarative content is $decl(\mathbf{Yellow}_1)$ = *"**Red** doesn't want to make a trade"*. If **Red** answers *"Yes"*, this answer validates the declarative content. However, the actual answer (**Red**$_2$) also validates the declarative content, while being supposedly a negative one. Different languages treat these constructions in different ways, in French it is possible to force the negation of a negative declarative content (therefore, to refuse the informative content of the question) by using the word *"si"* instead of *"oui"* or *"non"* (*"yes"* or *"no"*).

5.2 Complex answers and answer properties

Human communication heavily relies on logical inferences, ambiguities and, in general, pragmatics. Therefore, there are types of answers that are not as direct as the simple ones, as they don't give the requested information right away. The following develops on the four other descriptions from (Ginzburg, 2012).

5.2.1 Aboutness

When answers are not simple ones, different cases emerge. For polar questions, sometimes the answer is not a straight confirmation or a straight negation of the informational content. It is the case for conditional answers (*"if* `condition` *then yes"*) and answers under modalities (*"maybe"*, *"probably"*,*"possibly not"*).

Example 9 *DinG Corpus, aboutness*
Red$_1$ *Tu fais quelque chose ?*
Red$_1$ *"Are you doing something?"*

Yellow$_2$ *Ah ben j'ai construit 2 chemins si tu veux j'ai pas l'habitude de faire des trucs de fou comme ça*
Yellow$_2$ *"Well I built 2 roads you know I'm not used to doing crazy stuff like that"*

In example 9, **Red** asks a polar question. **Yellow** answers with a long sentence that amounts to a negative answer, but a human observer needs to make an inferential step to be able to deduce this: as **Yellow** is not used to *'doing crazy stuff like that'*, she will not be doing a lot of that, so she will not be doing something at the present moment. For discussions of these types of enthymematic reasonments, see (Breitholtz, 2014).

Wh-questions can be answered with disjunctive answers, presenting several possible short answers (in the context of example 5, *"The 3 or the 6, I don't remember"*) or with quantified answers (including

generalized quantifiers such as *"at most"*, *"a few"*, *etc.*).

Example 10 *Aboutness, DinG*

Blue$_1$	*Qui se fait souvent de la pierre ?*
Blue$_1$	*"Who often makes stone?"*
White$_2$	*Euh, je m'en suis faite une seule depuis tout à l'heure*
White$_2$	*"Uh, I've only made one since earlier"*

In this example, **White**'s answer allows us to rule her out of the set of players among whom we are looking for the ones that often make stone. Thus, this answer gives additional information but doesn't completely solve the issue raised by **Blue**. This explains why $\mathscr{R}_Q$ always contains a set of possible answers to the question Q.

Computationally, this corresponds to the creation of $\mathscr{R}_2 \subseteq \mathscr{R}_{\mathbf{Blue}_1}$ by adding the information brought by the predicates introduced by the answer.

Example 11 *Computation of $\mathscr{R}_{\mathbf{Blue}_1}$, see example 10.*
In this example, we simplify the representation and we do not consider the temporal relation, thus we remove the representation of "since earlier".

$$\begin{aligned}
[\![\mathbf{Blue}_1]\!] = \ & \lambda x.\exists e.\exists y.make(e) \wedge Ag(e,x) \\
& \wedge stone(y) \wedge Th(e,y) \\
& \wedge often(e) \\
\mathscr{R}_{\mathbf{Blue}_1} = \ & \lambda x.\exists e.make(e) \wedge Ag(e,x) \\
& \wedge often(e)
\end{aligned}$$

$\mathscr{R}_2$ is then built based on $[\![\mathbf{White}_2]\!]$:

$$\mathscr{R}_2 = \mathscr{R}_{Blue_1} \oplus [\![\mathbf{White}_2]\!],$$

where $\oplus$ is defined as a dynamic conjunction of the predicates introduced in the answer with an opening of the scope of λ. It works as if the λ was removed while the conjunction operates and goes back in once it is done.

$$\begin{aligned}
[\![\mathbf{White}_2]\!] = \ & \exists e.\exists y.make(e) \wedge Ag(e,\mathbf{White}) \\
& \wedge stone(y) \wedge Th(e,y) \\
& \wedge Amount(y,1) \\
\mathscr{R}_2 = \ & \lambda x.\exists e.\exists e'.\exists y.\exists y'.make(e) \\
& \wedge Ag(e,x) \wedge often(e) \\
& \wedge make(e') \wedge Ag(e',\mathbf{White}) \\
& \wedge stone(y') \wedge Th(e',y') \\
& \wedge Amount(y,1)
\end{aligned}$$

The property holds because $Amount(y,1)$ is about $\neg often(e')$. Thus $\mathbf{White}_2$ negates the question $\mathbf{Blue}_1$.

5.2.2 Strongly exhaustive answer

An answer A is *strongly exhaustive* for a question Q if and only if A is true and entails all the A_i that are simple answers to Q, (Ginzburg, 2012).

Intuitively, an answer is considered to be strongly exhaustive if independently from what comes next in the course of the dialogue, this answer cannot become more precise. It isn't necessary a simple answer, but the pragmatic environment of the dialogue allows us to approximate it as such. Thus, for a question Q, a strongly exhaustive answer $A = (\mathscr{R}', p)$ is such that $\mathscr{R}' = \mathscr{R}_Q$. $\mathscr{R}'$ is such that for all $A_i = (\mathscr{R}_i, p_i)$ a simple answer, $\mathscr{R}' \oplus \mathscr{R}_i = \mathscr{R}'$. Formally, this amounts to producing a fixed point on the objects $\mathscr{R}$.

We have not found examples of strongly exhaustive answers in DinG yet. Example 12 presents a constructed example of a strongly exhaustive answer in DinG's context.

Example 12 *Strongly exhaustive answer*

Red$_1$	*Tu veux quoi contre du blé ?*
Red$_1$	*"What do you want to trade for wheat?"*
Yellow$_2$	*De la pierre ou du bois*
Yellow$_2$	*"Stone or wood"*

Yellow$_2$ is true and entails both *"stone"* and *"wood"*, that are simple answers to **Red**$_1$.

5.2.3 Potentially resolving answer

Following (Ginzburg, 2012), a *potentially resolving answer* is an answer that either brings a simple answer, either shows that there is no answer.

Example 13 *Strongly exhaustive answer (self), DinG*

Yellow$_1$ *Alors qu'est-ce que je peux faire avec ça ?*
Yellow$_1$ *"So what can I do with this?"*

Yellow$_2$ *Euh ben rien en gros rien du tout*
Yellow$_2$ *"Uh, well, nothing, basically nothing at all"*

This characterisation of an answer is interesting as it seems to come directly from pragmatics. It finds an impact directly in our model. This definition supposes that the question is solved but doesn't specify the origin of the information that solves the question. It may come from the first component of the answering couple (the information itself), so from an effective answer. It may also come from the second component of the couple, the polarity. It is actually also possible that there is no answer. We make the choice here to suppose that there is an answer and that its informational content comes from $\mathscr{R}'$.

In example 13, **Yellow** both asks the question and gives the answer. Here, the resolution of the question comes directly from the information contained in the answer. The question/answer mechanism is covertly used both to support **Yellow**'s thinking process and to convey information to the other players.

5.2.4 Question/question relations

Knowing that it's possible to answer a question with a question forces us to consider the interactive aspect of dialogues, drawing us closer to the discursive modeling in the tradition of (Asher et al., 2003). This reference gives us the definition of a `Question elaboration` that stresses out the relation between two questions linked in a way such that any answer to the second one gives a strategy that leads to the solving of the query of the first one.

Example 14 *DinG Corpus, question/question*
White$_1$ *Tu as combien de moutons ?*
White$_1$ *"How many sheep do you have?"*

Blue$_2$ *Tu voudrais combien de moutons ?*
Blue$_2$ *"How many sheep would you like?"*

Intuitively, this means that it is possible to build $\mathscr{R}_{\mathbf{White}_1}$ but it is not possible to specify an answer element using $[\![\mathbf{Blue}_2]\!]$. It is not even possible to specify a subset of $\mathscr{R}_{\mathbf{White}_1}$ as a case of aboutness. The strategy here is to define a new $\mathscr{R}$ from $\mathscr{R}_{\mathbf{White}_1}$, using $[\![\mathbf{Blue}_2]\!]$, to allow the interaction to progress. $\mathscr{R}$ is such that $b \in \mathscr{R}$ if $\forall a.(b \vdash a) \wedge (a \in \mathscr{R}_{\mathbf{White}_1})$.

5.2.5 Other answers

The statistical and linguistic studies from (Blandón et al., 2019; Amblard et al., 2019) show the existence of two additional categories of answers: `Uncertain` and `Unknown`.

Example 15 *DinG Corpus, Uncertain*
Red$_1$ *J'ai le droit de le dire ça ? Ce que tu m'as piqué ?*
Red$_1$ *"Can I say that? What you stole from me?"*

Yellow$_2$ *J'en sais rien je pense c'est pas dramatique si les gens savent un peu euh ce que tu as fait*
Yellow$_2$ *"I don't know, I think it's not a big deal if people know a little bit about, uh, what you did"*

These concepts can be defined in our model. An `Uncertain` answer corresponds to the characterisation of an answer whose link with the question Q cannot be stated explicitly. Here, it is equivalent to saying that we don't know whether this answer is part of $\mathscr{R}_Q$, or, rather, that we don't know how to check whether it's the case. An `Unknown` answer means that we don't have access to the answer, either because of our lack of knowledge, thus making $\mathscr{R}_Q$ contain all the possible entities in the world, or because our knowledge base is not specific enough, making the cardinal of $\mathscr{R}_Q$ too big.

6 Clarification requests

(Purver, 2004) lists eight clarification request (CR) types, see table 2. Our study shows that the effect on preceding utterances of these different clarification requests in dialogue can be modeled using the functions `CF` and `CF+`, see table 3 for a summary of our approach. We group clarification requests by similarity of action on the preceding utterances.

Repetition requests (Wot, exp) `Wot` is a category of CRs that is used to request a repetition of an utterance or part of it.

Example 16 *DinG Corpus, Wot*
Blue$_1$ *Je suppose que personne ne veut du blé contre euh un bois euh un argile?*
Blue$_1$ *"I guess no one want a wheat for a wood uh a clay?"*

Category	Description	Example
Wot	words used to request repetition	Eh? / What? / Pardon?
Explicit (exp)	context-independent CRs	What did you say? / Did you say 'Bo'? / What do you mean 'leave'?
Literal Reprise (lit)	verbatim repetitions of the troubled utterance (while changing the interlocutor-oriented indexicals)	Did BO leave? / Did Bo LEAVE?
Wh-substituted Reprise (sub)	lit where a constituent is replaced by a *wh*-phrase	Did WHO leave? / Did Bo WHAT?
Reprise sluice (slu)	bare *wh*-phrase used as CR	Who? / What? / Where?
Reprise Fragments (RF)	bare phrase used as CR	Bo? / Leave?
Gaps	omitted targeted constituent	Did Bo ...?
Fillers	guess for the utterance intended in the unfinished antecedent sentence	A: Did Bo... B: Win?

Table 2: Categories of clarification requests, (Purver et al., 2003)

Yellow$_2$ *Quoi ?*
Yellow$_2$ *"What?"*

`exp` has the same effect as `Wot` but is produced by using an explicit question, that can be understood without knowing the exact context.

Example 17 *DinG Corpus, Explicit request*[6]
Blue$_1$ *Est-ce que quelqu'un voudrait m'échanger un blé contre un bois?*
Blue$_1$ *"Does someone want to trade a wheat for a wood with me?"*

Yellow$_2$ *C'est dans quel sens tu dis pas dans quel sens?*
Yellow$_2$ *"Which way you are not saying which way?"*

Both these categories of CR request repetition, rephrasing. `Wot` creates a new $\mathscr{R}$ that we know nothing about. It can be completely unrelated to the old one. `exp` creates a $\mathscr{R}_{new}$ that we can only characterize by $\mathscr{R}_{old} \cap \mathscr{R}_{new} \neq \emptyset$.

Example 18 *Explicit clarification request, DinG*
Red$_1$ *Quelqu'un veut m'échanger de l'argile ?*
Red$_1$ *"Does someone want to trade clay with me?"*

Blue$_2$ *Tu as de l'argile ou tu veux de l'argile ?*
Blue$_2$ *"Do you have clay or do you want clay?"*

Red$_3$ *Je veux de l'argile*
Red$_3$ *"I want clay"*

In example 18, **Blue** makes an explicit clarification request using a fully formed disjunctive question. Then, **Red**$_3$ is a simple answer.

Utterance reprises (lit, sub) `lit` is a repetition of the utterance that needs to be clarified. The interlocutor-oriented indexicals change and the prosody is used to stress out the interrogated part. `sub` acts as `lit`, it is a repetition of the utterance that needs to be clarified, but the interrogated part is replaced by a *wh*-phrase. `lit` and `sub` both act as `CF` on $\mathscr{R}$, on predicates and their arguments for `lit` and on thematic roles for `sub`.

Phrase reprises (slu, RF) `slu` is a bare *wh*-phrase, it is an anaphoric CR. `RF` is a bare phrase that is not a *wh*-phrase. `slu` creates a *wh*-question while `RF` creates a polar question, when combined with the previous utterance.

Example 19 *DinG Corpus, CR*
Blue$_1$ *2 pierres contre 1 argile ?*
Blue$_1$ *"2 stones for 1 clay?"*

Red$_2$ *2 pierres ?*
Red$_2$ *"2 stones?"*

These CRs shift the focus of the conversation while not changing the $\mathscr{R}$. To model this effect, we use $\texttt{CF}_{id}$: `CF` that takes as arguments $\mathscr{R}$, a predicate of $\mathscr{R}$, and the same predicate of $\mathscr{R}$. $\texttt{CF}_{id}$ leaves $\mathscr{R}$ unchanged.

[6] Is it that **Blue** gives a wheat to get a wood or the other way around?

Category	Effects of CRs on $\mathscr{R}$
Wot	$\mathscr{R}_{new}$
Explicit (exp)	$\mathscr{R}_{new}$
Literal Reprise (lit)	CF
Wh-substituted Reprise (sub)	CF
Reprise sluice (slu)	CF_{id}
Reprise Fragments (RF)	CF_{id}
Gaps	CF+
Fillers	CF+

Table 3: Effects of clarification requests on the relevant response set

Example 20 *Reprise Fragment, DinG*

Blue$_1$ *Quel goût, les chips ?*
Blue$_1$ *"The chips - what flavor?"*

Red$_2$ *Mais elles sont juste dehors*
Red$_2$ *"[but] they're right outside"*

Blue$_3$ *Non mais quel goût ?*
Blue$_3$ *"No, [but] what flavor?"*

In this example, **Blue**$_3$ is a question that cannot be understood without the context given by the previous dialogue turns. In that case, we use $\mathscr{R}_{Blue_1}$ to built the representation of **Blue**$_3$.

Gaps & fillers `Gap` is a CR that creates a question by omitting the interrogated constituent in an utterance. `Gap` can create a question from a declarative sentence or a question likewise. `Fillers` act as counterparts for `Gaps` as `Fillers` are attempts at guessing a way to complete incomplete utterances or `Gaps`. These CRs request additional information, so their effect on $\mathscr{R}$ is modeled by `CF+`.

7 Conclusions and further work

We introduce a model for the question-answer relationship and for the articulation of different types of utterances (questions, answers, clarification requests) in dialogue. Based on the descriptions given in (Ginzburg, 2012), we define the relevant response set $\mathscr{R}_Q$ w.r.t a question Q, which helps us to characterise the utterances that can follow Q. We also design set operators `CF` and `CF+`, which implement modifications in the content of sets. We use them to reflect the effect of clarification requests on preceding utterances. The modelisation of different types of answers and of clarification requests directly follows from the definitions of $\mathscr{R}_Q$ and `CF`, `CF+`, without having to add new properties or define new objects.

Though this point is not explicitly addressed here, the methodological and technical tools that are used to produce the logical semantic representations have been developed for years (see for example (Pogodalla, 2004)). The model presented here is quite simple as it is based on a small amount of elementary types and few objects. Yet, it allows us to gather several concepts that are difficult to conciliate such as compositional approaches to language and corpus-based observations modeling.

The perspectives for this work are now on one hand to apply the model on real-life data and on the other hand to make the computation of representations operational. However, though this last part is the most important one, it still needs for our model to be included in a more global system that takes into account different phases of dialogue and the way they articulate (Boritchev and Amblard, 2019b).

A substantial issue for our model is to include a proper treatment of Non-Sentential Utterances, which are important in the articulation of spontaneous speech. In the same perspective, we need to give a formal account of phatic expressions. Both phenomena seem to follow the same kind of process and need a specific treatment. Another relevant perspective is to propose an implementation that would be able to explicitly compute the answers to questions. To this end, we will need to include a dialogue parser to our pipeline.

Our next step will be to extend and polish this model. Some of the types we presented in section 4 may seem quite ad-hoc, but it is actually possible to to produce them in a systematic way. We also want to enrich our model by anchoring it in Inquisitive Logic (Ciardelli et al., 2017). Then, we want to address dynamicity-related issues in dialogue by taking into account the contexts and the common ground (Stalnaker, 2002): what does a player know and/or believe, what knowledge and/or belief is common to all the players, and how do those evolve as the multilogue unfolds. To these ends, we want to adapt Continuation Style Dynamic Semantics (De Groote, 2006) for dialogue and integrate the result in our model.

Acknowledgements

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE. We thank Philippe de Groote for his insight and his comments, Laurine Jeannot for her ideas and the discussions. We also thank our reviewers for all the useful comments and references.

References

Maxime Amblard. 2007. *Sémantic representations and generative grammar: the Minimalist Categorial Grammars*. Theses, Université Sciences et Technologies - Bordeaux I.

Maxime Amblard, Maria Boritchev, Marta Carletti, Lea Dieudonat, and Yiting Tsai. 2019. A Taxonomy of Real-Life Questions and Answers in Dialogue. In *SemDial 2019 - LondonLogue - 23rd Workshop on the semantics and pragmatics of dialogue*, London, United Kingdom.

Maxime Amblard, Karën Fort, Caroline Demily, Nicolas Franck, and Michel Musiol. 2015. Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Languages*, 55:91–115.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016a. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727, Portoroz, Slovenia.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016b. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727, Portoroz, Slovenia.

Nicholas Asher and Alex Lascarides. 1998. Questions in dialogue. *Linguistics and philosophy*, 21(3):237–309.

John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.

William Babonnaud. 2019. A Topos-Based Approach to Building Language Ontologies. In *Formal Grammar. 24th International Conference, FG 2019*, volume 11668 of *Lecture notes in Computer Science*, pages 18–34, Riga, Latvia.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

María Andrea Cruz Blandón, Gosse Minnema, Aria Nourbakhsh, Maria Bortichev, and Maxime Amblard. 2019. Toward dialogue modeling: A semantic annotation scheme for questions and answers. In *The 13th Linguistic Annotation Workshop (The LAW XIII)*.

Maria Boritchev. 2017. Approaching dialogue modeling in a dynamic framework. Master's thesis, Université de Lorraine.

Maria Boritchev and Maxime Amblard. 2018. Coffee or tea? Yes. In *The 22nd workshop on the Semantics and Pragmatics of Dialogue*, Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue, Aix-en-Provence, France. Laurent Prévot, Magalie Ochs and Benoît Favre.

Maria Boritchev and Maxime Amblard. 2019a. A compositional view of questions. In *WiNLP - Widening NLP - ACL Workshop*, Florence, Italy.

Maria Boritchev and Maxime Amblard. 2019b. Picturing Questions and Answers - a formal approach to SLAM. In Maxime Amblard, Michel Musiol, and Manuel Rebuschi, editors, *(In)coherence of discourse - Formal and Conceptual issues of Language*, (In)coherence of discourse Formal and Conceptual issues of Language. Springer. Language, Cognition and Mind.

Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, Göteborgs universitet. Humanistiska fakulteten University of Gothenburg. Faculty of Arts.

Harry Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Lou Burnard. 2000. Reference guide for the british national corpus (world edition).

Joan L Bybee and Suzanne Fleischman. 1995. *Modality in grammar and discourse*, volume 32. John Benjamins Publishing.

Alessandro Capone. 1997. *Modality and discourse*. Ph.D. thesis, University of Oxford.

Lucas Champollion. 2011. Quantification and negation in event semantics. *Discourse, Context, and Models*, 6:1–23.

Lucas Champollion. 2015. The interaction of compositional semantics and event semantics. *Linguistics and Philosophy*, 38(1):31–66.

Noam Chomsky. 1995. *The minimalist program*, volume 28. Cambridge Univ Press.

Noam Chomsky. 1999. *Derivation by phase*. ms, MIT.

Alonzo Church. 1940. A formulation of the simple theory of types. *The journal of symbolic logic*, 5(2):56–68.

Ivano Ciardelli, Floris Roelofsen, and Nadine Theiler. 2017. Composing alternatives. *Linguistics and Philosophy*, 40(1):1–36.

Robin Cooper and Jonathan Ginzburg. 2015. 12 type theory with records for natural language semantics. *The handbook of contemporary semantic theory*, page 375.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2-3):281–332.

M. J. Cresswell. 1976. Formal philosophy, selected papers of richard montague. *Philosophia*, 6(1):193–207.

Philippe De Groote. 2006. Towards a montagovian account of dynamics. In *Proceedings of Semantics and Linguistic Theory XVI*.

Markus Egg. 1998. Wh-questions in underspecified minimal recursion semantics. *Journal of Semantics*, 15(1):37–82.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Alice F. Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of Pragmatics*, 21(6):621 – 644.

Jonathan Ginzburg. 1995a. Resolving questions, I. *Linguistics and philosophy*, 18(5):459–527.

Jonathan Ginzburg. 1995b. Resolving questions, II. *Linguistics and Philosophy*, 18(6):567–609.

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

Jonathan Ginzburg and Ivan Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.

H. P. Grice. 1975. Logic and conversation. *Syntax and Semantics*, page 3:41–58.

Jeroen Groenendijk, Martin Stokhof, Johan Van Benthem, and Alice ter Meulen. 1997. Handbook of logic and language.

Julie Hunter, Nicholas Asher, and Alex Lascarides. 2018. A Formal Semantics for Situated Conversation. *Semantics and Pragmatics*, 11:1–52.

Manfred Krifka. 2001. For a structured meaning account of questions and answers. *Audiatur vox sapientia. a festschrift for arnim von stechow*, 52(2055):287–319.

Tuan Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2019. A gated self-attention memory network for answer selection. *arXiv preprint arXiv:1909.09696*.

Claudia Maienborn. 2011. Event semantics. *An international handbook of natural language meaning*, 1.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Arvid Österlund, David Ödling, and Magnus Sahlgren. 2015. Factorization of latent variables in distributional semantic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 227–231.

Sylvain Pogodalla. 2004. Computing semantic representation: Towards ACG abstract terms as derivation trees.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Matthew Richard John Purver. 2004. *The theory and use of clarification requests in dialogue*. Ph.D. thesis, University of London.

Jinfeng Rao, Hua He, and Jimmy Lin. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1913–1916.

David Schlangen. 2003. *A coherence-based approach to the interpretation of non-sentential utterances in dialogue*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering.

David Schlangen and Alex Lascarides. 2003. The interpretation of non-sentential utterances in dialogue. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

David Traum. 1994. A computational theory of grounding in natural language conversation.

Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. 2007. *Dynamic epistemic logic*, volume 337. Springer Science & Business Media.

Anne Xuereb and Jean Caelen. 2004. Un modèle d'interprétation pragmatique en dialogue homme-machine basé sur la sdrt. In *Workshop SDRT, TALN*, volume 4.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer extraction

as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 858–867.

Xuchen Yao and Yi Zhang. 2010. Question generation with minimal recursion semantics. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 68–75. Citeseer.

Dialogue Informing Syntax/Semantics: The Case of Afterthoughts

Stergios Chatzikyriakidis
Centre for Linguistic Theory and Studies in Probability,
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
stergios.chatzikyriakidis@gu.se

Abstract

In this paper, I look at the case of *afterthoughts* and claim that these can receive a natural syntactic explanation once they are seen as clarification strategies, more specifically, as answers to implicit clarification questions. I argue that this assumption paired with a framework that is suited to deal with structures that go beyond sentential syntax/semantics, in this case Dynamic Syntax, provides a natural explanation to the syntactic and intepretational intricacies associated with afterthoughts. On a more general level, the account advocates a view which methodologically uses dialogue modelling research to inform research on more core syntax/semantics issues.

1 Introduction

Looking at the syntactic literature on Right Dislocations (RD), one comes across a number of further categorizations. The one which has been prominent in a number of papers is the distinction between Backgrounded Right Dislocation (BRD) and Afterthoughts (ATs). This has been advocated by Averintseva-Klisch (2008a,b) for German, Ott and de Vries (2014); De Vries (2007) for Dutch, Chafe (1988) for English and Chatzikyriakidis (2017) for Greek among others. BRD functions as a content re-establisher device and has been claimed to involve comma intonation. On the other hand, ATs, as Averintseva-Klisch (2008a,b) claims, are clarification devices that carry the speaker's attempt to elucidate part of the preceding utterance/sentence and also exhibit period instead of comma intonation:[1]

(1) I know her, Ruth Kempson

(2) I know her... Ruth Kempson

[1]Comma intonation is noted with a single comma, and period intonation with three dots.

The two structures are not only different in their pragmatic/semantic import but also w.r.t their syntax. For example, in gender-marking languages, like German and Greek, gender mismatches can be attested between the dislocated element and its referent, something which is not possible for BRDs. The example below illustrates a gender-mismatch case from German taken from (Averintseva-Klisch, 2008b):

(3) Ich habe ihn$_i$ vorhin gesehen... das
I have him$_{MASC}$ before seen, the
Kleine$_i$ von der Nachbarn
little-one$_{NEUTR}$ of the neighbours

Besides the morphological mismatches, there are another two main differences between ATs and BRDs that I want to concentrate on in this paper: free positioning of ATs in the utterance sentence and optional additions. This gives us the following list containing three main properties relevant for ATs but not BRDs:

1. Morphological Mismatch
2. Free position in the utterance/sentence
3. Optional additions (e.g. "I mean") are possible

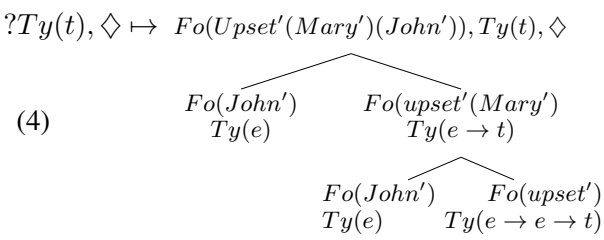
In what follows, I will argue that the corrective nature of ATs and its differences with BRDs can be explained once we make the assumption that ATs can be seen as clarification answers to an implicit question. Elaborating a bit, we claim that the speaker, by using an AT, tries to prevent a clarification question by the hearer, in effect by answering it. If this assumption is on the right track, then the minimum we need to model this idea, is a framework that is capable of going beyond the sentence level and deal with dialogue data. In this paper, I use Dynamic Syntax for this purpose.

The paper is structured as follows: in section 2, I provide a brief intro to Dynamic Syntax, concentrating on the features that are relevant for this paper. In section 3, I provide the main bulk of the analysis, putting forth an implementation of the idea which takes ATs to be answers to implicit questions. The difference between ATs and BRDs is discussed, and the main differences between are captured. There is also a discussion of specificational ATs and recursive ATs. In section 4, I conclude.

2 A brief Intro to Dynamic Syntax

The Dynamic Syntax (DS) framework (Kempson et al., 2001; Cann et al., 2005b) is a processing oriented framework, that has been successfully used both as a syntactic framework (Chatzikyriakidis, 2010; Bouzouita, 2008; Kempson and Kiaer, 2010; Gregoromichelaki, 2013; Seraku, 2013; Marten and Gibson, 2015), as well as a general framework for dialogue modeling (Gregoromichelaki et al., to appear; Eshghi et al., 2012, 2015; Kempson et al., 2016). One of the main ideas behind DS is that natural language syntax can be seen as the progressive accumulation of transparent semantic representations with the upper goal being the construction of a goal formula (expressed by means of a type requirement, of the form $?X$, where X is a type). This process is driven by monotonic tree growth, and attempts to model the way information is processed in a time-linear, incremental, word-to-word manner. The following example shows the beginning and the end stage of parsing the sentence *John upset Mary*:

(4)



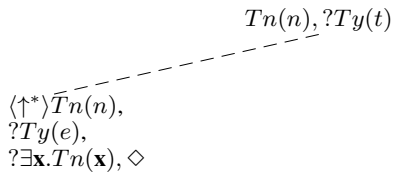
Note that we start with a requirement to obtain a propositional formula (expressed as $?Ty(t)$) and we end up with a tree encoding a propositional and where all its leaves have type and formula values.[2]

[2]Please see Kempson et al. (2001); Cann et al. (2005b) for a detailed exposition of the parsing process and the way composition works incrementally in DS.

2.1 Structural Underspecification and Parsing in Context

DS assumes structural underspecification to be a core feature of language processing. One of the basic mechanisms to encode this structural underspecification is the so-called unfixed nodes. These are nodes that have not yet found their position in the tree structure, their treenode address is underspecified:

(5) *ADJUNCTION: Introducing an unfixed node

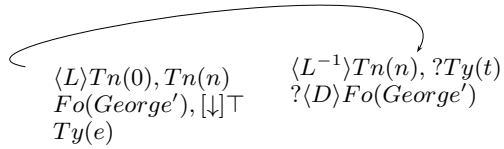


The above rule is introducing an unfixed node. It is structurally underspecified, since it does not carry a fixed treenode address (noted with the requirement $?\exists \mathbf{x}.Tn(\mathbf{x})$, a requirement to obtain a treenode address). The only thing w.r.t to its position it knows is that somewhere up above or at the current node, the address *Tn(n)* must be found. To give an example of where unfixed nodes will be used consider left dislocated structures in English like "John, I know". In this case, the dislocated NP is first parsed on an unfixed node and is only later resolved in the object position, once more information is obtained from parsing the rest of the utterance. I cannot go into more detail here as regards the use of unfixed nodes in dislocated structures, but the interested reader is directed to Kempson et al. (2001); Cann et al. (2005a); Chatzikyriakidis (2010); Kempson and Kiaer (2010); Gregoromichelaki (2013) for more information about dislocation in particular, and the use of unfixed nodes in general.

Besides the tree structures in which each sentence involves a single tree (regardless of tree embedding), DS also makes use of pairs of trees which are linked to each other via a relation called LINK. LINK structures involve two separate tree structures. The node from which the LINK starts can be seen as setting the context in which the LINKed tree is going to be parsed. Examples of LINK relations include relative clauses, in which case the relative clause is parsed within the context of the head noun or Hanging Topic Left Dislocation (HTLD) constructions in which case the HTLD sentence is parsed within the context of

having parsed the left-dislocated element first:

(6) George, I know him



A similar treatment has been proposed for Right Dislocation, the difference being that the LINK structure is now initiated from a type t complete node, i.e. a complete proposition. This idea has been used in the DS literature for BRD, with particular emphasis to pronoun doubling and clitic right dislocation in clitic languages like e.g. Greek (Cann et al., 2005b; Chatzikyriakidis, 2010). We will see that this idea is not enough to capture the idiosyncrasies of ATs. To this, we have to look at the way fragment answers and in general dialogue modeling is done in DS.

2.2 Split Utterances and Fragment Answers in Dynamic Syntax

There is a substantial body of work on formal Dialogue Modeling using DS. Here, we will mention some of the DS literature relevant for the needs of this paper, in particular papers that model split utterances and fragment answers. Two classic examples discussed in a number of papers, e.g. Purver et al. (2010); Kempson et al. (2011, 2012, 2016), are the ones shown below:

(7) A: Did you burn? B: Myself? No.

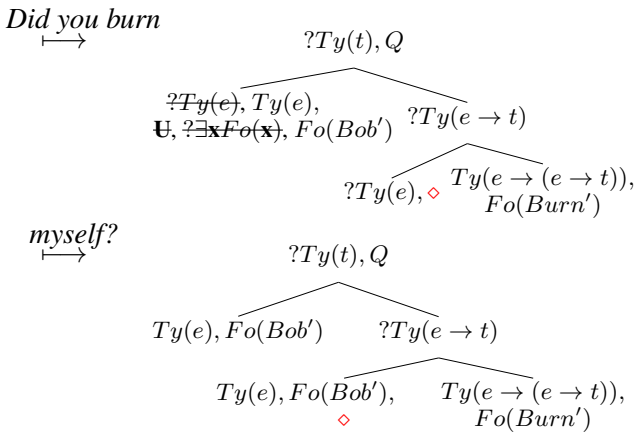
(8) Who hit Mary? John.

What is problematic in the first example, is that the full sentence arising out of the conjunction of the two utterances is ungrammatical, while perfectly fine in a dialogue setting. A formal model of syntax has to be able to accommodate these types of data as well. In DS, production and parsing work tightly together, and the same mechanisms are used in both. The only difference between the two is that during production there is also a subsumption check against a goal tree in every step of the derivation (Otsuka and Purver, 2003; Gregoromichelaki et al., 2012; Eshghi et al., 2011). At any point in the parsing process, the interlocutors can switch the roles of parser and producer at any time. Assuming the following lexical entry for the reflexive, it suffices to have an account of the problematic example:

myself **IF** $\langle\uparrow_0\rangle\langle\uparrow^1_*\rangle\langle\downarrow_0\rangle Fo(x)$
$Speaker(x)$
THEN $\texttt{Substitute}(\mathbf{U}, x)$
ELSE ABORT

The instructions presented in the lexical entry in this simple algorithmic format copy a formula from a local co-argument node onto the current node. This formula must satisfy the conditions set by the person and number of the uttered reflexive (naming the speaker). The result of parsing the split utterance in (7) is shown below:

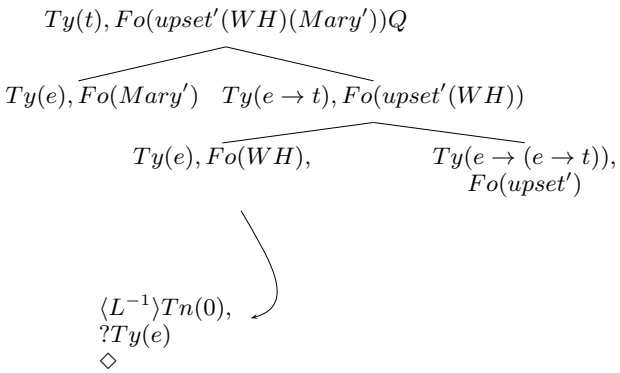
(9)



We see two trees. The first tree is the partial tree after the parse of *did you burn?*. A value has been provided for the metavariable U projected by the second person pronoun *you* (acting as content placeholders to be resolved later, and projected by the lexical entries of pronominals). This is basically a value identifying the hearer. The interlocutor can take on this structure and continue with *myself* with no problem. This is because the speaker in this turn was the hearer in the previous utterance. Given the lexical entry for *myself*, the value provided by *you* by the first participant to identify the hearer, can now be copied to the object node.

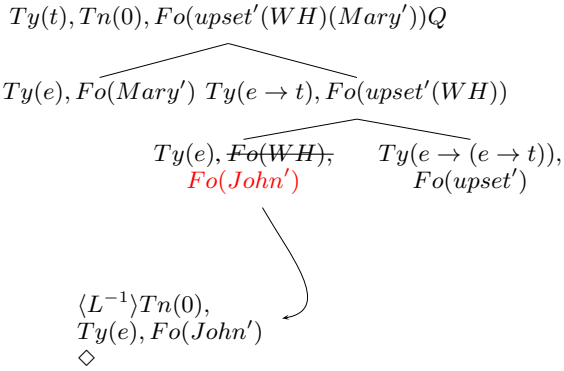
Lastly, let us look at (8), a fragment answer example. This proceeds as follows: the fragment answer is parsed within the context of the WH question, on a LINKed structure. Remember that the structure where the LINK starts can be seen as setting the context for the structure in LINK:

(10) Before parsing the fragment answer



The fragment *John* can be now parsed on the LINKed tree which expects a formula of type e:

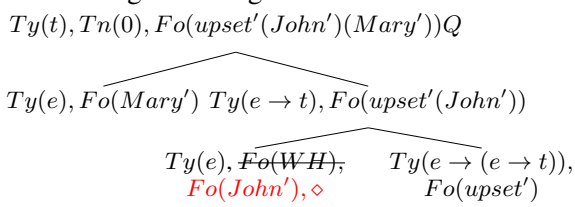
(11) Parsing the fragment and updating the WH metavariable to John



3 Afterthoughts as Fragment Answers to Implicit Questions

The account I want to pursue, as already mentioned, is one where ATs are seen as fragment answers to implicit questions. I will argue that by making this assumption, one can have a straightforward account of ATs that also predicts its syntactic idiosyncrasies. One notational issue: in what follows, I will not be showing the LINK transition between the WH question and the fragment answer as this was done in (11), but rather show the main tree after substitution the WH metavariable has been done. For example, the structure in (11) will be represented as:

(12) Parsing the fragment and substitution

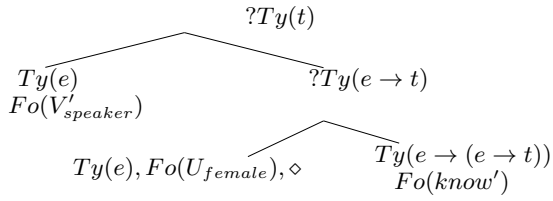


Let us start with a simple AT example:

(13) I know her... Ruth Kempson

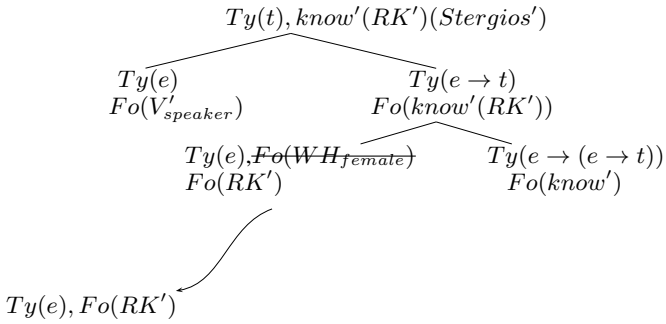
The first step is to parse up to the point of the AT, i.e. we parse "I know her":

Parsing "I know her"



At that point the AT comes into play. Assuming that this is an answer to an implicit question, the sentence parsed/produced so far is turned into a WH-question. In our case, the only difference is the substitution of the U metavariable with a WH metavariable. The AT is then parsed on a LINK structure connected with the main structure, provides a type e value, which is then copied to the main structure, substituting the WH metavariable with a proper value (RK'):

Parsing the AT

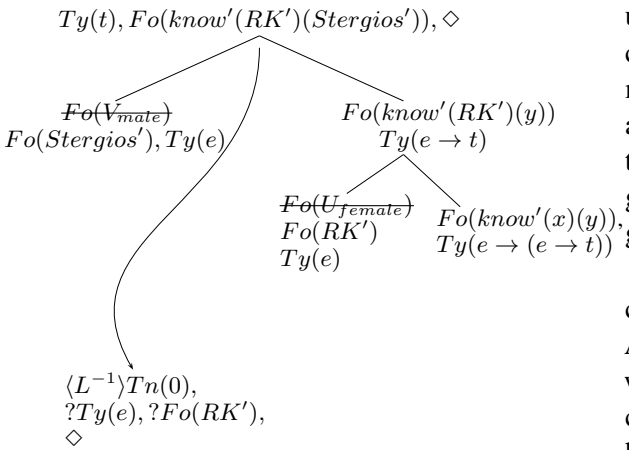


This is the general idea of how ATs function. To recap: ATs are taken to be a corrective mechanism, and as such are assumed to be answers to implicit clarificatory questions. A similar treatment to fragment answers can be used in the case of ATs. A framework that has the ability to handle fragment answers, should be able to handle ATs as well. In our case, we have exemplified this using the DS framework. The next question is whether this account makes the correct predictions as regards the behaviour of ATs in general, but also w.r.t. its differences with regular BRD structures. To this, we turn now.

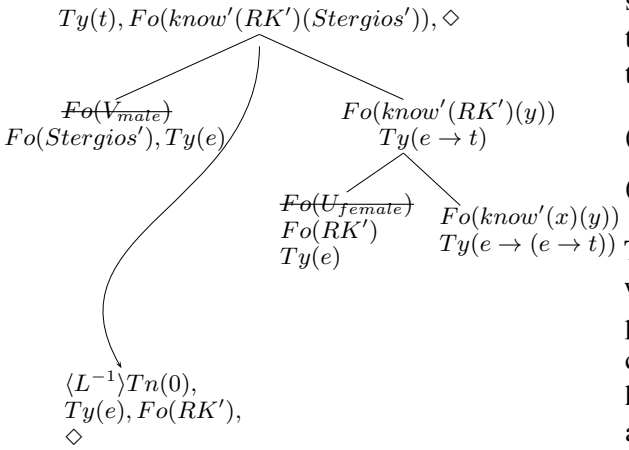
3.1 Afterthoughts and Backgrounded Right Dislocations

What is the difference between regular right dislocated structures, usually referred to as BRDs, and ATs? Well, in terms of their semantic import, as we have already seen, BRDs function as content re-establishers, whereas ATs is a clarification device. How is this to be reflected in the analysis? Well, first of all BRDs do not involve any implicit

questions, they are not parsed in the context of a clarificatory question. The way these structures have been handled in DS assume that a LINK relation is projected from a type t complete tree to a tree that has a requirement that a copy of one of the formulas in the main tree are found on the LINKed one. Let us explain with an example. We look at the same example 'I know her, Ruth Kempson', but on the BRD interpretation (with comma intonation). If BRDs are content re-establishers, what we have here is an optional element that re-establishes part of the content of the sentence 'I know her'. We first parse the sentence 'I know her' and then connect it via LINK to a tree that requires a copy of one of the formulas to be found in that tree, in our case RK':



The main tree has a complete formula that is the result of substituting the metavariables projected by the pronouns with proper values, RK' and $Stergios'$ respectively. There is a requirement for a formula value RK' to be found on the LINKed tree. This is exactly what the RD 'Ruth Kempson' will provide, eliminating the requirement:



Let us start with strict morphological agreement: ATs do not require it, BRDs do. Averintseva-Klisch (2008a) claims that this is the case at least for gender agreement, but does not give any other examples of morphological mismatch. We will concentrate on the gender mismatch case. Let us have a look at an example Averintseva-Klisch (2008a) discusses:

(14) Ich habe ihn$_i$ vorhin gesehen... das
I have him$_{MASC}$ before seen, the
Kleine$_i$ von der Nachbarn
little-one$_{NEUTR}$ of the neighbours

Our account of ATs predicts these gender mismatches quite naturally in the following sense: assuming that ATs are answers to clarification questions, then the relevant WH element in German in the above case is *Wen*. However, *wen* is gender neutral in the grammatical sense, i.e. it is compatible with any gender as a response. Thus, a proper update to the WH metavariable projected by *wen* can be in any gender, a fact that gives rise to gender mismatches. On the other hand, such mismatches are not possible in BRD constructions, given that the pronoun *ihn* will provide a masculine value for gender, that will be incompatible with any other grammatical gender value.

The next difference between BRDs and ATs concerns free positioning in the clause/utterance. ATs can appear freely within the utterance, whereas BRDs are restricted to the end of the clause. Explaining why BRDs exhibit this behaviour is easy: the assumption for BRDs, at least in the way these are handed in DS, is that they involve a LINK relation between a type t complete tree and a tree which needs a copy of one of the formulas in the complete tree. Details aside, the idea is that an utterance has to be considered in some way final, in the sense of providing all the necessary means to provide a full propositional structure, before the RD is parsed in BRDs. Consider the contrast below:

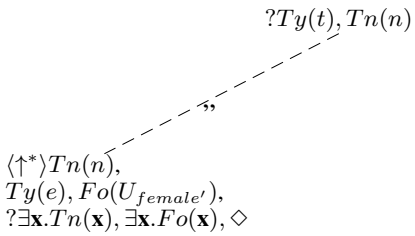
(15) *She, Ruth Kempson, is here

(16) She... Ruth Kempson, is here

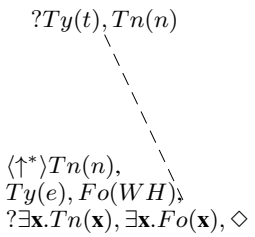
The BRD interpretation is not possible because when the RD comes into parse there is no complete structure to project the LINK from. In the case of the AT, the situation is different. What happens there is that the pronoun *she* gets updated and is turned into a clarificatory question. In terms of representation, what happens is that a regular metavariable, is turned into a WH metavariable, and eventually gets updated to the value RK' after *Ruth Kempson* is parsed. The parse can proceed as

usual from there. The three steps are shown below:

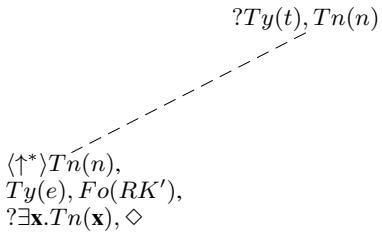
(17) Parsing the strong pronoun *she* in (16)



(18) Clarification question formation



(19) Parsing 'Ruth Kempson' within the context of the clarificatory question



Regarding the possibility of optional additions like *I mean*, *Ich meine* in German, the explanation seems to be straightforward under the proposed account, as well. Whatever our formal analysis is here for questions of the sort *who do you mean?*, will extend to our treatment of ATs with this type of additions. Naturally, there is no way for such optional structures to appear in BRDs, given our account of BRDs.

3.2 Specificational ATs

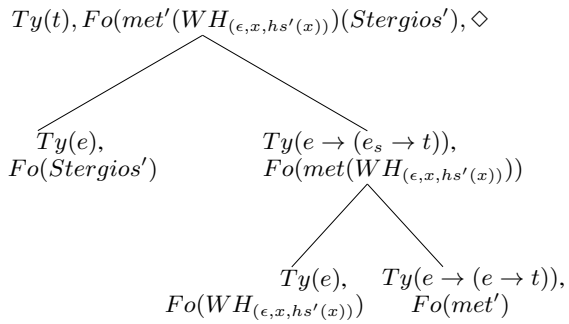
Specificational ATs are different to the type of ATs which we have been dealing with so far, and usually called identificational. The two structures are shown below, (20) a specificational AT and (21) an identificational AT:

(20) I met a hollywood star... John Travolta

(21) I know her... Ruth Kempson

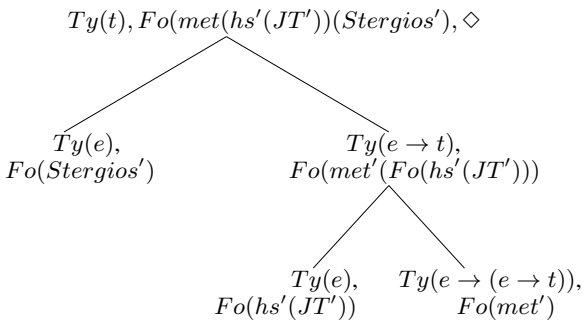
The same account provided so far can be used for specificational ATs, albiet with a minor difference: the implicit question in the case of specificational ATs will not involve a bare WH metavariable but a restricted WH metavariable. For (20) this will be the WH 'which hollywood star'. Thus, the context of the WH question will be the following before the AT is processed in (20):[3]

(22) Tree as context



The subscript on the WH metavariable says that the value that will update the WH metavariable will basically substitute the x metavariable in the epsilon calculus formula $\epsilon, x, hs'(x)$. Thus, the AT is parsed within this context and provides the substitution for x: [4]

(23) After parsing the AT 'John Travolta'



3.3 Afterthoughts and Recursion

Recursion is allowed w.r.t. ATs, i.e. more than one AT is possible in an utterance as witness the example below:

(24) I met her... Mary... yesterday

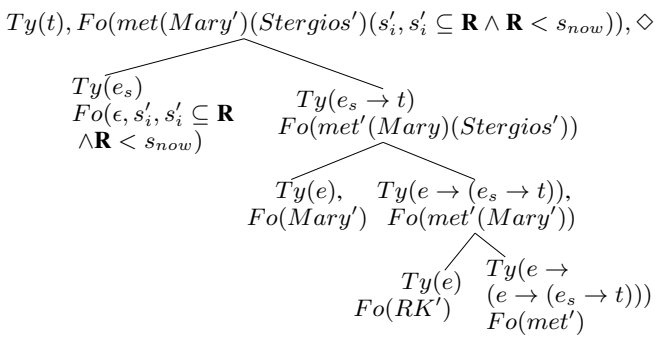
Assuming that ATs is a corrective strategy, this is to be expected if more than one aspect of the utterance needs to be further clarified. The account we have proposed here, which is based on the idea that ATs are basically fragment answers to clarification questions predicts this behavior of ATs. In order

[3] We use the epsilon calculus in the restriction. We cannot really go into details about the epsilon calculus here. It suffices to say that the epsilon calculus is a quantifier free system invented by (Hilbert and Bernays, 1939) and one can derive from the formula $(\epsilon, x, hs'(x))$ the formula $(\exists x.hs'(x))$ in predicate logic.

[4] JT stands for *John Travolta*.

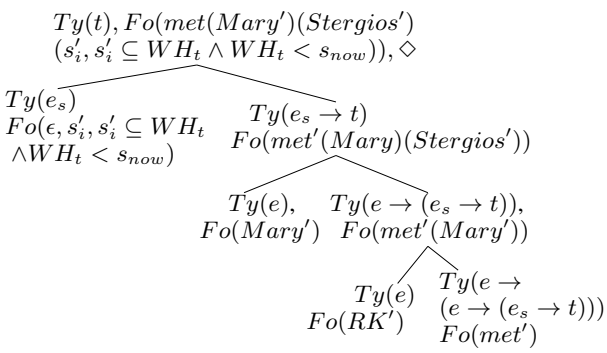
to discuss this example, we will have to make use of the situation/event node. This is the treenode where all tense/aspect information is assumed to be encoded in in DS. I have not been using it so far for reasons of simplicity, since it bore no difference to the actual account so far. However, it is needed for the example we are interested in. Let us see how this works. The tree below depicts the structure after the first AT (the details of how we reach this point are the same as with (13)):

(25) Parsing *I met her... Mary*:



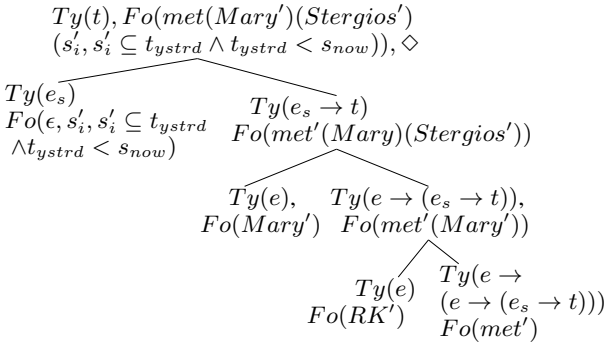
The extra situation argument node is of type e_s (s for situation) and encodes the relevant tense/aspect information. In the case of the simple past the situation is identified as a past situation. It is subsumed inside an interval R that is located in the past ($s_i' \subseteq \mathbf{R} \wedge \mathbf{R} < s_{now}$). Note that the interval is a metavariable, thus can be updated, should there be more specific information on the interval. What happens in the case of the second AT 'yesterday', is that the metavariable **R** is first updated to a WH metavariable ('when'):

(26) Clarification question *when did you meet Mary?*:



Parsing the AT will substitute the WH metavariable WH_t with a proper time interval value, noted here as t_{ystrd}:

(27) Parsing the second AT *yesterday*:



4 Conclusions and Future Work

In this paper I have discussed the case of afterthoughts from the perspective of dialogue modelling. In particular, I have argued that a natural explanation of afterthoughts and their syntactic/semantic properties can be provided once we make the assumption that ATs are answers to clarification questions. I have provided an implementation of this idea in Dynamic Syntax and have shown that a couple of the properties associated with afterthoughts, like morphological mismatches, recursion, free position in the utterance and optional additions.

The next step we want to take is to have a look at the issue of CRification in the sense of Ginzburg (2012) and its potential connection with ATs. In general, everything seems to be clarified in dialogue, and this is what our preliminary data indicate about ATs. If this is true, this will turn out to be a further confirmation that the account is on the correct track. A related issue to explore is the connection between ATs and overanswering.

On a more general level, it would be very interesting to check whether ideas coming from the literature on dialogue, can shed light to phenomena that have been considered difficult to handle in traditional syntactic/semantic formalisms.

Acknowledgments

This work is 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

Maria Averintseva-Klisch. 2008a. to the right of the clause. *Right Dislocation vs. Afterthought . En: C. Fabricius-Hansen, W. Ramm (ed.): Subordination*

versus Coordination in Sentence and Text. Amsterdam: Benjamins, pages 217–239.

Maria Averintseva-Klisch. 2008b. German right dislocation and afterthought in discourse. *Pragmatics and Beyond Series*, 172:225.

Miriam Bouzouita. 2008. At the syntax-pragmatics interface: clitics in the history of spanish. In R. Cooper and R. Kempson, editors, *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, pages 221–264. College Publications, London.

Ronnie Cann, Tami Kaplan, and Ruth Kempson. 2005a. Data at the grammar-pragmatics interface: the case of resumptive pronouns in English. *Lingua*, 115(11):1475–1665. Special Issue: On the Nature of Linguistic Data.

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

Wallace Chafe. 1988. Linking intonation units in spoken english. *Clause combining in grammar and discourse*, 1:27.

S. Chatzikyriakidis. 2010. *Clitics in four Dialects of Modern Greek: A Dynamic Account*. Ph.D. thesis, King's College, London.

Stergios Chatzikyriakidis. 2017. Afterthoughts in greek: Gender mismatches under a dynamic framework 1. *Journal of Linguistics*, 53(2):279–325.

Mark De Vries. 2007. Dislocation and backgrounding. *Linguistics in the Netherlands*, 24(1):235–247.

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

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

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.

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

E. Gregoromichelaki, R. Kempson, and R. Cann. 2012. Language as tools for interaction: Grammar and the dynamics of ellipsis resolution". *Linguistic Review*, 29(4):563–584.

Eleni Gregoromichelaki. 2013. Clitic left dislocation and clitic doubling: A dynamic perspective on left-right asymmetries in greek. *Rightward movement in a comparative perspective. Amsterdam: John Benjamins*.

Eleni Gregoromichelaki, Yo Sato, Ruth Kempson, Andrew Gargett, and Christine Howes. to appear. Dialogue modelling and the remit of core grammar. In *Proceedings of IWCS 2009*.

David Hilbert and Paul Bernays. 1939. *Grundlagen der Mathematik II*. Julius Springer, Berlin.

R. Kempson, E. Gregoromichelaki, and S. Chatzikyriakidis. 2012. Joint utterances in greek and english: implications for linguistic modelling. In *Proceedings of the 33rd Annual Meeting of the Department of Linguistics of Aristotle University*, pages 107–129.

R. Kempson, E. Gregoromichelaki, Wifried M.V, M. Purver, G. White, and R. Cann. 2011. Natural-language syntax as procedures for interpretation: the dynamics of ellipsis construal. In A. Lecomte and S. Tronçon, editors, *Ludics, Dialogue and Interaction*, number 6505 in Lecture Notes in Computer Science, pages 114–133. Springer-Verlag, Berlin/Heidelberg.

R. Kempson and J. Kiaer. 2010. Multiple long-distance scrambling: Syntax as reflections of processing. *Journal of Linguistics*, 46(01):127–192.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as mechanisms for interaction. *Theoretical linguistics*, 42(3-4):203–276.

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

Lutz Marten and Hannah Gibson. 2015. Structure building and thematic constraints in Bantu inversion constructions. *Journal of Linguistics*, 52(3):565–607.

Masayuki Otsuka and Matthew Purver. 2003. Incremental generation by incremental parsing. In *Proceedings of the 6th CLUK Colloquium*, pages 93–100, Edinburgh. CLUK.

Dennis Ott and Mark de Vries. 2014. A biclausal analysis of right-dislocation. In *Proceedings of NELS*, volume 43, pages 41–54.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the 'I's and crossing the 'You's: Context, speech acts and grammar. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, Poznań. Polish Society for Cognitive Science.

Tohru Seraku. 2013. *Clefts, Relatives, and Language Dynamics: The Case of Japanese*. Ph.D. thesis, University of Oxford.

Affordance competition in dialogue: the case of syntactic universals

Eleni Gregoromichelaki
University of Duesseldorf
King's College London
elenigregor@gmail.com

Stergios Chatzikyriakidis
University of Gothenburg
stergios.chatzkyriakidis

Arash Eshghi
Heriot Watt University
eshghi.a@gmail.com

Julian Hough
Queen Mary University of London
j.hough@qmul.ac.uk

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

Ruth Kempson
King's College, London
ruth.kempson@kcl.ac.uk

Jieun Kiaer
Oxford University
jieun.kiaer@orinst.ox.ac.uk

Matthew Purver
Queen Mary University of London
Jožef Stefan Institute, Ljubljana
m.purver@qmul.ac.uk

Mehrnoosh Sadrzadeh
University College London
m.sadrzadeh@ucl.ac.uk

Graham White
Queen Mary University of London
g.graham.white@gmail.com

Abstract

In this paper, we explore the idea that independently developed Dynamic Syntax accounts of dialogue and interaction fit well within the general approach of radical embodied and enactive accounts of cognition (REEC). This approach enables a rethinking of the grounding of linguistic universal constraints, specifically tree structure restrictions, as the outcome of affordance competition, a general REEC sociocognitive mechanism underpinning action selection. Given this subsumption, we argue such an approach opens up a whole new area of language-related dynamic systems research.

1 Introduction

Gregoromichelaki et al. (2019) claim that, in virtue of modelling natural languages (NLs) as actions licensing context-to-context transitions, Dynamic Syntax (DS) fits well within radical embodied and enactive accounts of cognition (REEC) (Bruineberg et al., 2019; Paolo et al., 2018). At present, there is a considerable effort going on to provide accounts of so-called "higher cognition" processes, like memory and imagination, from an enactive perspective that eschews the necessity of representational content (see, e.g., Froese and Izquierdo (2018); Briglia et al. (2018); Hutto and Peeters (2018); Werning (2020); Bruineberg et al. (2019); cf. Gallistel and King (2009)). However, major trends within REEC still remain underdeveloped with respect to the fine-grained details of the traditionally considered par excellence representational medium, NL. While broad outlines of an account have been provided, what is still missing are more specific models of NL-related phenomena standardly conceived as low-level syntactic and semantic issues. Yet in the absence of any attempt to incorporate NL activity in all its complexity within the REEC account as a general nonmodular view of cognition,

its claims remain seriously incomplete. This paper seeks to do three things: First we set out the general embedding of the DS framework within a REEC perspective, introducing the argument that a grammar formalism incorporating psycholinguistic (or "performance") insights and an externalist (inter)action-based view of semantic significance not only provides a natural basis for a model of conversational dialogue but also opens up the possibility of achieving an integration of current research in cognitive neuroscience and linguistic theorising. Secondly, regarding the finer details, we give a more detailed demonstration of how the DS framework, with its action-centric dynamic, has the potential to capture straightforwardly familiar morphosyntactic data for which static frameworks had to add auxiliary hypotheses of considerable complexity. But, on this basis, we can also go one step further towards explaining how the analysis of these phenomena carries over seamlessly when they are embedded in the dynamics of ongoing conversational dialogue. This is an advantage of the dynamic approach which is not available to static frameworks or to individualistic accounts both presuming a competence/performance dichotomy. We then address a putative counterexample to the REEC general claim of the dispensability of standard representational assumptions and the primacy of action. It concerns a robust structural constraint said to hold of all NLs (Kempson et al., 2016 a.o). We argue that such a restriction does not necessitate a representational explanation and can be seen instead as grounded in a combination of socio-cognitive constraints and general properties of dynamical systems (see, e.g., Silberstein and Chemero, 2012) hence fully commensurate with the view that all aspects of cognition are grounded in action without mandatory invocation of brain-internal mental representation.

2 Rethinking the nature of NL grammars

Dynamic Syntax, as set out in Kempson et al. (2001); Cann et al. (2005) and much other work since, is a grammar architecture whose core basis is incremental integration of the contribution of word sequences within a surrounding landscape of affordances. Affordances, under this perspective, are relations between possibilities for action provided by the environment (including the social milieu) and agent abilities available in a 'form of life' (Rietveld and Kiverstein, 2014 a.o.). Within this view, we take the grammar as an integral part of the sociocognitive environment into which humans are enculturated via natural dialogue and interaction. Thus the grammar, as it is being developed and constantly revised, is not seen as qualitatively different from other constraints influencing the environment-engaging behavioural patterns of individuals and groups in the achievement of their aims and goals. It is of course possible to reify these processes of linguistic engagement with the world and assign them abstract structure articulated in representational terms, as in the standard grammatical models in theoretical linguistics or models intended to explicitly teach linguistic skills. Indeed, folk theories of language and thought (at least in some societies) undeniably possess such conceptions of what language is, and we do not doubt that such conceptualisations affect linguistic behaviour. However, we believe that such conceptions are neither basic, nor universal (see, e.g. Linell, 2005). Instead, they constitute further (meta)affordances available within particular social groups for engaging with the available first order NL resources and constraining more systematically the relevant action possibilities.

Grammar – which for us includes what are standardly distinguished as syntax and semantics/pragmatics – constrains human (inter)action by providing a source of normativity, of what is right or wrong, of what makes sense or not, and, hence, what is perceived as rewarding or not, relative to particular social practices. For this reason, grammar is not a construct encapsulated within an individual brain or mind. Following Wittgenstein's well-known arguments against the existence of "private languages" and against an interpretative conception of rule-following (see, e.g., Wittgenstein, 1953), we assume that normative constraints apply within a public domain of expression (even in cases where we privately rehearse responses to a simulation of such interactional challenges in the public domain).

Concomitantly, in the REEC affordance literature, a distinction is made between individual *abilities* which are non-representational capacities allowing individuals to perceive and pick up what opportunities and restrictions are available in their sociomaterial environment and *affordances* which are potential interactions with the environment available within practices or 'forms of life' independently of any particular individual agent.

Grammars, in our view, operate at the level of regulating actions within practices, both linguistic and non-verbal, and, for this reason, are part of the public landscape of affordances available to interacting agents in each particular instance of engagement with the environment. Whether agents interact solely with the physical environment or with other agents, their actions (which not only reveal but actually constitute their conceptualisations) are enabled and restricted by normative constraints. Such constraints which are non-representational ways regulating behavioral patterns (systems of habits) are imposed by the various cultural groups agents inhabit or wish to associate with. Due to membership in various such cultural groups, in each occasion of engagement, agents' actions in turn "enact" and hence modify or enrich the normative constraints available in the current practice.

Words, both as forms and meanings, and syntactic constructions are established patterns of actions that can be fitted in across various 'language games' to enact the nature of the current activity ('semantics') subject to normative judgements emanating from sources outside the acting agent (despite the fact that sophisticated agents can simulate internally such externally sourced judgements). This results in a system constantly in flux (Cooper, 2012) but with enough emergent stability in each particular occasion to underpin agent coordination in the service of various purposes.

In accordance with these assumptions, the DS syntactic engine, including the lexicon, is articulated in terms of goal-driven actions (see also van Benthem, 2011) accomplished either by giving rise to expectations of further action opportunities, by exploiting contextual resources, or by being abandoned as unviable in view of more competitive alternatives. Thus words, syntax, and morphology are all modelled as affordances, opportunities for (inter-)action, produced and recognised by interlocutors to perform step-by-step coordinated mappings from perceivable stimuli (phonological strings) to concept-constituting action patterns (routines, *macros*) or vice-versa (Gregoromichelaki et al., 2019).

The substance of the DS framework is given by a specialised dynamic modal logic (PDL, Propositional Dynamic Logic) whose state-transition language describes a process of gradual unfolding of a diagrammatically laid out relational structure modelling the landscape of salient affordances as a Directed Acyclic Graph (DAG), see Fig 1. DAG nodes (Interaction Control States, ICS) are in turn structured states modelling hubs leading further embedded paths of potential interpretations/productions. Interpretation or production relies on the process of actualising selected paths among the potential provided within local graph structures in the form of trees.

In formal terms, such path traversals at the intermediate level of DAG nodes are enacted and constrained by means of employing in the constraint language specialised modal operators that restrict the trajectories of parsing/generation to incrementally emerging tree-shaped routes. The emergent licensed trajectories are models of sets of constraints articulated within the logic of finite trees (LOFT, Blackburn and Meyer-Viol, 1994). Each node of a tree-structure model is in turn inhabited by graphs resembling feature structures whose attributes are modelled by modal operators bearing linguistically relevant labels like type, $\langle Ty \rangle$, treenode address, $\langle Tn \rangle$, and $\langle Fo \rangle$ for content. In current versions of DS, content values are Record Types (henceforth RT) originating in Type Theory with Records (TTR, Cooper and Ginzburg, 2015) under construction (Eshghi et al., 2013; Hough, 2015; Purver et al., 2011). RTs' inherent underspecification fits well with the requirement to model the potential for indefinite enrichment across various dimensions. RTs (standing for "concepts") can in turn be conceived as mini-grammars of the DS kind articulating affordances for engagement with aspects of the environment (Gregoromichelaki et al., 2019), or potential for interaction with others (Eshghi et al., 2017a; Eshghi and Lemon, 2014). In a parallel effort, Sadrzadeh et al. (2018) show how combining a Vector Space Semantics methodology with DS can model incremental construction of ad hoc concepts to resolve issues of ambiguity and underspecification, allowing for pervasive variability in NL interpretation.

In both these versions of DS, DAGs map out the potential transitions globally available for selection in a particular context of interaction while, locally, at each DAG node (ICS), the potential constructions and transitions are constrained by LOFT axioms that ground tree axiomatisation principles. As DS is a model-theoretic formalism (Pullum and Scholz, 2001), all the inferential

activity modelling incremental parsing/production is defined at the level of constraints articulated through the vocabulary of the relevant logics. The transformation of such PDL and LOFT descriptions is accomplished via general and lexicon-driven mappings. These model patterns of established and relatively stable sequences of basic actions (*macros*). Such macros effect the progressive projection of content, i.e., further opportunities for action, for emergent trajectories following the time-linear order of presentation of a linguistic string. The paths traversed by means of the LOFT modal operators, reflect the constrained structuring of physical and cognitive actions that take place during linguistic processing. Some of the outcomes of such actions can be reified as predicate-argument structures by assuming an external, God's eye, non-modal-logic point of view that abstracts from the process of inducing and traversing such structures. This is a useful and often needed perspective, both for theory construction and for practical human purposes like teaching, metalinguistic reflection and many others but, in our view, it should not obscure the most basic subpersonal level of processing from which it emerges.

2.1 Rethinking syntax as process

Bringing this framework to bear on current views within REEC, Gregoromichelaki et al. (2019) argue that DS is compatible with a view of NLs as activities, "languaging", rather than manipulation of internal knowledge structures that define arbitrary mappings from sound to propositional symbolic representations. This is because DS articulates state-transition mechanisms relative to an ever-evolving interaction context, meshing with nonverbal actions and, for this reason, without needing to only license fixed propositional contents associated with sentence strings. Crucial in this perspective is incorporating in the formal framework the social normativity expressed by the grammar. The native incrementality and action-orientation of DS, as demonstrated by Kempson et al. (2016) amongst others, is well-suited to implement the fine-grained modelling of conversational dialogue dynamics. In particular, we are able to capture the wholly fluent manner in which co-participants in a dialogue switch roles, share utterance responsibility, or effect repair, as the formalism permits the distribution of all syntactic and semantic dependencies across more than one agent.

From this viewpoint, in dialogue, what can be considered as "complete sentences" may emerge through a sequence of fragmentary contributions, with each participant adding some fragment to a partial structure.

(1) Alex: There. To the left.
Hugh: Got it . . .
Hugh: but it's heavy
Eliot: without the lid?
Hugh: if you fancy burning yourself.

In dialogue that is, employing "fragments" and switching of roles between speaking and hearing is widespread because not only the physical context but also the interlocutors' actions and attitudes contribute to the construction of meaning and the achievement of (joint) goals. In such circumstances, incomplete linguistic dependencies can act as prompts soliciting the interlocutor to provide not only physical objects but also utterance continuations:

(2) A: The the ...
B: Here.
A: Thanks. And get a a
B: mattock. For breaking up clods of earth
[BNC (modified)]

(3) Jack: I just returned
Kathy: from
Jack: Finland [Lerner 2004]

Each such contribution can add unproblematically to whatever partial structure has been set out so far, irrespective of whether or not what precedes it is a full sentence. The effect is one of a rich potential for interactivity between dialogue participants.

Moreover, fine-grained morphosyntactic and semantic constraints, with seemingly no functional motivation, apply across the exchange of turns and sharing of utterances so that the initiation of a turn by one participant constrains the options available for the next. For example, in the Romance languages, Modern Greek, and other language families, a well-known morphosyntactic restriction on clitic clusters[1] is the Person Case Constraint (PCC; Perlmutter, 1970; Bonet i Alsina, 1991). In its most general standard variant, the PCC states that

[1] Clitic clusterings are cases in which weakened pronouns occur in a fixed sequence characteristically immediately preceding or following the verb.

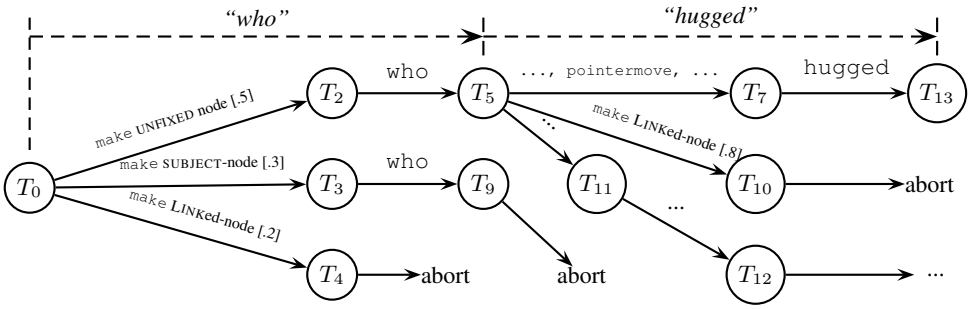


Figure 1: DAG: *Who hugged*

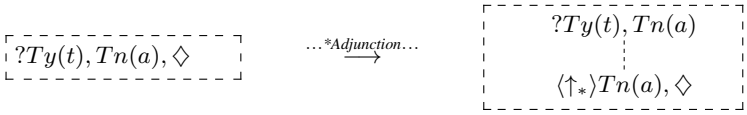


Figure 2: Introducing an 'unfixed' node

Figure 3: *who*-induced actions

a dative clitic cannot co-occur with a 1st/2nd person accusative clitic:

(4) *A: Irthe xtes o Giorgos ke*
A: came yesterday the George and
tis to edose
her$_{\text{CL-DAT}}$ it$_{\text{CL-ACC}}$ gave
'A: Yesterday George came and he gave it to her'

(5) *A: *O Giorgos tu se*
the George him$_{\text{CL-DAT}}$ you$_{\text{CL-ACC}}$
edixe.
showed
'A: George showed you to him' [Greek]

(6) *A: *Le me ha dado*
him$_{\text{CL-DAT}}$ me$_{\text{CL-ACC}}$ has given
'A: S/He has given me to him.' [Spanish]

Notably, such morphosyntactic restrictions also apply across turns:

(7) *A: Irthe xtes o Giorgos ke*
came yesterday the George and
tis...
her$_{\text{CL-DAT}}$
A: 'Yesterday George came and to her ...

B: to edixe?
it$_{\text{CL-ACC}}$ show
B: 'he showed it to her ?'

(8) *A: Irthe xtes o Giorgos ke*
came yesterday the George and
tu...
him$_{\text{CL-DAT}}$
' A: Yesterday George came and to him ...

*B: *se edixe?*
you$_{\text{CL-ACC}}$ showed?
B: 'he showed you to him?'

The attempted continuation of the clitic cluster in the completion + confirmation request in (8) above is unacceptable because the PCC is violated. This shows that the grammar coordinates simultaneously and seamlessly the complementary and multifunctional actions of multiple agents while they are involved in interaction without imposing abstract licensing restrictions that concern single individuals acting on their own. For example, there is no requirement imposed by the grammar that turns need to consist of (overt or covert) complete sentences/propositions to be licensed.

However, as can be seen here, e.g. in (2) where the appropriate word is sought, the set of affordances in a context are perceived differentially by each agent or group of agents depending on their level of attunement to the relevant practices that constitute the source of such affordances, their skills, their attentional state, and previous experiences. Available affordances will

also be partitioned and probabilistically ranked depending on the current concerns and purposes of the agents involved. Agent-relevant affordances are called 'solicitations', i.e., a subset of the available affordances that stand out with respect to particular agents (Rietveld et al., 2018; Bruineberg et al., 2019). We extend this notion to apply to groups of agents. Such demarcated group solicitations within so-called 'fields of affordances' are modelled in DS by the time-linear unfolding of the DAG during conversational interaction, see Fig. 1 (see also Sato, 2011; Eshghi et al., 2013; Hough and Purver, 2012).[2]

Focussing attention at the local level of DAG nodes (ICSs), the first step is to consider the mapping between a string of words, in DS, a sequence of triggers for macros, and the DAG-transformations that it induces. The specialised PDL backbone of DS operates by means of prediction of available open paths of development constrained by so-called *requirements* (these are indicated by ? appearing in front of any available grammatical action). The constraints thus induced by the grammar (and other contextual factors) define a range of goals for the next steps of processing. Agents attuned to the grammatical practices available in a particular context can pursue the most relevant goal defined in this context guided by a process of affordance competition (Cisek and Thura, 2019). For example, a potential starting point of such a process can be a node state as displayed in Fig 2. Here we find the indicator of the current focus of attention, the pointer $\diamondsuit$, and a prediction reflecting the expectation that a proposition ($?Ty(t)$) can be developed. This expectation will lead to many alternative predicted paths of achieving this development that can be displayed as shown in Fig 1. In Fig 2 we pursue one of those paths, the topmost path in the DAG, which is an option made available by the English grammar: a macro called **Adjunction* can introduce a radically structurally underspecified ('unfixed') node predicted to be needed to accommodate the processing of a content-underspecified element like *who* as seen in Fig 3. The latter processed first in a sentence can end up in multiple argument positions eventually so its contribution, a metavariable notated as **WH**, needs to be held in memory until a suitable place is found for it later in the process. Therefore, the identification of the node that accommodates the macro associated with *who* does not specify immediate dominance relations with respect to the root node, as shown in the illustration, Fig. 2-3. This is indicated by the dashed line. The underspecified relation is expressed via the modal operator $\langle\uparrow_*\rangle$ appearing in front of the root node identifier, $Tn(a)$ specifying that the relation between the two nodes is one of (non-immediate) dominance[3].

As can be seen in the topmost path of the DAG, eventually the verb *hug* will be processed and its subject requirement will enforce the contribution of *who* to be specified as such. In DS this means that the unfixed node will now become fully specified with respect to all dominance relations. Once such a DAG path has been successfully traversed, the state reached will record a tree-structured topology of information with no words left to be parsed/generated any more (see Kempson et al., 2001, 2016; Hough, 2015, for details).

2.1.1 A universal constraint?

The constraint on development of such structures which is then proposed for all stages of every DS process is that there be only one copy of any node type at each individual stage of a particular DAG path. Indeed, it has been argued in major support of the framework that the abandonment of universal constraints on structure urged by some (Christiansen and Chater, 2008; Evans and Levinson, 2009; Bybee, 2010; Haspelmath, 2020) can be reversed by the shift to a dynamic system that operates with partial trees and their incremental introduction. This is because universal structural constraints can be shown to hold due to the inferential principles of LOFT, the constraint articulation basis of describing tree graph transitions in DS. So the proposed restriction is a general constraint which follows from the logic of partial tree construction: each node in a partial structure at any stage of the development

[2] In order to simplify presentation, the available macros have been significantly condensed and schematically mentioned through the more central effects they induce; number values indicate toy probability rankings for particular paths; ellipsis (...) indicates that multiple steps have been omitted as they have been judged as irrelevant to the point we wish to make.

[3] $\langle\uparrow_*\rangle$ is the modal operator expressing an immediate dominance relation, $\langle\uparrow_*\rangle$ is the operator which, using the Kleene * operator, expresses the weaker dominance relation. A yet more restricted variant is also defined requiring resolution of the under-specified dominance relation within a single sequence of functor relations, hence within a local predicate-argument structure

process has a unique identity with respect to other nodes in that structure; and this restriction applies equally to *structurally underspecified* nodes whose precise position in the emergent structure is yet to be established. Conceptually, this is a desirable property of the system because we assume that the information accumulated on a node is always underspecified and can be enriched at various stages of processing, even long after a node has been initiated. The restriction on a single uniquely identifying node address ensures that information encountered later in the processing of a string can have access to some particular already existing node, even if such a node is radically underspecified in terms of either position or other properties. Consequently, once a node 'unfixed' relative to particular domain has been introduced no other node can claim a matching underspecified identification. The effect is a "no-copy" principle for any node type. However, as we are going to see presently sometimes this restriction gets in the way, especially with free word-order languages that can exploit word order for information structuring purposes. In attempting to circumvent the effects of this restriction, linguistic patterns might appear that might give the impression of functionally unmotivated arbitrariness and taken as evidence for the "autonomy of syntax" and the particularity of the "language faculty". However, as we are going to argue such patterns do not provide any such evidence. Instead, the confluence of various other affordances and constraints explain the observed patterns.

2.2 Syntactic and morphosyntactic puzzles resolved

This dynamic of processing incorporating parsing and generation under the same formalism has been applied to a range of puzzles concerning syntactic and semantic/pragmatic phenomena as well as their interaction. For example, in verb-final languages, despite very free ordering of all constituents other than the verb, long-distance dependencies are subject to the strict constraint: more than one long-distance "movement" relation is licensed only if the items "dislocated" to the left periphery are co-arguments of some embedded verb. However, locally such co-arguments retain flexibility of ordering with respect to each other as in simple clauses (Pritchett, 1992; Koizumi, 2000):

(9) *Masami-ni* *prezento-o* *Kiyomi-ga*
Masami$_{DAT}$ prezento$_{ACC}$ Kiyomi$_{NOM}$
katta
bought [Japanese]
Kiyomi bought a present for Masami

(10) *Masami-ni* *prezento-o* *John-ga*
Masami$_{DAT}$ prezento$_{ACC}$ John$_{NOM}$
Kiyomi-ga *katta-to* *omotteiru*
Kiyomi$_{NOM}$ bought-COMP believed
John thought Kiyomi bought Masami a present

This flexibility of ordering and the left periphery restriction is characteristic of the verb-final language pattern, by report the most common language pattern among natural languages (Dryer, 2013). But it also occurs in languages with more flexible word orders like Modern Greek where the verb along with the arguments can occur in any order. At first sight, it might appear that this freedom of ordering and piling up of coarguments in the left periphery will be as problematic for DS as it is for other frameworks (see e.g. Koizumi (2000)). If sequences of noun phrases may occur in any order, or, in the embedded case, the left periphery can host any number of arguments, there would appear to be, in either case, a violation of the "no-more-than-one" underspecified relation constraint.

However, as argued in (Kempson and Kiaer, 2010), within DS, there is a simple explanation of this puzzle. The DS system is able to distinguish degrees of underspecification across all grammatical notions, from syntactic to semantic domains. In the case of node position, fine-grained variants of 'unfixed' relations are defined depending on the domain within which the underspecification is confined. One such variant is the case of a node specified solely as dominated by the root node, so an underspecified relation within that global domain. Another variant that naturally arises due to the locality defined by a verb and its arguments is underspecification relative to a more restricted co-argumental domain. And naturally there are transitive implicative relations among such definitions with the global less specified transitions being able to be updated into the more locally-constrained ones (Cann et al., 2005). Interaction with other simultaneous grammatical constraints now solves the puzzle of free ordering.

The solution is provided by considering the role of case morphology. Case in DS has a crucial processing and semantic purpose, namely, the introduction and restriction of further possibilities

of action, instead of just being seen as a feature-matching reflex of some overarching syntactic structure. What this means for free constituent order in languages such as Japanese and Modern Greek is that the simple monoclausal sequence involves the iterated process of introducing an unfixed relation followed by an immediate update step to a specific fixed position provided by processing the case-marking. Once this latter step has been taken, it immediately enables the licensing again of a new underspecified relation this time concerning a newly introduced node. Moreover, this process can also apply within the province of a currently unfixed node with the more embedded underspecification variant concerning a more local domain than the one defined by the overarching underspecified node. However, what is not possible in such a case is that the processing of arguments from various clauses will intermingle. This is because the only underspecification that is available in such circumstances is the more local variety which can only be updated within a local domain. Thus the composite effect of excluding long-distance dependencies from multiple sources but with local ordering flexibility is then directly predicted through the licensed feeding relation between the two differently constrained types of underspecified relation. This modelling of local word order freedom with immediate update despite the lack of a predicate in verb-final languages is notably confirmed by experimental work establishing the incremental nature of Japanese sentence processing by Witzel and Witzel (2016) who demonstrate experimentally that both structural and semantic reference judgements are incrementally made before the verb is processed.[4]

Moreover, surprising confirmation of the restriction on underspecified dependencies and its potential evasion comes from its applicability to the superficially wholly unrelated morphosyntactic phenomenon of clitic clustering. As we saw earlier in (4)-(8) due to the PCC, arbitrary lexical gaps in clustering paradigms are attested. The restriction on unfixed nodes, namely the fact that no more than one treenode with the same address is possible, gives us a natural explanation of the PCC effects because it forces us to look at the observable properties, the morphology, and the interpretation of clitic morphemes within each particular language with no presumption of any allegedly universal syntactic structuring that has to be imposed contrary to the observable facts.

The effects of case are involved again. Dative clitics are in general notoriously underspecified. In Romance, this underspecified nature of the dative can be traced back to Latin, where dative marking is characterised as ambiguous (van Hoeke, 1996) in the sense that, in and of itself, dative does not determine a fixed hierarchical position in the structure. Similar diachronic considerations also apply for other languages that show the same PCC effects, e.g. Greek (Chatzikyriakidis, 2010; Chatzikyriakidis and Kempson, 2011). In DS terms now, given this underspecification of the contribution of the dative, the plausible assumption to make is that dative clitics are processed on an unfixed node, so as to allow variable interpretations of the dative-marked argument depending on the clausal context. Furthermore, 1st and 2nd person accusative clitics in Spanish and many other Romance languages are syncretised with the dative, i.e. the same morphological form is used for both the 1st and 2nd person accusative, as well as the 1st and 2nd person dative clitics. For this reason, 1st and 2nd person accusative clitics can also be considered as underspecified and, under DS assumptions, as processed on an unfixed node too.

Under these underspecification of form assumptions and given the "no copy" constraint, we can predict that any combination of a dative clitic with a 1st or 2nd person accusative clitic will be disallowed. This is because, as we said, two distinct unfixed nodes of the same type cannot be introduced in a single domain. As a result, the ordering Dative-1st/2nd person in, e.g., Modern Greek, will appear as unacceptable due to incompatible information coming from the two types of clitics accumulating on a single node, see (8). On the other hand, 3rd person accusative clitics are not syncretized with the dative and, furthermore, they are always interpreted as direct objects. In DS terms, this means that they are processed within a fully specified structure, rather than on an unfixed node. Thus, the "no copy" constraint does not interfere with the processing of combinations of a dative and a 3rd person accusative clitic, see (7).[5]

[4] See also (Kiaer, 2007, to appear) for extensive discussion and experimental evidence from Korean.

[5] There are a number of variants of the PCC, e.g. the weak PCC version, which allows combinations of 1st and 2nd person pronoun clitics, but this variability has been shown to be afforded by the formal machinery of the system as well (Chatzikyriakidis and Kempson, 2011).

The significance of these results concerning syntactic and morphosyntactic restrictions is that these are observable low-level facts which are usually taken as sui-generis and arbitrary without any grounding in human interaction, meaning, or processing, hence apparently warranting complex and unavoidable stipulation in distinct components of the grammar. Current individualistic linguistic theories justify this split on the basis of notions like competence-performance, modularity (Fodor, 1983), and Marrian computational vs algorithmic level distinctions (Marr, 1982; Steedman, 2000; Kobele, 2012). Within the DS framework, in contrast, these patterns along with their interactional effects (see e.g. (8)) are seen to fall out in virtue of modelling syntax as the progressive shaping of the landscape of affordances for interpretation and production in the everyday coordination of action. Such normative morphosyntactic constraints are explainable within such a grammar because they can be seen as historically sedimented practices routinising the most frequently taken up processing paths (macros), whether as a parser or a producer. Such routinisations are not qualitatively distinct from the other grammatical or lexical constraints given the flexibility afforded by the grammar (as modelled by the various paths within the DAG) even within single instances of interaction. Conceptualised as normative constraints, such macros are necessarily independent from individual NL users and constitute affordances within the processing environment (the 'form of life'). However, in order to be taken up appropriately, the processing agent needs to be attuned to their potential, i.e., they need to appear as 'solicitations' to the agents involved. But the possibility exists that such affordances are potentially inadequately grasped either because the agents do not possess the skills (attunement, 'abilities') required or because the concerns of individual agents and groups prioritise other normative considerations in the competition among which affordances to pursue. Hence the constant potential for innovation and change as well as flexibility and adaptability resulting from novel combinations of affordances pursued. Nevertheless, the availability of such macros in a practice-sharing 'form of life' allows agent coordination in particular interactional episodes due to their operation as manifestly joint relevant affordances (joint 'solicitations').

3 NL within the REEC perspective

There is a range of approaches within REEC usually grounded within dynamical systems models of socio-cognitive phenomena (Chemero, 2009). Most of those espouse non-representational accounts of perception and action, so-called "lower cognition" or "basic minds". However, even advocates of radical enactive perspectives stop short from extending this approach to "higher order" cognition, especially language (see e.g. Clark, 2016; Hutto and Myin, 2012). Moreover, even accounts that aim to develop NL models compatible with dynamical and complex systems approaches like connectionism, due to the individualistic perspective they adopt, have suggested occasionally that neural network implementations strengthen the competence/performance distinction and support the assumption of symbolic representations, albeit of an emergent nature (Prince and Smolensky, 1997).

Against this view, a lot of current REEC research (Rietveld et al., 2018; Bruineberg et al., 2019; Paolo et al., 2018) aims to integrate all levels of cognitive activity within the enactive approach. And DS sides with this approach in that the constraints defined through the logic underlying the framework constitute additional affordances available in the sociomaterial environment of human interactions and are not confined within individual brain structures (Gregoromichelaki, 2018).

DS constraints concern the *process* of comprehension/production in a social context as aspects of general perception/action mechanisms implemented via sensorimotor feedback loops with the environment that do not necessarily engage representational constructs (even though reification of such subprocesses and representational abstraction is also considered possible).

Under such a view, all grammatical dependencies are able to function as word-by-word incremental coordinating devices, i.e., affordances, for either interlocutor. They can do that either by providing forward momentum in the conversation. For example, see (2) earlier for a determiner-noun dependency whose satisfaction completes an already initiated statement but also provides a platform for an enrichment and further specification of the expressed content. Similarly in (3) a preposition-prepositional complement dependency functions as a query speech

act indicator to induce the interlocutor to provide the completion as an answer to the question (Gregoromichelaki et al., 2011, 2013a). Moreover, grammatical constraints can disrupt what is expressible and in what manner by the interlocutor as in (8). In all these instances, grammatical dependencies can function coordinatively sufficiently and efficiently irrespective of whether or not coherent units like the standard representational constructs of 'sentences' or 'propositions' are ever derived or not:

(11) Hester: It's for me.
Mrs Elton, the landlady: And Mr Page?
Hester: is not my husband. But I would rather you continue to think of me as Mrs. Page. [The Deep Blue Sea (film)][6]

(12) Jem: Mary, whatever it is you think you know you mustn't speak of it, not if you want to stay safe.
Mary: says the horse-thief
[*Jamaica Inn*, BBC Transcripts]

(13) A: SOMEONE is keen.
B: says the man who slept here all night
[*The A-Word*, BBC Transcripts]

Under DS assumptions, in (11), we don't need to invoke pro-drop in English to explain the missing subject of Hester's utterance. Instead, we can implement the intuitive explanation that the subject is provided due to the fact that the utterance is a continuation of Mrs Elton's interrogative non-sentential query in the previous turn. For both utterances, there is no reason to assume that speech acts are licensed only if interlocutors can derive fully specified propositional contents and/or sentential structures embedded under speech act predicates. Neither us, as a third party reader, nor Hester, nor Mrs Elton herself, have particular predicates in mind when encountering and understanding "And Mr Page?" in (11) since its affordances mesh with the rest of the affordances in our respective contexts and allow us to go on interacting with the utterance or its utterer. And it is precisely this lack of predicate, not only in the surface structure but also in the conceptual structure, that allows Hester to make use of the previous utterance to build her own which is thus fulfilling the pending affordance for specification that has been introduced by the verbless utterance.

In contrast, in (12)-(13), the previous utterance, which appears to be a fully-specified independent sentence and proposition, is taken over by the interlocutor at the next processing stage. As a result, it becomes embedded under a verbum dicendi thus losing its supposed independence as a "complete thought" and becoming instead part of a report that uses the original speaker themselves as the 'animator' (Goffman, 1981) of the message now subsumed under the authorial control of the second speaker but remaining, nevertheless, a joint construction. Hence the effectiveness as a hostile move of this construction in English. But to achieve the modelling of this effect it is essential that we do *not* assume that there is a *copy* of the original utterance embedded under the reporting verb and under the second speaker's voice. Instead, the second speaker's utterance employs the original speaker's own words against them, and that is the whole point of the construction (Gregoromichelaki, 2018)

So, instead of propositional knowledge of rules and manipulation of representations, the abilities of individual agents for interacting efficiently with others under the guidance of grammatical constraints could be characterised as subpersonal mechanisms which allow access to a normative landscape of affordances. Access to this landscape is then mediated by inducing a range of predictive goals (solicitations) to be fulfilled by either interlocutor in the very next steps which they will be taking. The skills required to take advantage of solicitations in such a context do not presuppose any 'rational' high-order individualistic inference, standardly taken to be the basis of all successful human communication (Clark (1996); Sperber and Wilson (1995) and many others).

Instead, the task of selecting appropriate actions is taken over by affordance competition, which operates at the much lower level of sensorimotor contingencies (Cisek and Thura, 2019; Anderson, 2014; Rietveld et al., 2018). However, this presupposes that not only does the grammar incorporate processing features like incrementality and predictivity but also that the grammar provides a shared "workspace" (Kempen, 2014) for both production and comprehension to operate and interact (Gregoromichelaki et al., 2013b).

In addition, non-linguistic practices need to

[6] Along with natural data, constructed data from literature, film scripts etc. are particularly relevant in this context as they show that such constructions are not "speech errors" that can be easily dispensed with.

be integrated with NL actions to contribute their qualitatively identical types of constraints (Gregoromichelaki, 2018). Evidence that morphosyntax directly interacts with embodied situational affordances of every day action comes from elliptical constructions. For example, in case-rich languages such as German and Greek, elliptical fragments necessarily display the form suitable to what might have been a complete sentence formulated in response, even when there is no obvious antecedent of a verbal sentential form– in the German example in (14) the accusative form *den Arzt* seamlessly blends with other situational affordances to constrain future action to the effect that someone should call a doctor:[7]

(14) *A and B see a woman lying on the floor:*
A to B: Schnell, den Arzt/#der Arzt
Quick, the doctor$_{ACC}$*/#the doctor*$_{NOM}$

This shows that what is called an 'antecedent' in linguistic accounts of ellipsis need not have linguistic form, and can in fact be an action concurrent with the utterance, or even in the future of it as in commands like (14). Moreover, physical actions can combine with linguistic actions to seamlessly contribute constituents like subjects and objects to sentences (see, e.g., Slama-Cazacu, 1976):

(15) while sitting on the piano: He doesn't play [PLAYS TUNE]. But rather he plays –he does it better than I do– [PLAYS TUNE - SINGS]
[Clark, 2016]

It seems then that NLs are not underpinned by distinctive mechanisms residing within individual minds encapsulated from the demands and contingencies of interaction with the sociomaterial environment.

3.1 Rethinking NL Autonomy

The question that then remains is how to think of alleged sui-generis properties of NL morphosyntax that have been adduced as arguments for claims like "the autonomy of syntax" hypothesis. Within DS, this question is pertinent regarding the status of claimed universal NL processing constraints like the "no-copy" restriction. While this remains an issue for much further development, we suggest that, within a domain-general framework like DS, there are grounds for seeing this restriction as a general control property of socio-cognitive coordination systems (control in the cybernetic sense of 'regulation', e.g. Bickhard (2009); Carver and Scheier (2012)) and, therefore, indeed a general property of physical systems involving the interaction of multiple simultaneous processes.

In this connection, Anderson (2014) attributes to individual brain mechanisms ('abilities') the function of action control via perception/action feedback loops with the environment without necessary representational mediation. Like DS, he sees "higher" cognitive processes like NL as developments of already existing brain capacities, rather than as species-specific cognitive adaptations. The brain is modelled as a connectionist network, superimposed on the physical brain structure, inspired by Smolensky (1986)'s architecture but without the individualistic representational interpretation originally assigned to this model. The state transitions and attractor landscapes of this network control the interaction of the individual and the environment without presuming that the brain constructs a model of the world it interacts with. As in REEC's view of skills, abilities, and dispositions, the contribution of the individual brain is seen as complementary to the equally significant roles of the whole body and the environment. Brain activity on its own is in no way to be considered as enacting conceptually-transparent reactions to the environment. Some sequences of states can be assigned interpretations as individuated action goals but it is only at the end of a sequence of preparatory stages during which multiple partially activated goals coexist and interact with environmental inputs that such an abstractive interpretation becomes possible. Accordingly, the states that the network passes through are not assigned representational contents but are individuated by their responsiveness to particular inputs and the behaviours that they enable. This control of behavior architecture also dispenses with a central decision module that processes abstract goals prior to action. Instead, different patterns of activation at each state track the gradual convergence of the totality of the control mechanism towards a particular action, that is, they track the "decision" process to pursue a selected subset of the available affordances that compete for realisation within an agent's environment. What is im-

[7] For a critique of views that purport to reduce all fragments to a sententialist form of explanation, see Kempson et al. (2018).

portant for us here is that, at preliminary stages, affordance competition in Anderson's interpretation is implemented by the fact that neural patterns of activation might reflect simultaneously multiple partially-activated but incompatible options. This is only possible before final convergence to the selected action goal which can only consist of the pursuance of multiple but compatible affordances. Architecturally, the potential for the presence of incompatible options arises because various incompatible affordances can engage the activation of the same network nodes as long as they are not fully activated. At the final update stage of convergence (i.e. when a "decision" has been reached), the partial specifications of multiple action goals will be necessarily eliminated if their full activation patterns are incompatible, i.e., if they require the same network nodes for their realisation. The nonlinear update function ensures this. Only affordances with the strongest activation, having been gradually reinforced through environmental inputs, will survive. Their prevalence means that they will assume the role of controlling action as immediate agent goals while disrupting and inhibiting the activation patterns of any competitors.

We suggest that the 'no-copy' constraint on LOFT tree-path traversals is of the same type and aetiology. The graph structure imposed by LOFT instantiates an action control mechanism within the DS conception of NLs as grounded in goal-directed action. Restrictions such as the 'no-copy constraint' then provide a restricted architectural bottleneck to facilitate action selection in view of multiple competing action opportunities in the environment. To take a concrete example, we assume that an 'unfixed node' state can accommodate multiple affordances as goals. In the case of the Greek clitic clusters (see (4)-(8)), during production, multiple gender, person, and number specifications, compete for realisation at the sequential cluster positions. Each clitic morpheme at a particular cluster position when processed will trigger a combination of affordances (macros) appearing on a single treenode. But given the multiplicity of competing specifications, the eventual combination of affordance annotations on any treenode corresponding to a single cluster position will be the outcome of a convergence after affordance competition. In this sense, each treenode's specification of action goals is the equivalent of fully-specified activation patterns which cannot be realised by neurons serving overlapping action specifications. Similarly, in DS, a node can only accommodate compatible selections of multiple macros as requirements for further update, for example, a macro associated with morphological marking for gender 'female' can combine with a macro for '3rd person singular' but not with a macro for gender 'masculine'. If inconsistent macros come to appear as goals at the same node the result will be anomaly (technically implemented in PDL as reaching a state where absurdity, $\perp$, holds or, equivalently, that the ABORT action is implemented). The reason is that there will be no possibility of pursuing the action paths indicated as subgoals by those macros and reaching a consistent state, hence the action control mechanism provided by the grammar will have failed its purpose.

As a result, in a language like Modern Greek with its otherwise generally free word order we do nevertheless find phenomena of strict ordering and epiphenomenally arbitrary positional exclusions. The processing reason for these phenomena though is that any underspecified clitic morpheme will need to be processed on an unfixed node whose specifications will be a particular combination of compatible affordances. Any other similarly underspecified clitic cannot then follow because its morphological specifications will annotate the same node leading to inconsistency. For this reason, whichever clitic morpheme is initially more strongly favoured by the contextual expressive needs of the interlocutor will win the competition among the various candidate combinations and will get to occupy the relevant cluster position thus excluding seemingly unaccountably the appearance of other morphemes not only in the current position but also in sequentially later ones. The only resort remaining for the expression of such a defeated and thus excluded combination of specifications will then be to appear not as a clitic but as a full NP, pronominal or nominal.

So, affordance competition is the underlying cause for action selection, action control, both facilitation and obstruction, with NL surface structural constraints being subsumed under an overall behaviour-guidance and general systems-control architecture.

In closing, we note that, besides capturing non-determinism and its structural effects in the updates achieved by morpho-syntactically induced

actions, a further integral task in this program is to capture the enormous range of interpretation possibilities which words display, a challenge often treated as peripheral for grammar frameworks. However, in our view, given the domain-generality of DS, the same mechanisms set out here can account for contextual modulation of word meaning in terms of updates selecting, through affordance competition, the relevant aspects of meaning of words in context. Words can be seen as triggering packages of sets of potential actions (macros) formalised as TTR types (Gregoromichelaki et al., 2019).

Accounting for contextual shift in meaning, current DS work on the incorporation of a vector-based semantics for the affordances words provide is ongoing, with the goal of modelling the pervasive variability of word meaning and its incremental resolution, without invariably invoking lexical ambiguity. This approach builds on previous work on compositional distributional semantics for pregroup grammars and the Lambek Calculus (Coecke et al., 2010, 2013), and for Combinatory Categorial Grammar (CCG) (Maillard et al., 2014; Moot, 2018). There is current work in this area, where distributional meanings of words are combined with their grammatical types, albeit more informally (e.g. Baroni and Zamparelli (2010)). But these do not consider an incremental setting with underspecified nodes in the grammatical construction, a challenge which our program addresses. In this vein, Purver et al. (forthcoming) suggest two distributional counterparts to underspecificational notions, and present experimental results regarding affordance competition that results in optimal choice in the disambiguation task

4 Conclusion

The shift of emphasis from static structural generalisations to actions is central here. The goal of NL processing and NL interaction is not to generate a sentence structure or proposition as such: it is to promote progressive context-updates via pursuing selected action opportunities (affordances) yielding further affordances guided by the various practices constraining and enabling the actions of the interlocutors. Given the DS architecture's domain-generality, constraints arising from non-verbal practices blend seamlessly with morpho-syntactic and semantic constraints. Context-updates ensuing from NL sequential signals follow constrained trajectories modelled as graph-transitions of restricted shapes, with an intermediate level of tree-structured paths. This structuring accounts for apparent idiosyncratic morphosyntactic constraints because it channels processing towards accomplishing interpretation and production in an incremental manner by restricting the available predictions. Considerable but not unconstrained flexibility is achieved due to various types of licensed underspecification and updates and action selection requirements are implemented despite the multiplicity of action opportunities in the environment. During the process of formulating or comprehending an utterance, the initial availability of various options for update is gradually restricted via incremental culling of options. This is carried out as constraints accumulate from various sources and at various stages, with a process of affordance competition serving as the necessary action selection mechanism. This mechanism replaces "high order" cognitive individualistic and internalist notions like intention recognition, mind-reading, and planning (see also Mills and Gregoromichelaki, 2010; Silberstein and Chemero, 2011; Froese, 2018) and implements relevance considerations at the low-level of subpersonal mechanisms of action control. Within this perspective, it is notable how this action control mechanism applies across sources of constraints bridging the levels of both "lower-" and "high-level" cognition.

We believe that, by abandoning representational individualism and grasping the potential that rethinking language as an action system offers, we can explain in an integrated way the dynamics of dialogue interaction and the seamless interweaving of NL utterances with nonverbal actions. Additionally, given that NL constraints are here subsumed under a theory of general action, the practices guiding NL users are a source of normativity independent of themselves, hence avoiding "private language" Wittgensteinian objections. On the other hand, the established practices guiding action need to be enacted afresh every time in new contexts, and either reinforced if successful or modified if ineffective (see, e.g., Eshghi et al., 2017b). We can then get a handle on both the apparent stability of a language for its users and yet its endless potential for ongoing variability and change. Furthermore, combining these perspectives, the ontogenetic trajectory of acquiring a first

language can then be seen not as maturational development executing a genetic program but as a grounded process crucially involving interaction with the social environment and scaffolded by non-verbal expressive and communicative joint actions like gesture and turn-taking (Clark and Kelly, to appear).

Acknowledgements

We wish to thank three anonymous reviewers for their careful reading and thought-provoking comments which led, in our minds, to significant improvements to this paper. We would also like to thank the Technical Programme Chairs for important practical assistance.

Chatzikyriakidis and Howes were supported by grant 2014-39 from the Swedish Research Council, (VR), 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. Howes was also supported by the VR grant 2016-0116 – Incremental Reasoning in Dialogue (IncReD). Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

References

M Eulalia Bonet i Alsina. 1991. *Morphology after syntax–pronominal clitics in romance*. Ph.D. thesis, Massachusetts Institute of Technology.

Michael L. Anderson. 2014. *After Phrenology*. Cambridge University Press.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge MA. Association for Computational Linguistics.

Johan van Benthem. 2011. *Logical dynamics of information and interaction*. Cambridge University Press.

Mark H Bickhard. 2009. The interactivist model. *Synthese*, 166(3):547–591.

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.

J. Briglia, P. Servajean, A. H. Michalland, L. Brunel, and D. Brouillet. 2018. Modeling an enactivist multiple-trace memory. ATHENA: A fractal model of human memory. *Journal of Mathematical Psychology*, 82:97–110.

Jelle Bruineberg, Anthony Chemero, and Erik Rietveld. 2019. General ecological information supports engagement with affordances for 'higher' cognition. *Synthese*, 196(12):5231–5251.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridgge University Press.

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

Charles S Carver and Michael F Scheier. 2012. Cybernetic control processes and the self-regulation of behavior. In *The Oxford handbook of human motivation*, pages 28–42. Oxford University Press New York, NY.

Stergios Chatzikyriakidis. 2010. *Clitics in four dialects of Modern Greek: A dynamic account*. Ph.D. thesis, King's College, London.

Stergios Chatzikyriakidis and Ruth Kempson. 2011. Standard Modern and Pontic Greek person restrictions: A feature-free dynamic account. *Journal of Greek Lingusitics*, pages 127–166.

Anthony Chemero. 2009. *Radical Embodied Cognitive Science*. MIT Press.

Morten Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Brain and Behavioral Sciences*, 31:489–509.

Paul Cisek and David Thura. 2019. *Reach-to-Grasp Behavior*, chapter Neural circuits for action selection. Routledge.

Andy Clark. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, Oxford.

Eve Clark and Barbara Kelly. to appear. Constructing a system of communication with gestures and words. In Goldin-Meadow and Morgenstern Aliyah, editors, *Gesture in language*. Mouton de Gruyter.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Bob Coecke, Edward Grefenstette, and Mehrnoosh Sadrzadeh. 2013. Lambek vs. lambek: Functorial vector space semantics and string diagrams for lambek calculus. *Annals of Pure and Applied Logic*, 164(11):1079 – 1100. Special issue on Seventh Workshop on Games for Logic and Programming Languages (GaLoP VII).

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1–4):345–384. Special Issue: A Festchrift for Joachim Lambek.

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

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

Matthew S. Dryer. 2013. Order of subject, object and verb. In Matthew S Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Arash Eshghi, Julian Hough, and Matthew Purver. 2013. Incremental grammar induction from child-directed dialogue utterances. In *Proceedings of the 4th Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 94–103, Sofia, Bulgaria. Association for Computational Linguistics.

Arash Eshghi and Oliver Lemon. 2014. How domain-general can we be? Learning incremental dialogue systems without dialogue acts. In *Proceedings of Semdial 2014 (DialWatt)*.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017a. Bootstrapping incremental dialogue systems from minimal data: the generalisation power of dialogue grammars? In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017b. Interactional Dynamics and the Emergence of Language Games. In *Proceedings of ESSLLI 2017*.

Nicholas Evans and Stephen Levinson. 2009. The myth of language universals: language diversity and its importance for cognitive science. *Brain and Behavioral Sciences*, 32:429–492.

Jerry A. Fodor. 1983. *The Modularity of Mind*. MIT Press.

Tom Froese. 2018. Searching for the conditions of genuine intersubjectivity: from agent-based models to perceptual crossing experiments. In Albert Newen, Leon De Bruin, and Shaun Gallagher, editors, *The Oxford handbook of 4e cognition*, pages 163–186. Oxford University Press, Oxford, UK.

Tom Froese and E. J. Izquierdo. 2018. A Dynamical Approach to the Phenomenology of Body Memory: Past Interactions Can Shape Present Capacities Without Neuroplasticity. *Journal of Consciousness Studies*, 25(7-8):20–46(27).

C. Randy Gallistel and Adam Philip King. 2009. *Memory and the Computational Brain*. Wiley-Blackwell, Oxford, UK.

E. Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press.

Eleni Gregoromichelaki. 2018. Quotation in dialogue. In *The Semantics and Pragmatics of Quotation*, pages 195–255. Springer.

Eleni Gregoromichelaki, Ronnie Cann, and Ruth Kempson. 2013a. On coordination in dialogue: subsentential talk and its implications. In Laurence Goldstein, editor, *On Brevity*. OUP.

Eleni Gregoromichelaki, Christine Howes, and Ruth Kempson. 2019. Actionism in syntax and semantics. In *Gothenburg Proceedings of Conference on Dialogue and Perception.* CLASP.

Eleni Gregoromichelaki, Ruth Kempson, Christine Howes, and Arash Eshghi. 2013b. On making syntax dynamic: The challenge of compound utterances and the architecture of the grammar. In Ipke Wachsmuth, Jan de Ruiter, Petra Jaecks, and Stefan Kopp, editors, *Alignment in Communication: Towards a New Theory of Communication*, Advances in Interaction Studies (series editors: Kerstin Dautenhahn, Angelo Cangelosi), pages 57–85. John Benjamin.

Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, Greg J. Mills, Ronnie Cann, Wilfried Meyer-Viol, and Patrick G. T. Healey. 2011. Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1):199–233.

Martin Haspelmath. 2020. Human linguisticality and the building blocks of languages. *Frontiers in Psychology*, 10:1–10.

Willy van Hoeke. 1996. The Latin dative. In William van Belle and Willy Langendonck, editors, *The Dative Volume 1: Descriptive Studies*, pages 3–38. John Benjamins.

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

Julian Hough and Matthew Purver. 2012. Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the 16th SemDial Workshop on the Semantics and Pragmatics of Dialogue (SeineDial)*, pages 136–144, Paris, France.

Daniel D. Hutto and Erik Myin. 2012. *Radicalizing Enactivism: Basic Minds without Content*. MIT Press.

Daniel D. Hutto and Anco Peeters. 2018. The Roots of Remembering: Radically Enactive Recollecting. In Kourken Michaelian, Dorothea Debus, and Denis Perrin, editors, *New Directions in the Philosophy of Memory*, pages 97–118. New York: Routledge.

Gerard Kempen. 2014. Prolegomena to a neurocomputational architecture for human grammatical encoding and decoding. *Neuroinformatics*, 12(1):111–142.

R. Kempson and J. Kiaer. 2010. Multiple long-distance scrambling: Syntax as reflections of processing. *Journal of Linguistics*, 46(01):127–192.

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

Ruth Kempson, Eleni Gregoromichelaki, Arash Eshghi, and Julian Hough. 2018. Ellipsis in Dynamic Syntax. In Jeroen van Craenbroeck and Tanya Temmerman, editors, *Oxford Handbook of Ellipsis*. Oxford University Press.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax*. Blackwell.

Jieun Kiaer. 2007. *Processes and Interfaces in Syntactic Theory: the case of Korean*. Ph.D. thesis, King's College London.

Jieun Kiaer. to appear. *Pragmatic Syntax*. Bloomsbury.

Gregory M Kobele. 2012. Eliding the derivation: a minimalist formalization of ellipsis. In *Proceedings of the 19th International Conference on HPSG*.

Masatoshi Koizumi. 2000. String vacuous overt verb raising. *Journal of East Asian Linguistics*, 9:227–85.

P. Linell. 2005. *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. Psychology Press.

Jean Maillard, Stephen Clark, and Edward Grefenstette. 2014. A type-driven tensor-based semantics for CCG. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 46–54, Gothenburg, Sweden. Association for Computational Linguistics.

David Marr. 1982. *Vision: A Computational Approach*. Freeman & Co, San Francisco.

G. Mills and E. Gregoromichelaki. 2010. Establishing coherence in dialogue: sequentiality, intentions and negotiation. In *Proceedings of SemDial (PozDial)*.

Richard Moot. 2018. Combining logical and distributional methods in type-logical grammars. *Journal of Language Modelling*, 6:288–317.

Ezequiel A. Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. 2018. *Linguistic Bodies: The Continuity between Life and Language*. MIT Press.

David M Perlmutter. 1970. Surface structure constraints in syntax. *Linguistic inquiry*, 1(2):187–255.

Alan Prince and Paul Smolensky. 1997. Optimality: From neural networks to universal grammar. *Science*, 275(5306):1604–1610.

Bradley L. Pritchett. 1992. *Parsing with Grammar: Islands, Heads, and Garden Paths*, pages 321–349. Springer Netherlands, Dordrecht.

Geoffrey Pullum and Barbara Scholz. 2001. On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In *Proceedings of the 4th International Conference on Logical Aspects of Computational Linguistics (LACL)*, pages 17–43. Springer.

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

Matthew Purver, Mehrnoosh Sadrzadeh, Ruth Kempson, Gijs Wijnholds, and Julian Hough. forthcoming. Incremental composition in Distributional Semantics. *Journal of Logic, Language and Information*.

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*, page 41. Oxford University Press.

Erik Rietveld and Julian Kiverstein. 2014. A rich landscape of affordances. *Ecological Psychology*, 46.

Mehrnoosh Sadrzadeh, Matthew Purver, Julian Hough, and Ruth Kempson. 2018. Exploring semantic incrementality with dynamic syntax and vector space semantics. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2018 - AixDial)*, Aix-en-Provence.

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.

Michael Silberstein and Anthony Chemero. 2011. Dynamics, Agency and Intentional Action. *Humana Mente*, 4(15):1–19.

Michael Silberstein and Anthony Chemero. 2012. Complexity and Extended Phenomenological-Cognitive Systems. *Topics in Cognitive Science*, 4(1):35–50.

T. Slama-Cazacu. 1976. Nonverbal components in message sequence: 'mixed syntax'. In W.C. McCormack and S.A. Wurm, editors, *Language and Man: Anthropological Issues. Mouton, The Hague*, pages 217–227.

Paul Smolensky. 1986. *Parallel distributed processing*, volume 2, chapter Neural and conceptual interpretation on PDP models. MIT Press, Cambridge.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*, second edition. Blackwell.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA.

Markus Werning. 2020. Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation. *Review of Philosophy and Psychology*, 11(2):301–333.

L. Wittgenstein. 1953. *Philosophical Investigations, Trans. G.E.M. Anscombe*. Oxford: Blackwell.

Jeffrey Witzel and Naoko Witzel. 2016. Incremental sentence processing in japanese: a maze investigation into scrambled and control sentences. *Journal of Psycholinguistic Research*, 45:475–505.

A Multi-level Framework for Understanding Spoken Dialogue Using Topic Detection

Andrew P. Valenti, Ravenna Thielstrom, Felix Gervits, Michael Gold, Derek Egolf, Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University, Medford, MA 02155
{andrew.valenti, ravenna.thielstrom, felix.gervits, michael.gold, derek.egolf, matthias.scheutz}@tufts.edu

Abstract

Spoken Dialogue Systems (SDS) are used to interact with intelligent agents through natural language. Speech processing errors may cause the system to fail to generate an appropriate response. In this paper, we present a novel framework for understanding spoken dialogue in which utterance analysis is escalated through a multi-level system *according to the feedback retrieved at the syntactic, semantic, and contextual/topic level*. Analysis is applied incrementally at each level as the system attempts to resolve the uncertainty surrounding utterance interpretation. We demonstrate how our multi-level approach can be integrated with other SDS components to improve its ability to recognize spoken task commands. We evaluate this by comparing the interpretation accuracy of utterances from two task domains given as input to an SDS, under two experimental conditions: one with the multi-level framework and one without.

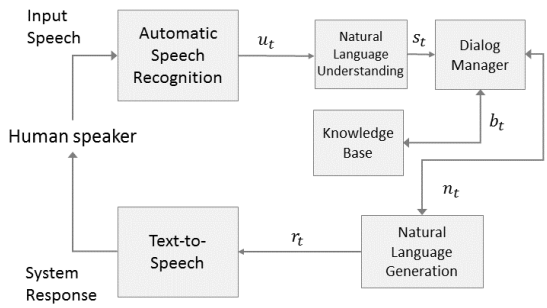


Figure 1: Typical spoken dialogue system. At each turn t, input speech is converted to an utterance, u_t, which the Natural Language Understanding component maps to an internal representation, s_t of the human's intent. The Dialogue Manager uses this to update the agent's belief state in the Knowledge Base, b_t and then infers a natural language form, n_t from Natural Language Generation which initiates a response, r_t to the Text-to-Speech component.

1 Introduction

Humans use Spoken Dialogue Systems (SDSs) to interact with intelligent agents using speech-based natural language (De Mori, 1997; Zue and Glass, 2000; Jokinen and McTear, 2009). Figure 1 shows components typically found in such systems (Scheutz et al., 2019; Young et al., 2013). Here, the Automatic Speech Recognizer (ASR) recognizes the human's utterance, u_t, and sends it to the Natural Language Understanding (NLU) component. However, problems may arise in the ASR component which can propagate through the system and cause it to fail to generate an appropriate response. For example, the ASR may recognize the word "Iraq" instead of "a rock" (Sarma and Palmer, 2004), or it may hear a novel word it has not yet learned (Scheutz et al., 2017). Alternatively, the user may believe the system to be capable of retrieving the weather report when its domain is retrieving movie listings; in such a case, the system will need to respond to the user's out-of-domain (OOD) request (Tur et al., 2014). Finally, in a multilingual environment the SDS may switch between different languages, e.g., a robot that a human can query in English or Japanese to initiate a Wikipedia search (Wilcock and Jokinen, 2015).

The ASR cannot recognize what it does not know about, and in the cited examples the researchers solved this problem by extending the ASR vocabulary or by adjusting the prior probability of the hypothesized word sequences. However, detecting and interpreting the user's true intention, and selecting an appropriate response given noisy human speech and ASR transcription errors, requires a method for communicating between SDS components. For example, the NLU, Knowledge Base, and Dialogue Manager components can request the recognition subsystem to reinterpret the utterance in the event of say, a processing failure.

In this paper, we show a novel framework for understanding spoken dialogue in which utterance analysis is escalated through a multi-level system *involving interpretation on syntactic, semantic, and contextual/topic levels* (see Figure 2). Analysis is applied incrementally at each level as the system attempts to resolve the uncertainty surrounding utterance interpretation. Links to other SDS components from each of the levels can affect the agent's beliefs and, conversely, other components can signal the framework to reinterpret the utterance. This may occur in the context of, for example, a new topic. To our knowledge, no other approach has demonstrated the use of such a multi-tiered system for improving accuracy in the SDS' ability to recognize spoken task commands. This paper proceeds as follows. In Section 2, we discuss prior approaches to resolving out-of-domain requests, using context to improve ASR and parser performance, and learning novel words. Section 3, situates these approaches in our framework. We also discuss how *topic detection* is used to determine context. Section 4, discusses the framework demonstration and evaluation, using utterances drawn from two task domains in two conditions: one with the framework and one without. Finally, Section 5, examines the advantages, disadvantages, and limitations of this approach and framework improvements.

2 Related Work

Research in improving NLU in task-oriented dialogue systems and intelligent agents can be motivated as follows. One way to ensure reliable performance of speech recognition for SDSs is to make a *closed-world* design assumption, and limit their operation to well-defined domains (for examples, see Lane et al. (2005)). This could be accomplished by representing the dialogue model as a finite state system using a pre-defined state transition network, which assumes that the dialogue is known in advance (McTear, 1998). This approach is not resilient to input outside the agent's domain, and so frame-based dialogue systems have been proposed. In these, the model attempts to fit the dialogue into frame slots (i.e., a "form") corresponding to an action or utterance (Xu and Rudnicky, 2000). However, these systems struggle when utterances fail to fit into a frame. Finally, agent-based systems have been developed for more advanced dialogue management with interactive agents (Allen and Perrault, 1980; Cohen and Perrault, 1979). These approaches typically categorize utterances into *speech acts* (Searle and Searle, 1969) and perform inference or planning based on a mental model of the interlocutor. Such systems make use of robust symbolic representations of an agent's beliefs, desires, intentions, and other properties, allowing for mixed-initiative interaction (Rao and Georgeff, 1991). This is the dialogue management approach used in the present system.

It is desirable that the human be able to communicate in a natural and flexible manner with the agent. To enhance usability, NLU systems are built on *open-world* assumptions. In these systems, the user may provide both in-domain and OOD inputs, the latter of which may be unsupported by the system. Accepting OOD inputs could lead to errors propagating through the system, which may lead to undesirable responses unless it can reliably distinguish between the two and process them accordingly. Context detection is one approach researchers have used for OOD. Veale et al. (2013) discuss a method for applying top-down contextual bias based on the expected dialogue turn to a neural speech recognition system to improve its performance. Sarma and Palmer (2004) compute the likely contexts of all words in an ASR system vocabulary by performing a lexical co-occurrence analysis using a large corpus of output from the speech system. This is used to find the likely context for query words, and the system uses this to identify similarly-sounding, but erroneous query words.

Topic detection may also be used to infer context. Lane et al. (2006) proposed a detection framework which makes use of the classification confidence scores of multiple topics and applies a linear discriminant model to perform in-domain verification. Lane et al. (2005) describe an architecture which combines topic detection with topic-dependent language models for use in a multi-domain SDS. According to the researchers, their approach allows the user to freely switch among domains while maintaining a high-level of accuracy.

However, topic approaches use a *bag-of-words* which, along with those that are feature-based (Tur et al., 2014), have difficulty dealing with unknown words, e.g., rarely used expressions and neologisms. To overcome this problem, Oh et al. (2018) describe a method in which OOD sentences occurring in a dialogue are detected based on sentence

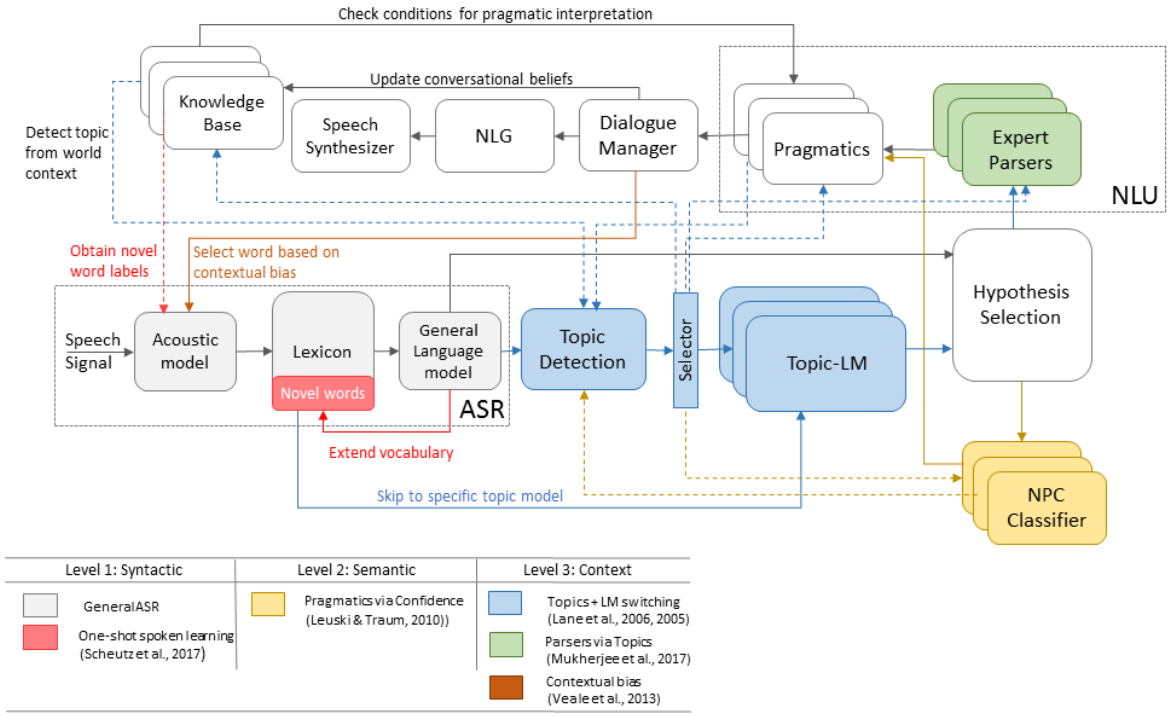


Figure 2: Multi-level framework for understanding spoken dialogue encompassing prior approaches.

distances. The distances are measured by sentence embedding vectors using RNN (Recurrent Neural Network) encoders and incorporate an attention mechanism.

Alternatively, Scheutz et al. (2017) describe a mechanism for detecting the intentional use of novel words in a *one-shot learning* system. Here, the ASR is modified such that when an unknown token is generated by the acoustic model, its corresponding word-level unit is discovered from the acoustic features. A nearest-neighbor classifier is used to determine whether the discovered unit represents the first member of a new word-class of the vocabulary and, if so, the class and example are added; otherwise it is added to an existing class.

In addition to using context to switch among language models, topic modeling can be applied to syntactic SDS components. Mukherjee et al. (2017) use Latent Dirichlet Allocation (LDA) to improve parser performance across multiple domains. LDA is used to find the topic structure in a document, which is a single sentence here. The sentence is assigned to the most likely topic and an "expert" parser for the topic is trained for syntactic analysis.

For situations where the domain is constrained, yet the user will be using natural language with its attendant disfluencies and irregularities, the ASR is likely to not recognize domain-specific commands. For this type of system, Leuski and Traum (2010) describe a statistical classification component which, in order to automate natural and flexible human-agent dialogue, identifies the best response to user input by estimating the closest possible match within its training set if a precise match cannot be found.

Finally, Chen et al. (2013) describe how information from multiple non-ASR components in their conversational spoken language translation system can be combined with strong baseline ASR error detector features and used to improve overall ASR error rate. The system contains built-in error detection modules that pinpoint regions in the input where the ASR is likely to fail, including a confidence estimator of the language translation (i.e., English-Iraq). In addition, the posterior word probabilities from a named entity detector is used to improve out-of-vocabulary word recognition.

In this section, we reviewed literature representing the main approaches to resolving OOD inputs and improving ASR performance. The contextual approaches (e.g., topic modeling, word co-occurrence, statistical classification) have the effect of changing the prior probabilities of the trained ASR by making a selection from multiple language models (Mukherjee et al., 2017; Lane et al., 2006, 2005; Sarma and Palmer, 2004), biasing the ASR

word hypothesis (Veale et al., 2013), or discriminating among similar interpretations (Leuski and Traum, 2010). An alternative approach is to extend the ASR vocabulary when a novel instance of a word class is detected as in Scheutz et al. (2017). Chen et al. (2013) used a combination of the two approaches. In the following section, we will discuss how these approaches have been integrated with some, but not all, of the components of the SDS framework.

3 Multi-level Framework

Our framework consists of three possible levels of utterance interpretation: syntactic, semantic, and context. The purpose of the first level is to generate an interpretation of the user's meaning using a *syntactic* analysis of the utterance. For example, an utterance such as "pick up the red box" can be easily parsed along strict syntactic rules, generating a semantic form for the utterance by means of these syntactic rules. However, a more complicated, non-grammatical utterance such as "the red box, pick that up maybe" may violate expected syntactic rule structure. Thus, we attempt to handle unexpected utterances with a second level of interpretation. This level bypasses any syntactic assessment entirely; it attempts to guess directly at the *semantics* of the user's intent by finding which group of semantically-linked utterances is most similar in word content to the target utterance. Here, a classifier is used to select the most likely interpretation based on the utterance's similarity to previous utterances which the classifier has been trained on. For example, "the red box, pick that up maybe" is assessed as most similar to "pick up the red box", and is assumed to have the same meaning (and therefore the same semantic form). Both of these levels return a confidence score to allow the Hypothesis component to select between the semantic form produced by the classifier and the semantic form produced by syntactic analysis (in this case, the semantic form may be in predicate logic, i.e., *pickUp(self,box(red))*). At the third level, the *context* of the utterance is analyzed and used to restart and inform connected components to reinterpret the utterance using, for example, a new language model, classifier, or parser.

In the legend of Figure 2, we situate selected prior work in the framework, assigning them to the syntactic, semantic, or context levels in accordance with their approach to improving SDS performance. We place Scheutz et al. (2017) "One-shot Spoken Learning" in the syntactic level as they assume an unrecognized speech token may be a novel word. After a pattern analysis of the acoustic features, their system attempts to place the new token in the vocabulary and interprets its grammar function within the syntactic context of the rest of the sentence. This flow is shown by the solid red connections in the figure. However, to recover the label of the word so that the agent's text-to-speech system can say it back, the phonemic sub-units within the word feature must be recovered and mapped to the pronunciation dictionary. The dotted red line indicates the required connection for this capability.

We situate the NLU system described by Leuski and Traum (2010) in the second, semantic level as it makes no assumption that the syntactic form is correct. Level 2 uses a statistical classifier to generate multiple similar interpretations of the utterance, selecting the one with the highest confidence score and sending it to the Pragmatics component for intention analysis; these connections are shown by the solid gold lines in the figure. The classifier component could ask for a back-off and reinterpretation of the utterance if the highest confidence score falls below a specific threshold; the connections for this additional capability are indicated by the dashed gold lines.

One particular benefit of this multi-level framework is that it allows for the combination of a rule-based approach to NLU (the parser) with a machine-learning approach to NLU (the classifier). Alone, a rule-based approach hardly ever reaches the broad accuracy of state-of-the-art systems that use machine learning. However, with lack of explainability being a key weakness of machine learning, the machine-learning approach by itself rarely offers any clear introspective explanation of how or why it arrived at its results, unlike a rule-based approach which can provide the rules and syntactic definitions behind the reasoning of any particular parse. With a trained classifier able to outperform the strict limitations of the rule-based parser by finding the closest possible expected sentence to the user input, and a parser able to provide syntactic explanation for any expected sentence, this multi-level architecture can surpass the weaknesses of both approaches.

The basis for the context level 3 is its ability to use topic detection to infer utterance context and thus we situate the hierarchical topic classification

of Lane et al. (2005) in that level. In their implementation, the researchers describe a system which can detect in- and out-of-domain utterances, and freely switch among several topic-dependent language models. In the figure, the connections and components for this system are shown in blue; however, we indicate by the blue dashed line that there could be additional connections that could further improve the interpretation. New connections from the Knowledge Base, Pragmatics, and Classifier components allow them to request a back-off and reinterpretation of the utterance by inferring its context through topic detection. Connections from the selector back to those components can signal that an alternative, topic-dependent classifier, pragmatics, KB model should be used.

We also situate in level 3, the system described in Mukherjee et al. (2017) which creates topic-specific datasets that are then used to train expert parsers. This system is shown in the green box in the figure without a solid line connection to the Selector because the researchers have evaluated the expert parsers individually and do not specify a method for freely selecting from among language models. The green dashed lines show the connections to Topic Detection and from the Selector to indicate this added capability.

Finally, we situate the dialogue contextual bias signal system described in Veale et al. (2013) in level 3. Rather than using a topic model to infer bias, the authors use the knowledge of common dialogue exchange patterns contained in the dialogue Manager to develop a bias signal for the ASR component (shown as a solid brown line in the figure). This is used to change the words' prior probabilities in the ASR, influencing word selection according to dialogue context. The authors describe this system as a biologically plausible cognitive model based on human perceptual decision making. As such, it provides an interesting avenue for further research into human-like ways to improve speech recognition.

3.1 Implementation

We implemented the multi-level framework shown in Figure 2 in the DIARC cognitive robotic architecture (Scheutz et al., 2019). The implementation consists of the ASR, Topic Detection, Selector, Topic-Dependent Language Models, Parsers, Classifier components, and Pragmatics. The ASR is based on the chain model developed using Kaldi for the ASpIRE Challenge and trained on Fisher English (Harper, 2015; Povey et al., 2011). For parsing, we used a symbolic, rule-based parser, and for the classification component, an implementation which is part of the NPCEditor platform (Leuski and Traum, 2010).

Figure 3 shows that the processing flow begins with a spoken utterance that is transcribed imperfectly by the ASR component. In a single-level system, the utterance follows the blue line to the syntactic parser, which attempts to parse the meaning into a semantic form and send it along the blue line to the pragmatics component for extraction and further processing. In the multi-level system, the utterance is sent both to the parser and classifier (brown lines), which then each send the pragmatics component their own interpretation of the utterance's meaning along with a confidence score. For the classifier, this confidence score rates the utterance's similarity to the closest utterance in the classifier's training set (though if this score is below a predefined confidence threshold (0.6) it rounds the score to 0). For the parser, the confidence score is either 1.0 to signify that it was able to find a valid parse, or 0, in cases where no parse was able to be found.

The pragmatics component determines the highest score between the parser and classifier and picks the respective interpretation for further processing by the rest of the dialogue system. If the parser fails and the classifier cannot find a semantic interpretation with $score > 0.6$, then the brown line on the "No" branch is followed to the Topic Detector. If the topic has changed, the Selector will switch to a generic language model that is a mixture of topic unigrams, and the utterance will be reinterpreted. If there was no topic change, the framework assumes there cannot be a valid interpretation, and will generate an appropriate response to the user. The generic LM is used so that the Topic Detector has a basic utterance to which it infers a topic distribution. The distribution is used to select a domain-specific LM, and the utterance is reinterpreted using the new LM.

3.2 Topic Detection and Model Training

Prior to evaluating the multi-level framework, we trained the Topic Detector and tested its ability to distinguish utterances drawn from one domain or the other as follows (see Section 4 for a discussion on the domains used). We used Latent Dirichlet

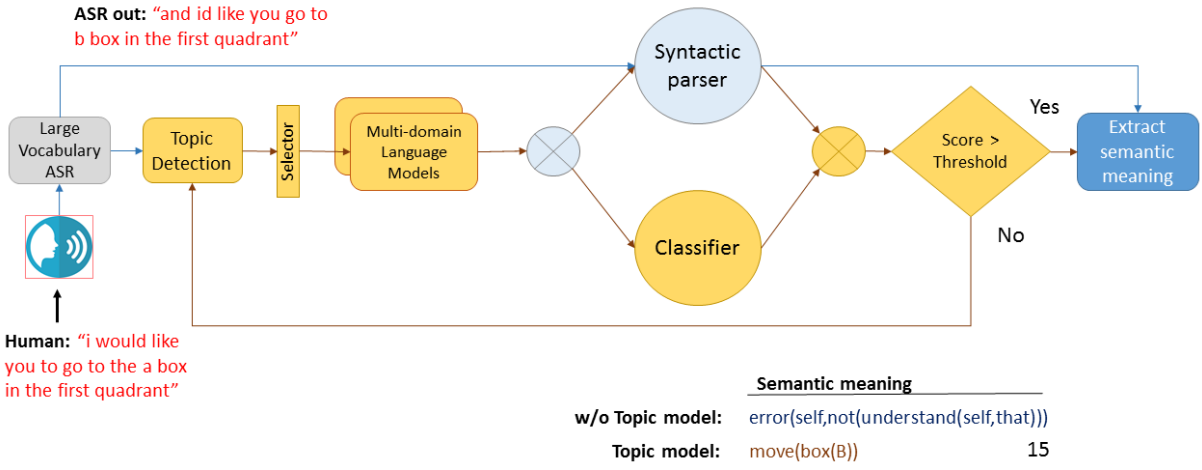


Figure 3: Multi-level processing flow. A single-level system assumes the utterance transcription received from the ASR is entirely accurate or has a single interpretation (blue lines). The multi-level system acknowledges ASR output may not be in exact expected form, and uses a classifier to find a close match. If the classifier fails, the topic is inferred and the language model and other components are switched to a new context (brown lines).

Allocation (Blei et al., 2003) to infer topics in a heterogeneous collection of textual data. We used the Gensim (Řehůřek and Sojka, 2010) implementation of LDA to train the Topic Detector and extract the topic distribution from the utterances. We used the default hyper-parameters and set the number of topics equal to the number of domains used in the training and test datasets, i.e., two.

We used the trained LDA model to create the Topic Detector. It detects a topic shift from one utterance to another by comparing the Kullback–Leibler (KL) divergence between the two topic distributions (Kullback and Leibler, 1951). If the difference is above a pre-determined threshold, a shift is indicated. This is used by the topic detection component to signal the Selector to switch to a new language model. As stated previously, this mechanism can be extended to select among alternative parsers, classifiers, knowledge base components, etc.

We used k-fold validation with $k = 5$ to train and test the Topic Detector's ability to distinguish utterances drawn from one domain or the other. Each fold contains utterances from both domains. To prepare the folds, we selected utterances from both domains, and placed them in 5 equally sized sets. We then trained and evaluated the topic detection model 5 times, selecting a different fold for evaluation every time and training on the other 4 folds. The training subsets consist of utterances from both domains.

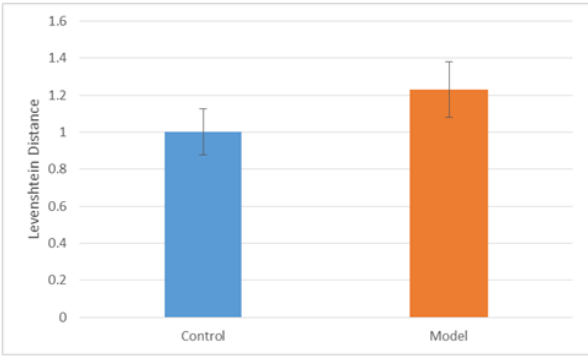
The testing subset consists of at most 100 randomly paired utterances from the domains. Since the number of pairs increases exponentially with the size of the original list, this limit is imposed on the maximum number of sentence pairs to select. If a pair of sentences come from different files, then this is considered ground truth of a topic change, and vice versa. Sentence pairs are fed to the topic model, one after the other, and the predicted topic change is compared against the ground truth topic change. If the KL divergence of the two sentences is equal to or exceeds the input KL threshold (0.5), then this is considered a topic change, and otherwise not.

4 Framework Demonstration

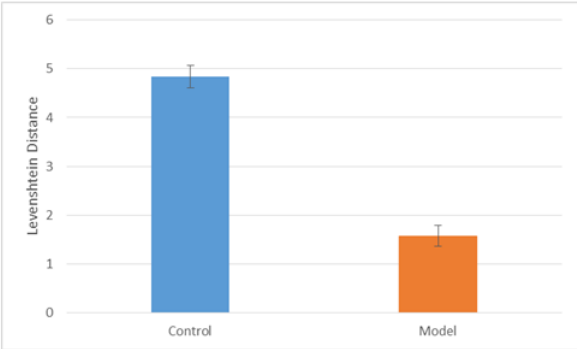
For model validation, we evaluated its accuracy in interpreting natural language utterances from different human-robot tasks. The goal was to compare our multi-level system to one which only used syntactic parsing. Using the data collected in other human-robot interaction experiments, we obtained two separate corpora of utterances used to instruct a robot in a specific task environment. For this implementation, we trained the topic detector on two topics. In the first domain, *SpaceStation*, participants ($n = 26$) gave commands in natural language to control several robots repairing components of a space station (Gervits et al., 2020). Out of 663 utterances, including duplicates, 363 unique utterances remained, out of which 50 were withheld for the test dataset and 313 were used to train the LDA model. In the second domain, *Diorama*, participants ($n = 33$) taught new skills to a robot learner

using natural language (Bennett et al., 2017). Out of 680 utterances, including duplicates, 525 unique utterances remained, out of which 50 were withheld for the test dataset and 475 were used to train the LDA model. The sentences from the SpaceStation domain comprised one document and those from the Diorama domain comprised another. The LDA model was trained on a union of the two document collections. We set up a pipeline for incoming utterances wherein each utterance would be processed in parallel by two different systems: our three-tier framework which included topic-switching and a bag-of-words classifier, and a control framework which possessed only the baseline tier of interpretation through syntactic parsing.

In both the control and multi-tier systems, parser rules were written by hand based on 100 utterances from each corpus, which were also used to train the ASR language models. For the control framework, these syntactic parsing rules were combined into one parser dictionary, whereas for the multi-tier system, the parser swaps between topic-specific dictionaries at the topic detector's signal. Similarly, the control system's language model was trained on the combined set of 200 utterances, while the multi-level system contained two separate models each trained on 100 utterances from the distinct tasks. For the multi-level system, two different classifiers were trained on the two different sets of utterance training data that had been hand-labeled with the correct semantic interpretation in predicate 1st-order logic form for each utterance. This pipeline was fed a test set of 50 utterances from each corpus (100 utterances total) that were withheld from the training data. Relevant output such as utterance transcription, semantic form, and in the multi-level framework's case, topic identification, were logged and manually annotated by the experimenter with the ground truth values of these variables: the correct transcription of each utterance, its intended interpretation in symbolic 1st-order logic form, and the task (topic) from which it originated.



(a)



(b)

Figure 4: (a) Mean token-based Levenshtein distance (LD) between each utterance and its ASR transcription. (b) LD between the ground truth meaning of an utterance in predicate form and the model interpretation.

4.1 Results

To investigate the differences in accuracy between the transcriptive and interpretive abilities of the multi-level and control frameworks, we compared the output of each framework per utterance to its respective ground truth value using a Levenshtein distance (LD) metric, which measures word dele-

tions, insertions, or substitutions. For transcription accuracy, we found the token-based LD between each utterance and the ASR transcription of that utterance[1]. In the multi-level framework, the $mean\ LD = 1.23$, and for the control, the $mean\ LD = 1.00$. We verified that the variances were homogeneous and then conducted a paired two-tailed t-test which showed no significant difference in transcription accuracy between frameworks ($t(99) = -1.707, p > .05$). Similarly, for interpretation accuracy, we found the LD between the ground truth meaning of an utterance in predicate form and the system's semantic interpretation of that utterance to measure our dependent variable: how close the predicate of the utterance interpretation was to the ground truth. For example, the ground truth utterance "pick up the red box" has a predicate form (semantic meaning) of *pickUp(self,box(red))*. In cases where the system was unable to come up with any interpretation, the LD defaulted to 6. In the multi-level framework, $mean\ LD = 1.57$ and for the control, $mean\ LD = 4.84$. We verified that the variances were homogeneous and then conducted a paired two-tailed t-test found a significant difference in these means ($t(99) = 12.425, p < .001$), demonstrating that the multi-level framework is significantly much more accurate than the control.

The control framework was generally only able to accurately identify the semantic form of the utterance in cases where the utterance matched exactly to the grammatical rules specified in the parser. In addition, it occasionally generated the correct semantic form in cases where it misrecognized an unexpected utterance as an expected utterance with the same meaning. For example, the utterance "hey robot one come fix this tube" was misrecognized as "fix the tube" by the control framework and thus correctly parsed with the meaning of *repairTube()*. In other cases of transcription inaccuracy, the control framework occasionally recognized portions of utterances from the wrong task corpus. For example, it recognized the utterance "robot one go to left four" as "robot one go left", thus interpreting the utterance as *move(left)*. While the diorama task has a need for this level of directability in the robot's movement, the space station does not, and the action *move(left)* is used exclusively in the diorama task. However, the control framework had no reason to suppose that this command was less likely to be uttered in this context, either at the speech recognition level or at the parsing level. In all other cases, the control framework was unable to parse the unexpected utterance, even if the ASR Component transcribed it completely accurately. In contrast, the multi-level framework, even if unable to get a wholly accurate transcription, was generally able to come up with a semantic form which, if not exact, was fairly close to the intended interpretation, on average only off by one or two arguments. The topic identifier in the multi-level framework also correctly identified the topic of the utterance 93% of the time.

[1]This was token-based instead of character-based because we did not wish to reward misrecognition of a shorter word over that of a longer word, i.e., "canned" and "can" vs "canned" and "tanned"

5 Discussion

We hypothesized that the multi-level framework would perform better in interpreting semantic meaning of utterances. The results show that there was a statistically significant difference between the semantic interpretation accuracy of the multi-level framework and the control framework, with semantic interpretation being more accurate in the multi-level framework (Figure 4b). In addition, there was no significant difference between the transcription accuracy of the multi-level framework and the accuracy of the control framework (Figure 4a). The control framework performed very well on expected utterances (speech that matched the syntactic structure of utterances from the training set), but very poorly on unexpected utterances. Its success at semantic interpretation was binary: either 100% or 0% certainty. In contrast, the multi-level framework was able to infer in uncertain situations due to the classifier, leading to greater overall interpretation success. Although executing a wrong command could be potentially worse than not understanding a command at all, having an uncertain estimate of what the user wants is better than no estimate. Rather than executing the command, further error recovery could begin based on information from other components. For example, if the agent's certainty regarding its interpretation is not above a certain threshold, its dialogue manager could initiate a confirmation or clarification request, or its knowledge database could be solicited for contextual information or dialogue history that might resolve uncertainty. This would be a direction to explore in future work.

We note that even in the control framework, the

LM was trained on a selective portion of data only containing the two topics. In contrast, the Aspire Chain Model default LM is used for general dialogue. Thus, there is not a substantial difference in the ASR word error recognition (WER) between the two. When run with the default, the control's WER is far worse. This will vary depending on how specific to the task the utterances are. For example, utterances like "go to left four" or "drive forward pushing box c" are transcribed as the irrelevant phrases "gonna last for" and "dry forward pushing box see", while "what are you doing right now" and "knock down the yellow tower" are recognized correctly. The success of the multi-level framework depends on whether or not the topic is identified correctly. If the topic is misidentified, all other output from the system will also be incorrect. Though a 93% success rate of topic identification appears good, the success rate might decrease as more topics are introduced. For this reason, focus should be placed on how other components of an autonomous dialogue system can be integrated into the process of topic identification, so that the burden is not solely placed on ASR. There are several examples of how this can be handled. For example, if the parser and classifier both fail to come up with an interpretation above some threshold of certainty, they could prompt the topic component to switch the topic to the second-place choice and attempt another pass at NLU, or additional information could be solicited from the system's knowledge base about the dialogue history, previous goals, goal status, or world state that may further assist with topic identification. The Dialogue Manager can also initiate clarification requests in case the above methods fail.

We used only classifier and parser confidence scores to evaluate interpretation, but other metrics might be considered. For example, the ASR's word-level transcription confidence could be used ahead of the interpretation components to signal an earlier switch to another language model. Chen et al. (2013) supplemented ASR confidence scores with additional metrics such as: LM perplexity, number of competing words, and acoustic model deviation from true scores. They also used acoustic-prosodic features for improved confidence in the ASR-hypothesized word boundary detection. This might have the greatest benefit for the framework when given OOD utterances. These utterances are often broken up into multiple in-domain words and thus, word insertions are frequent, making up about 40% of word errors.

6 Conclusion

Natural language interaction with SDSs can result in errors which propagate through the components, causing interpretation of utterance meaning to fail. We developed a multi-level framework in which utterance analysis is escalated according to feedback received at the syntactic, semantic, and topic level. We situated this framework in the context of prior research in improving speech recognition and natural language understanding and showed how they have been integrated with some, but not all, of the components of our framework. In a demonstration in which humans used natural language to initiate commands controlling robots in two separate domains, we showed how these approaches can be integrated with other SDS components. We found improved accuracy in the SDS' ability to interpret spoken task commands. By integrating multiple different potential routes for understanding into the dialogue system, we allow for better recovery across the system.

References

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

Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6589–6594. IEEE.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Wei Chen, Sankaranarayanan Ananthakrishnan, Rohit Kumar, Rohit Prasad, and Prem Natarajan. 2013. ASR error detection in a conversational spoken language translation system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7418–7422. IEEE.

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

Renato De Mori. 1997. *Spoken dialogues with computers*. Academic Press, Inc.

Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz.

2020. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *AAMAS '20: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.

M. Harper. 2015. The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 547–554.

Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1):1–151.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Ian Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2006. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):150–161.

Ian R Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2005. Dialogue speech recognition by combining hierarchical topic classification and language model switching. *IEICE transactions on information and systems*, 88(3):446–454.

Anton Leuski and David Traum. 2010. Practical language processing for virtual humans. In *Twenty-Second IAAI Conference*.

Michael F McTear. 1998. Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit. In *Fifth International Conference on Spoken Language Processing*.

Atreyee Mukherjee, Sandra Kübler, and Matthias Scheutz. 2017. Creating POS tagging and dependency parsing experts via topic modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 347–355.

Kyo-Joong Oh, DongKun Lee, Chanyong Park, Young-Seob Jeong, Sawook Hong, Sungtae Kwon, and Ho-Jin Choi. 2018. Out-of-domain detection method based on sentence distance for dialogue systems. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 673–676. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Anand S Rao and Michael P Georgeff. 1991. Modeling rational agents within a bdi-architecture. *KR*, 91:473–484.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Arup Sarma and David D Palmer. 2004. Context-based speech recognition error detection and correction. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 85–88. Association for Computational Linguistics.

Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. 2017. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*.

Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*, pages 165–193. Springer.

John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Gokhan Tur, Anoop Deoras, and Dilek Hakkani-Tür. 2014. Detecting out-of-domain utterances addressed to a virtual personal assistant. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Richard Veale, Gordon Briggs, and Matthias Scheutz. 2013. Linking cognitive tokens to biological signals: Dialogue context improves neural speech recognizer performance. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Graham Wilcock and Kristiina Jokinen. 2015. Multilingual wikitalk: Wikipedia-based talking robots that switch languages. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–164.

Wei Xu and Alexander Rudnicky. 2000. Task-based dialog management using an agenda. In *ANLP-NAACL 2000 Workshop: Conversational Systems*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Victor W Zue and James R Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180.

How to Say Things with Hand Gestures

Sandy Ciroux
Universität Konstanz
sandy.ciroux@uni-konstanz.de

Abstract

Speech act theory has been around and undergoing revisions for more than 50 years. Theories of gestures, on the other hand, are newer, as they started to emerge about 30 years ago and have since then gained more and more attention. This paper wishes to align to this growing interest in gestures by proposing an experimental pilot study of illocutionary acts produced with the help of so-called co-speech hand gestures. This study in fact brings together the two subfields of linguistics aforementioned. The experiment consists in two face-to-face conversations, each between two people. Using ELAN and the Linguistic Annotation System for Gestures (see Bressem et al., 2013) for my analyses, I found that (1) contrary to iconic and metaphoric gestures, which are unable to produce illocutionary acts by themselves, deictic gestures have some potential for being used without any speech in the production of a well-formed utterance, (2) the words accompanying a deictic gesture are often redundant with it (and not the other way around), (3) only deictic gestures are able to realise non-assertives and (4) it seems possible for deictic gestures to both contribute to the propositional content and realise the illocutionary force.

1 Introduction

Speech act theory was first developed by Austin in *How to Do Things with Words* in 1962 and then elaborated by Searle, who also coined its name. This theory explains how an addresser can produce acts by means of uttering words.

Whenever we make an utterance, we produce a series of acts: an utterance act (uttering of words), a propositional act (referring and predicating) and an illocutionary act (Searle, 1969). Together, the utterance act and the propositional act form the locutionary act of an utterance, which in turn forms its illocutionary act together with the illocutionary force (intention in issuing the utterance) (ibid.).

The literature usually recognises five main categories of illocutionary speech acts (each containing many subcategories): assertives, commissives, declaratives, directives and expressives. No two illocutionary acts have the same force. For instance, an example of directive acts that do not share the same illocutionary force is the act of requesting versus the act of ordering. This contrastive example is further discussed below.

Another act, parallel to the illocutionary act, that is discussed in the literature is the perlocutionary act, that is the act of producing a particular effect in the addressee.

The next section sketches out a brief review of what speech act theorists and other scholars interested in communication said about multimodality.

2 Background

On top of producing acts by means of uttering words, addressers can also produce illocutionary acts non-verbally, such as betting, warning and ordering, to name but a very few (Austin, 1962). Additionally, Searle (1969) highlights that, though linguistic communication is not restricted to words, any element needs to be intended to be taken as linguistic.

A certain communicative element (e.g. a gesture or tone of voice) can also affect the illocutionary force of an utterance. Consider the following example discussed by Bierwisch (1980):

(1) John, give a dollar to the bar keeper.

This utterance can be either a polite request or a

strict order depending on the accompanying gesture(s) and/or tone of voice (Bierwisch, 1980). I believe one can indeed suspect that if, for instance, the utterer uses a palm up open hand gesture to point at the bar keeper and moves her arm with a reduced and decelerated movement, it will most probably be understood as a request. However, if she uses her fully stretched index finger with an enlarged and accelerated movement, the utterance will presumably rather be interpreted as an order. A similar example will be discussed in the analysis section.

Addressers can also use deictic gestures towards objects in order to make their addressee turn and look at them. For instance, if someone says 'that chair' with a pointing gesture, it will most likely make the addressee look at the chair. In fact, gestures can also be used to locate objects, events, places and the like (Clark, 1996). Moreover, as Clark (1996) puts it, '[s]peakers gesture with their voice, hands, arms, face, eyes, and body in order to describe, indicate, and demonstrate an indefinitely wide range of objects, events, states, and properties' (p. 256).

Wharton (2009) further underlines that both facial expressions (aka facial gestures, i.e. smiles, eyebrow flashes, frowns and so on) and hand gestures indicate the addresser's internal state by conveying their attitudes to the proposition expressed:

(2) Lily (with a stern facial expression, in an angry tone of voice, gesturing furiously): You're late!

In this example, the aggressiveness of Lily's gestures shows how angry she is.

Additionally, Cuffari (2012) aptly pinpoints that some gestures accompanying speech have the same functions that words can fulfil. Because speech and gesture contribute in very different ways to fulfilling their functions, 'the confluence of both phenomena makes a distinct and significant contribution to the whole speech act' (Cuffari, 2012, p. 618).

Similarly, Sperber and Wilson (1995, 2012b) point out that utterances sometimes need a higher-level explicature that can be achieved by a mixture of describing and indicating, i.e. by words and by paralinguistic cues (e.g. tone of voice, gaze, head movements, facial expressions, hand gestures,...). Take this as an exemplification:

(3) I have been here for two hours.

'Here' indicates some restriction on the area the addresser is referring to. However, when we use a deictic word such as 'here', it is not always clear if we are talking about the specific room we are in or rather about the whole building, for instance. This can be made more salient by the use of hand gestures or gaze (Sperber & Wilson, 2012b).

Some other times, one can also communicate without using words at all. To illustrate this point, Sperber and Wilson (2012a) give the following example: Mary is angry with Peter and does not want to talk, but when Peter tries to communicate, she can (1) look angrily at him and clamp her mouth firmly shut or (2) look angrily at him, put a finger to her lips and whisper 'Shhh'. In these two alternatives (among many others), Mary uses gestures that conventionally express that one wishes one's addressee to be silent (or discreet).

On top of that, Sperber and Wilson (2015) state that facial expressions can trigger conversational implicatures and thus have an impact on the message the addressee will understand:

(4) Passenger: What time is the next train to Oxford?
Railway official: 12.48.

With a neutral facial expression, the answerer in this example does not implicate anything. However, with a specific facial expression, she may implicate that the train is leaving soon or that the platform is far away. A suitable facial expression would give the addressee the opportunity to act accordingly to try and catch his train.

On another level, the works discussed below tackle the issue of the nature of the function relating a gesture accompanying a propositional content and the speech act itself. The main concern is whether gestures express a propositional content (by adding some content-related information) or realise the illocutionary force of the communicative act (e.g., as mentioned above, by transforming what could be seen as a request into an order or vice versa).

Fein and Kasher (1996) see three possible functions: (1) gestures vary with both the illocutionary force and the propositional content, (2) gestures vary only with the illocutionary

force or (3) gestures vary just with the propositional content. Fein and Kasher's (1996) study (on comics gestures) suggests that gestures vary mainly with the illocutionary force of an utterance.

For his part, Kendon (2000) notes that, not only do addressers use gestures to provide additional content or precision regarding pragmatic aspects of the utterance, but they can also change their gestures in order to recast an utterance as an other speech act. Gestures indeed do take part in the production of communicative acts in a variety of ways.

Now, Bressem et al. (2013) highlight that gestures with a representational function (e.g. iconic or metaphoric gestures) or a referential function (e.g. deictic gestures) can express propositional content, while gestures with a performative function (i.e. gestures that indicate the type of communicative act the addresser is engaged in, e.g. deictic gestures) realise the illocutionary force. All in all, they recognise three types of gestures in terms of their relation to communicative acts: (1) gestures expressing a propositional content, (2) gestures relating to the illocutionary force and (3) gestures affecting the perlocutionary force. Note that the study presented in this paper is not concerned with the latter.

3 Research Questions

To shed some more light on the role of gestures in the production of communicative acts, the following sections report on a pilot experiment into both the place and the contribution of hand gestures, namely it (1) addresses the question to what extent verbal and non-verbal components are needed in the making of illocutionary acts and (2) investigates whether and which hand gestures do contribute to the illocutionary force or to the propositional content. In other words, the research questions can be formulated as follows: Can iconic, deictic or metaphoric gestures be used without speech to produce an illocutionary act by themselves? Are deictic, iconic or metaphoric gestures redundant with the word(s) they accompany or is this the other way around, i.e. are words redundant with the gesture they accompany? In the making of what kind of illocutionary acts are these gestures used? Do these gestures contribute to the propositional content or rather to the illocutionary force?

4 Method

The pilot experiment was conducted in French (because it was easier to find participants since I was at a French-speaking university at the time of the experiment) and examined the use of deictic, iconic and metaphoric gestures produced during dyadic conversations, as those are the co-speech hand gestures that are communicative and thus apt for the production of illocutionary acts.

Basing myself on the literature, I have discriminated between deictic, iconic and metaphoric gestures as follows. Deictic gestures are pointing gestures that locate and spatialise people or objects. Both iconic and metaphoric gestures depict some aspects of something by virtue of their resemblance to it. The difference between them resides in that iconic gestures are directly formally related to the semantic meaning of a linguistic unit, so much so that they can depict features of what is described, that are or are not expressed in words, while metaphoric gestures are indirectly and abstractly related to the meaning of the linguistic unit they refer to and of which they depict the concept, that has no physical form.

There exists other types of hand gestures such as beats or emblems, neither of which has been considered for analysis because (1) beats are not seen as communicative and (2) emblems are not considered co-speech gestures, as they are often used alone as complete utterances (see McNeill & Levy, 1982; McNeill, 1985; Kendon, 1986; Cassell et al., 1999; Haviland, 2000; and Enfield, 2012 among others for detailed descriptions of all the types of hand gestures mentioned above).

4.1 Participants

Four participants, grouped into two pairs, took part in the experiment. At the time of the experiment, they were studying towards a master's degree in linguistics at the Université Libre de Bruxelles. All four of them were told they were taking part in a study on the way people express themselves so that they were naïve to the exact purpose of the study. They all participated voluntarily.

4.2 Materials

Participants were presented with a list of thirteen topics, more specifically social issues, which were picked because they were assumed to be easy to be talked about. Participants were

instructed to select and discuss at least two topics from that list. The topics were as follows: (1) inclusive writing in French, (2) the bilingual policy in Brussels, (3) unequal pay according to sex, (4) feminisation of trades names, (5) immigration policy in Belgium, (6) free tuition fees in universities, (7) gymnastics classes in schools, (8) the teaching of religions in schools, (9) the neutralisation of genders in language, (10) the place of advertising in the public sphere, (11) wearing ostentatious religious symbols in the public sphere, (12) the impact of the internet/new means of communication on love relationships and (13) the impact of the internet/new means of communication on the pupil-teacher relationship.

4.3 Procedure

The assignments (i.e. instructions and list of topics) were given on a sheet of paper taped to the table in front of the participants so that they would not pick it up and would still have their hands free to gesture. A camera recording the interactions for later analysis was placed on the same table. The participants were kept standing in the hope that this would lead to production of more gestures. In order for them to be at ease with both their speech and gestural production, each pair had twenty minutes for conversation. Each participant knew their interlocutor so that they were more comfortable and hopefully produced more gestures than if they had to converse with someone with whom they were not at all acquainted. To make sure that the participants were really at ease, (1) they were offered the opportunity to get to know me, the experimenter, while enjoying a sandwich before the experiment and (2) they were left alone during the whole interaction.

4.4 Data Treatment

The first five minutes of the two video-taped conversations were left out of the analysis because they were used by the participants to familiarise themselves with the task and get involved in the debate. The remaining fifteen minutes of each conversation were annotated for analysis.

The annotations were made using the software ELAN and the Linguistic Annotation System for Gestures (Bressem et al., 2013). This system has been designed to be compatible with ELAN and, on top of considering gesture phases, recognises four parameters of hand gestures: hand shape, orientation, movement and position of the hand. Each of these parameters are then evaluated according to subcategories in terms of specific features (see Bressem et al., 2013 for a detailed description).

The analysis was carried out in three steps: (1) I annotated all the features (i.e. phases and parameters) of gestures leaving out the sound production at first, (2) then I concentrated on the gestures' relation with speech and (3) I analysed the communicative acts produced. As the sole experimenter involved in the analyses, I did the evaluation twice (with a seven-month interval) to ensure the correctness and reliability of the annotations.

5 Analysis

For each example in this section, the italicised part corresponds to the analysed gesture in its entirety, that is from the initial position to the rest position (see Kendon, 2004; and Ladewig & Bressem, 2013 for a detailed description of all gesture phases).

5.1 Deictic Gestures

All in all, there were ten occurrences of deictic hand gestures.

The first pair of participants was discussing unequal pay according to sex and women in politics to begin with. They had then gone somewhat off-topic, and so the addresser in (5) points at her addressee when she says 'tu' (*you*) for the second time.

(5) *Après tu m tu m'fais penser euh*, quand t'as dit un truc tout à l'heure […].
"*Then you m you make me think er*, when you said something earlier […]."

The deictic produced in this example seems redundant with the word it accompanies (i.e. both gesture and word seem to express the same meaning) but actually has two intertwined pragmatic functions: (1) it allows the addressee to better understand that he is the one who said something to which his interlocutor would now like to react and (2), because the addressee was starting to talk about something else, the addresser uses this gesture to indicate that she really wished to comment on her addressee's previous point. This gesture could actually draw the addressee's attention without speech (i.e. without the words used during the whole gesture

sequence), and is a contribution to the illocutionary force because, without the gesture, it would be a mere act of asserting some general state of affairs, but with the gesture, it can be interpreted as an act of informing someone in particular. This deictic can also be seen as a means of politely asking the addressee to stop talking for a moment so that the addresser can properly comment on the addressee's previous point before turning to an other idea. If the second interpretation is accepted, we can argue that the gesture in (5) transforms what would be an assertion without the gesture into a directive, namely a polite request. So in the former case, the illocutionary act is the same, that is an assertive, but the illocutionary force varies from asserting to informing. While in the latter scenario, the illocutionary act (and consequently, of course, the illocutionary force itself) changes from assertive into directive. Since this gesture has a performative function in both cases, it undeniably impacts the illocutionary force of the whole utterance.

The same participants were now talking about the neutralisation of genders in language. In (6), after an inaudible remark from one of them, the other participant points at her interlocutor with a palm up open hand gesture when she says 'bien' (*well*).

(6) *Oui, bien dit* !
"*Yes, well said*!"

Because such a gesture conventionally shows openness, it is redundant with the speech but helps the addressee to better understand that there is agreement on the addresser's side with what he said before. Because it refers to the addressee and what he said previously, this deictic gesture has a referential function and would actually do a good job without any words, as it alone already expresses the propositional content. Note that, if this gesture were to be used without any words at all — which, as just suggested, could be the case — it would additionally realise the illocutionary force and thereby also have a performative function. Either way, the communicative act is an expressive.

Before the inaudible remark previously mentioned, the participant had actually talked in the direction of the camera. In (7), the co-participant imitates her interlocutor by turning towards the camera. She makes a palm lateral open hand away from body gesture with both hands to point at it between 'T'as fait un' (*You were*) and 'face public' (*facing the public*) so as to indicate where the 'public' is.

(7) *T'as fait un … face public*.
"*You were … facing the public*."

This gesture gives information as to the spatialisation, which helps the addressee to better visualise what the addresser is referring to. The deictic in this example therefore definitely has a referential function. What is redundant in this case is the spoken part, as the pointing gesture appears before it and asserts exactly the same propositional content.

The participants were now discussing the wearing of ostentatious religious symbols in the public sphere. One of them was arguing that, religion being a private matter, no one ought to order someone to do something on the basis of one's beliefs. To illustrate her point, the addresser produces the quotation in (8), where she is imitating someone ordering a religious person to kneel by pointing towards the ground, which renders the description more vivid for the addressee.

(8) Et tu peux pas dire : « Ah *il faut que tu fasses euh que tu te mettes à genoux et que* tu applaudisses ».
"And you cannot say, 'Ah *you have to do er to kneel and to* applaud'."

As it alone already produces a directive, the gesture has a performative function and is consequently a contribution to the illocutionary force no matter what. Now, if there was only speech without the gesture, it could be a suggestion. On the other hand, if we only had the gesture production, it would undoubtedly lead to the interpretation of an order. But, as in this case speech is actually not central, it is clear that the gesture also expresses an order. The addresser points with a fully stretched index finger and the movement is enlarged and accelerated, which, as I would like to argue, are features linked to giving (strong) orders rather than making (kind) suggestions. Note that the gesture realises the illocutionary force of the quoted utterance only and not that of the whole utterance issued by the addresser, that is the illocutionary force of the utterance that includes the quoting frame is in no way affected by the gesture.

The example given below is similar to (8) in that the addresser places her addressee in the role of a person who is told what is good and what is bad by pointing at him as she says 'tu' (*you*).

(9) C'est comme montrer aux gens : « Tu vois, ça c'est bien *et c'que tu fais, c'est mal* ».
"It's like showing people, 'You see, this is good, *and what you do is bad*'."

As with (8), the gesture in (9) affects only the quoted utterance and not the utterance comprising the quoting frame. If speech was absent from the quoted utterance, the gesture alone could not express the whole proposition. But the gesture is also indispensable, as it emphasises the word it accompanies and in fact has a performative function. It contributes to the illocutionary force because, without the gesture, the quoted utterance would be an act of informing, but with it, it is an act of stating, which is a bit stronger. The gesture here really serves as a means to make the (impersonated) addressee realise that what *he* does is bad.

In (10), the addresser points at her forehead as she discusses the fact that making a young Muslim girl wear a scarf is equivalent to writing on her forehead that she will have to live as a woman all her life, that is under the conditions one stereotypically associates with Muslim women.

(10) Surtout, *tu lui écris sa condition d'femme sur le front quoi hein*.
"Mostly, *you write her her condition as a woman on the forehead huh*."

The gesture used here seems redundant with the words, but with it, the addressee's attention is drawn on the forehead, which emphasises the whole point. Alone, this gesture would not make sense of the whole proposition, but if speech were partially absent, the gesture would undoubtedly contribute to the propositional content. In any case, this deictic has a performative function and definitely contributes to the illocutionary force, or in this case to the strength, as it transforms an assertion into a claim.

The following utterances in this subsection were produced by the other pair of participants. They first decided to debate the impact of the internet/new means of communication on love relationships. More specifically, they were talking about Tinder. In (11), the addresser points at herself as she says 'moi je' (*I*).

(11) *Après euh moi j'ai une fille j'ai une une amie*, qui était très proche avant […].
"*But then er I have a girl I have a a friend*, who was close to me back then […]."

Here speech cannot be left out, but the gesture is useful because, with it, the addressee better understands that the addresser is currently referring to herself. The gesture thus has a referential function and reinforces the propositional content expressed in words. Now, whether with or without the gesture, it is an act of informing.

The participants are now talking about the teaching of religions in schools, and the addresser who utters (12) points at herself while she says 'moi' (*I*).

(12) *Moi c'est moi j'ai toujours été en école catholique*.
"*I it's I I've always been in a Catholic school*."

As with (11), the gesture in (12) has a referential function and is, just as the word it accompanies, part of the propositional content since it basically highlights who is speaking. In this case, the deictic gesture indeed does not have any performative function since the act of informing is not modified by the gesture. Therefore, it cannot be seen as a means to alter the illocutionary force. Also, this gesture could not be understood without speech.

The second group of participants picked up a third topic: the place of publicity in the public sphere. One of them underlined the crucial absence of publicity concerning autistic children outside the university walls. The participants had indeed previously seen such an advertisement on the next door. In (13), the addresser points in the direction of the door she is talking about.

(13) Y a pas de pub pour les enfants autistes euh *comme on a vu sur la porte à côté*.
"There is no publicity for autistic children er *as we saw on the next door*."

As the spatialisation of the door is not given in words, the gesture adds information to the propositional content. And actually, speech could partially be absent without altering the assertion,

as the place of the advertisement is known to both interlocutors. The utterance would indeed still make sense without 'sur la porte à côté' (*on the next door*) since the referential deictic gesture gives all the information that is needed. As with the latter two examples discussed above, the deictic gesture in (13) cannot realise the illocutionary force since it has a referential function only.

In (14), the addresser points in the exact same direction while talking about the corridor.

(14) J'trouve que, *quand t'es dans c'couloir par exemple* […].
"I think that, *when you're in this corridor for example* […]."

As with the spatialisation of the door in (13), the information about the spatialisation of the corridor is not expressed in words, and the gesture is therefore not redundant since it adds that missing piece of information crucial for a good understanding of the propositional content of the statement. For their part, the words 'dans c'couloir' (*in this corridor*) could be absent without changing the point the addresser wishes to make, as the gesture alone refers to what is commented upon. Here again, because the deictic gesture has a referential function only, it cannot realise the illocutionary force.

To sum up, (5) and (8) could have been used without any words, and in both cases the gesture has a performative function and realises the illocutionary force of the utterance. In (9) and (10), though the gesture has the same function and also realises the illocutionary force, words are indispensable. It is the other way around in (7), (13) and (14): as they could have been used without speech, the gestures in these examples are far more crucial than words, but they do not contribute to the illocutionary force in any way whatsoever and instead have a referential function. The expressive in (6) is a bit special in that it could have been produced without words and, in such a case where speech would have been absent, the deictic would have been a contribution to both the propositional content and the illocutionary force of the whole utterance. For their part, (11) and (12) illustrate examples where (1) speech can actually not be left out and (2) the accompanying gesture, which has a referential function, has no impact on the illocutionary force.

5.2 Iconic Gestures

This subsection discusses two examples of iconic gestures.

The first group was discussing inequality between men and women. In (15), the addresser was specifically recounting an anecdote concerning French politician Christiane Taubira, which illustrates that some women are not respected.

(15) Christiane Taubira, qui qui vient de je n'sais plus quelle région outre-A *outre-Atlantique euh de France tu vois genre euh* […].
"Christiane Taubira, who who comes from I don't know which region across the *across the Atlantic er of France you know like er* […]."

While saying 'France', the addresser depicts the country with a palm down open hand gesture with her hand making a circular movement. As all five fingers are fully stretched, which renders an idea of width, it indicates that the country is big. So, thanks to the gesture, the addressee acquires additional information (concerning the size of the country) with regard to the assertion. Since the iconic gesture in this example adds content-related information and has a representational function, it contributes to the propositional content. Nonetheless, speech is also important here, as the gesture would not make sense alone.

The following utterance was produced while the second pair of participants was talking about new technologies. In (16), the addresser is imitating someone who sticks tape on their computer's camera in order to avoid being videotaped unawares.

(16) *Tu vois, genre, j'ai toujours trouvé ça ridicule les gens qui mettent des des scotchs sur leurs ordinateurs*, mais je l'fais maintenant quoi.
"*You know, like, I always thought it was ridiculous people who put sticky tape on their computers*, but I do it now."

As she mentions the tape, the addresser depicts its position and size with the help of her fingers: her thumb and index are bent parallel to each other. These pieces of information (i.e. position and size) are not expressed in words and the gesture is thus crucial, but so is the spoken part in order to make a well-formed assertion. This

gesture only impacts the propositional content and is not able to realise the illocutionary force, as it has, as with the previous example, a representational function.

The two iconic gestures discussed in this subsection are very similar in that, in both of these examples, the gesture is essential, as it gives pieces of information not expressed in speech. Nonetheless, neither of them could be understandable alone. As the gesture has a representational function in the two cases, it is impossible for either iconic to impact the illocutionary force.

5.3 Metaphoric Gestures

Eight metaphoric gestures are discussed in this last subsection.

Let us consider again the first group of participants. The addresser in (17) is talking about equality between men and women and the degree of access they can gain in politics and companies.

(17) En fait, j'pense que *c'est l'accessibilité ... J'crois que* […].
"Well, I think that *it's access ... I think that* […]."

In this example, the addresser illustrates ACCESS by making a large palm up open hand gesture towards her interlocutor, which echoes the concept for a better understanding on the part of the addressee. The gesture thus has a representational function and is therefore part of the propositional content, that is of the stated viewpoint. This metaphoric gesture cannot make sense alone.

The participants were now discussing the idea that one cannot just force change on languages because they evolve naturally of themselves.

(18) *Les langues genre tendent toujours à la simplification tu vois, de plus en plus, tu vois*.
"*Languages like they always tend to simplification you know, more and more, you know*."

When the addresser says 'tendent' (*tend*), she moves her hand in front of her, so much so that her arm is fully stretched. Then she repeats the movement on 'de plus en plus' (*more and more*) to indicate that languages become more and more simple, which helps the addressee to visualise the evolutionary process of languages. In other words, as with (17), the metaphoric gesture illustrated here has a representational function and contributes to the propositional content only. The gesture, though very helpful, could not be used by itself and needs speech to produce a coherent assertion.

Following the discussion on the Islamic veil mentioned earlier, the addresser in (19) is reflecting on the problems a Muslim woman may encounter because of her religion once she is trapped in a cycle of beliefs.

(19) Parfois, quand t'es *quand t'es dans l'engrenage, tu t'rends pas compte de ces choses-là*.
"Sometimes, when you're *when you're in the cycle, you don't realise those things*."

When she says 'engrenage' (*cycle*), the addresser uses a metaphoric gesture that imitates a cyclic movement to echo her words. This gesture thus has a representational function. It could be used without speech, but the addressee may understand a (close but) slightly different idea. For example, this gesture could be interpreted as meaning 'brainwashed' (instead of 'in the cycle'). With both speech and gesture the addressee, however, gets the whole picture and the stated propositional content is crystal clear.

The following utterances were issued by the second group. Recall these participants discussed, amongst two other topics, the impact of new technologies on relationships. The utterance in (20) was issued while they were more specifically talking about Tinder.

(20) À un moment, *ça a pris d'l'importance*.
"At one point, *it took on importance*."

The addresser in this example metaphorically shows to what extent applications such as Tinder took on importance by using both hands with a palm up away from body gesture, fully stretched fingers and enlarged movement indicating amplitude, which indirectly relates the gesture to the word it accompanies, namely 'importance'. The gesture thus has, here again, a representational function and thus contributes to the propositional content. As with (19), the gesture could be used alone but with a possible slight change in the interpretation of the statement: one could understand 'it exploded' instead of 'it took on importance', which is a bit farther on the continuum.

One of the participants now develops the idea that the problem with new technologies does not come from the devices we use but rather from the fact that people are not properly educated to correctly use them. In the following example, as the addresser says 'éduquer' (*educate*), she makes a rolling movement that metaphorically indicates improvement, thence education.

(21) Et d'éduquer d'*éduquer à ça en fait.*
"And to educate to *educate on that actually*."

The gesture in (21) could not make sense without any speech, but it helps the addressee to visualise the concept EDUCATION. This gesture therefore has a representational function and indeed does not affect the illocutionary force, as the act is in all cases an act of stating.

In (22), the addresser argues that there are much safer communication platforms than the new Messenger that Facebook created for children.

(22) *Y a d'autres plateformes de communication beaucoup plus sûres.*
"*There are other much safer communication platforms*."

Here the addresser metaphorically illustrates COMMUNICATION by using both hands, palms facing each other vertically, with crooked fingers in a spiral movement to indicate connection, which helps the addressee to better visualise the concept. Once again, the gesture is part of the informative propositional content, as it has a representational function, and it could not be understood alone.

The addresser in the example below is saying that computing is too abstract for her, so much so that it escapes her. She illustrates ABSTRACTION by making a large gesture with flat hands away from her body.

(23) *C'est tellement abstrait pour moi, l'informatique, que euh* pff […].
"*It's so abstract for me, computing, that er* pff […]."

Since the concept of ABSTRACTION is mentioned in words, the gesture is redundant with part of the spoken component of the informative utterance. This metaphoric gesture has a representational function and is part of the propositional content. Here the gesture could make sense without speech but not without altering it. Without the accompanying word, the gesture could be interpreted as meaning 'complicated' (instead of 'abstract').

The utterance considered hereafter was produced while the participants were discussing publicity about autistic children and research on the subject.

(24) *T'es tellement coupé du monde dans la recherche.*
"*You're so cut off from the world in research*."

In this example, the addresser metaphorically shows a wall with a palm vertical open hand gesture to indicate that researchers are cut off from the world. For emphasis, the gesture is repeated in a zigzagging movement, which visually indicates to the addressee that there is a clear-cut separation between researchers and the rest of the world. Redundancy with speech is therefore undeniable even if, if used alone, this gesture could express REJECTION, which is still pretty close to the idea expressed in words. As with the other examples in this subsection, the gesture in (24) contributes to the propositional content since it has a representational function, as it gives a visual illustration of what is claimed in words.

As a conclusion, the gestures produced in (19), (20), (23) and (24) could have made some sense without words but not without a possible change in interpretation. As for the gestures in (17), (18), (21) and (22), there is no way they could have been used alone. None of the examples discussed in this subsection realise the illocutionary force. Instead, they all contribute to the propositional content, as all of them have a representational function.

6 Results

No participant used gestures alone, whether deictic, iconic or metaphoric. Consequently, no gesture produced an illocutionary act by itself, that is without speech production. However, six of the deictics could have been understood without any accompanying words and four metaphoric gestures could have rendered a close idea of what was expressed orally. The difference most certainly lies in that deictics are usually conventionalised, whereas iconic and metaphoric gestures are not.

Regarding redundancy, only two iconic and two deictic gestures added extra information. Nonetheless, it is worth highlighting that, for most deictics, it is actually not the gesture that is redundant with the words it accompanies. In fact, for those deictics that could have been used alone, it is the accompanying words that are redundant.

Now, except for one or two directives (depending on interpretation) and one expressive, all gestures were used in the making of assertive acts.

With regard to the contribution of gestures, this study has confirmed that iconic and metaphoric gestures, that is gestures that have a representational function, cannot realise the illocutionary force and typically contribute to the propositional content only, as they exclusively depict (some) aspects of what is being talked about. The deictics that have a referential function also contributed to the propositional content. That being said, four deictics (i.e. those that have a performative function) realised the illocutionary force (or strength in one case). Nonetheless, in some cases where speech could have been (partially) absent, the deictic could have been linked to both the propositional content and the illocutionary force at the same time. This observation broadens the general idea — discussed in the background section — that deictics can be linked either to the propositional content or to the illocutionary force. This should, however, be further investigated in future research concentrating on the use of deictic hand gestures more specifically.

7 Discussion

The literature on speech act theory (and related) does more than mentioning that, apart from words, so-called non-verbal communicative elements, such as sighs, facial expressions, body movements and the like, can also be used to produce communicative acts (see Austin, 1962; Searle, 1969; Bierwisch, 1980; Clark, 1996; Wharton, 2009; Cuffari, 2012; and Sperber & Wilson, 1995, 2012a, 2012b, 2015, to mention but a few). However, it seems fair to say that speech act theory, though it has been discussed for more than half a century now, has not given much room for an in-depth study of hand gestures. This paper did not have the pretence of doing them full justice, but the purpose was at least to give gestures a (bigger) voice in this specific subfield of pragmatics.

This paper described a qualitative study of communicative hand gestures produced during conversations and their relation to the propositional content and/or the illocutionary force. In the pilot experiment presented here, I was open to the study of all illocutionary acts but found that participants tend to extensively produce assertives more than other communicative acts. Besides that, and as already mentioned in the previous section, no gesture was used without speech even though the words accompanying the gesture were sometimes redundant with it. For these two reasons, further research should design experiments enabling the investigation of both assertives and non-assertives that would be made up of gestures only. More concretely, because (1) deictics showed some potential for being used in the production of directives and expressives, (2) most of them could have been understood without words and (3) some of them contributed to the illocutionary force, this preliminary investigation leads to the question whether deictics could produce directives, expressives or any other acts without any speech. In fact, if the deictic gesture has a performative function and if it alone expresses content-related information, then it would both realise the illocutionary force and contribute to the propositional content.

Remember I have argued that, if the index is fully stretched and the movement is enlarged and accelerated, the gesture will be interpreted as expressing an order rather than a suggestion, for instance. In more general terms, it means that, depending on their shape (e.g. fully stretched index finger versus palm up open hand) and movement character (enlarged versus reduced and accelerated versus decelerated), deictics express different forces. It is thus not farfetched to further hypothesise that, if used alone, two deictics varying in shape and movement character would be interpreted as two (slightly) different illocutionary acts. And again, these deictic gestures would both contribute to the propositional content and realise the illocutionary force simultaneously. To verify this hypothesis, further research should design a study that investigates the use of pro-speech deictic hand gestures.

Acknowledgements

This paper is based on my second master's thesis, which was supervised by Philippe De Brabanter, whom I thank for commenting on the thesis itself as well as on a previous draft of this paper. Thank you also to three anonymous reviewers for giving feedback on my manuscript and to Nicole Dehé for providing useful remarks on a few sections, especially the method section. Any remaining potential inaccuracies are my sole responsibility.

References

Austin, John L. (1962). *How to do things with words*. Oxford: Clarendon Press.

Bierwisch, Manfred (1980). Semantic structure and illocutionary force. In John R. Searle, Ferenc Kiefer, & Manfred Bierwisch (Eds), *Speech act theory and pragmatics* (*Texts and Studies in Linguistics and Philosophy, 10*) (pp.1-35). Dordrecht, Boston: D. Reidel.

Bressem, Jana, Ladewig, Silva H., & Müller, Cornelia (2013). Linguistic annotation system for gestures. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, & Sedinha Tessendorf (Eds), *Body - language - communication: An international handbook on multimodality in human interaction* (*Handbücher zur Sprach- und Kommunikationswissenschaft (HSK), 38*(1)) (pp.1098-1124). Berlin, Boston: De Gruyter Mouton.

Cassell, Justine, McNeill, David, & McCullough, Karl-Erik (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*(1), 1-33.

Clark, Herbert H. (1996). *Using language*. Cambridge: Cambridge University Press.

Cuffari, Elena (2012). Gestural sense-making: Hand gestures as intersubjective linguistic enactments. *Phenomenology and the Cognitive Sciences, 11*(4), 599-622.

Enfield, Nick J. (2012). *The anatomy of meaning: Speech, gesture, and composite utterances* (*Language, Culture & Cognition, 8*). New York, Cambridge: Cambridge University Press.

Fein, Ofer, & Kasher, Asa (1996). How to do things with words and gestures in comics. *Journal of Pragmatics, 26*(6), 793-808.

Haviland, John (2000). Pointing, gesture spaces, and mental maps. In David McNeill (Ed.), *Language and gesture* (*Language, Culture & Cognition, 2*) (pp.13-46). Cambridge: Cambridge University Press.

Kendon, Adam (1986). Some reasons for studying gesture. *Semiotica, 62*(1/2), 3-28.

Kendon, Adam (2000). Language and gesture: Unity or duality? In David McNeill (Ed.), *Language and gesture* (*Language, Culture & Cognition, 2*) (pp.47-63). Cambridge: Cambridge University Press.

Kendon, Adam (2004). *Gesture: Visible action as utterance*. Cambridge, New York: Cambridge University Press.

Ladewig, Silva H., & Jana, Bressem (2013). A linguistic perspective on the notation of gesture phases. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, & Sedinha Tessendorf (Eds), *Body - language - communication: An international handbook on multimodality in human interaction* (*Handbücher zur Sprach- und Kommunikationswissenschaft (HSK), 38*(1)) (pp.1060-1079). Berlin, Boston: De Gruyter Mouton.

McNeill, David (1985). So you think gestures are nonverbal? *Psychological Review, 92*(3), 350-371.

McNeill, David, & Levy, Elena (1982). Conceptual representations in language activity and gesture. In Robert J. Jarvella & Wolfgang Klein (Eds), *Speech, place, and action: Studies in deixis and related topics* (pp. 271-295). Chichester: John Wiley.

Searle, John R. (1969). *Speech acts: An essay in the philosophy of language*. London: Cambridge University Press.

Sperber, Dan, & Wilson, Deidre (1995). *Relevance: Communication and cognition*. Oxford: Blackwell.

Sperber, Dan, & Wilson, Deidre (2012a). A deflationary account of metaphors. In Deidre Wilson & Dan Sperber (Eds), *Meaning and relevance* (pp.97-122). Cambridge, New York: Cambridge University Press.

Sperber, Dan, & Wilson, Deidre (2012b). Introduction: Pragmatics. In Deidre Wilson & Dan Sperber (Eds), *Meaning and relevance* (pp.1-28). Cambridge, New York: Cambridge University Press.

Sperber, Dan, & Wilson, Deidre (2015). Beyond speaker's meaning. *Croatian Journal of Philosophy, 15*(44), 117-149.

Wharton, Tim (2009). *Pragmatics and non-verbal communication*. Cambridge: Cambridge University Press.

Demo and Breaking News Presentations

Towards Situation- and User-Adaptive Voice Output: Classifying Driver Personality in Context

Daniela Stier
Mercedes-Benz AG &
Ulm University, Germany
daniela.stier@daimler.com

Wolfgang Minker
Ulm University
Germany
wolfgang.minker@uni-ulm.de

Ulrich Heid
Hildesheim University
Germany
heidul@uni-hildesheim.de

Abstract

The current trend towards natural, intuitive speech interaction requires intelligent dialogue systems, that flexibly react according to an individual *user* in the interaction *context*, *e.g.*, by adapting the language style of voice output. A precondition for the efficient realization of such adaptive system behaviour is the reliable identification of these parameters. Against this background, this work serves as a basis for the development of user-adaptive in-car voice output and presents our classification results based on acoustic features.

1 Introduction

Intelligent Spoken Dialogue Systems (SDSs) are expected to provide an increasingly natural and intuitive human-machine interaction, *e.g.*, by flexibly adapting their voice output to the requirements of an individual user. An interaction as efficient as possible based on the human model (*cf.* interactive alignment model; Pickering and Garrod, 2004; Branigan et al., 2010) is particularly relevant in situations where voice-based communication with an SDS runs as a secondary task. Here, speech needs to be produced and processed in parallel and should not distract from the primary task, such as driving a car. In such dual-task scenarios, the development of an adaptive SDS depends both on individual user characteristics and the context of interaction. From a user's point of view, for instance, the characterization of user language in the interaction context is essential for the design of adaptive voice output strategies. Studies on the language used by drivers, *e.g.*, revealed clear differences in complexity depending on the manifestation of their personality traits (Stier et al., 2020a) and the driving situation (Stier et al., 2020b). From the SDS perspective, a requirement for user- and situation-adaptive voice output is the reliable identification of both the user and the driving context. While the automatic identification of personality traits in interpersonal or human-machine interaction has already been investigated extensively (*e.g.*, Mairesse and Walker, 2007; Sidorov et al., 2014; Subramaniam et al., 2016), to our knowledge no research exists which extends this aspect by the interaction context of a primary task. In this paper we present our results of this fundamental work. Our approach can be summarized in the following steps:

1. Collect spoken data for diff. driving contexts
2. Extract acoustic features
3. Create user clusters based on personality traits
4. Build and test statistical models for driving situations and user clusters based on features

2 Data Collection and Processing

We collected data as WoZ experiment in a fixed-base simulator (Mercedes E-Class) with 180° monitor. After an initial training in the car (5min), each participant followed a lead vehicle at distance of 100m and drove manually with SAE 0 on a highway (100km/h) and in a city (50km/h; 8min each). There was light oncoming traffic and additional traffic lights placed each km in the city, which changed to green when the participant approached to prevent motion sickness. In order to collect as much data as possible, the simulated voice-based interaction was limited to system-initiated small talk questions of the WoZ (*e.g.*, "What is your favourite hobby and how did this hobby come about?"). Each participant was instructed in advance to answer spontaneously and in his or her own language style (i.e. without overemphasis, as with a human interlocutor). Each participant was recorded on video with a camera installed in the vehicle (A-pillar).

Overall, 44 subjects (28 male, 16 female) with an average of 43 years (*sd* 13.24) participated and self-assessed their Big Five Personality traits on five-point Likert scales (German questionnaire; Rammstedt and Danner, 2016). Our data collection comprises 376 answers for the highway (*mean* 62.67, *sd* 18.67) and 331 for the city (*mean* 55.17, *sd* 16.93).

Table 1: Overview of extracted features and their variations.

Spectral Centroid	mean and standard deviation of spectrum
Energy Difference	difference in energy between low (<500Hz) and high (>500Hz) frequency
Intensity, Pitch	maximum, mean, minimum and standard deviation of intensity and pitch
MFCCs and deltas	mel-frequency cepstrum coefficients (16 features) and changes (16 features)
Tempo	mean speaking rate (in beats per minute)

Table 2: Distribution of participants and answers.

UC	**1**	**2**	**3**	**4**	**5**	**6**	$\sum$
Sub.	6	6	8	4	12	8	44
Age	41.7	41.3	45.6	41.8	43.5	41.9	42.9
O	3.72	2.90	3.14	4.00	3.75	3.63	3.52
C	3.96	3.85	4.46	4.69	3.94	3.92	4.09
E	3.56	3.54	3.73	4.75	3.78	3.70	3.78
A	3.17	3.53	4.05	3.63	3.81	4.15	3.77
N	2.83	2.46	1.89	1.59	2.00	2.75	2.26
H	50	60	65	40	95	66	376
C	47	51	53	37	87	56	331

Note: O-N= Big Five traits, H= number of answers on highway, C= number of answers in city

2.1 Acoustic Feature Extraction

The voice recordings were tailored to the individual user responses ranging from 3 to 400s (*mean* 41.78, *sd* 40.19) without further processing. Based on Landesberger et al. (2020), we extracted 43 features (Table 1) using `Praat` (Boersma and Weenink, 2020) and `librosa` (McFee et al., 2015).

2.2 Labelling User Personality and Context

Personality manifests multiple traits simultaneously. Thus instead of investigating traits individually, we combined the assessed Big Five traits als clustering variables and obtained six user clusters (UCs; `SPSS` 2-step clustering, av. silhouette 0.4, size ratio 3.25). We additionally distinguished the driving context (DC) during which a response was recorded. Table 2 summarizes details of our data.

3 Classifying User and Situation

We performed stratified five-fold cross-validation and compared performance measures (accuracy, precision, recall, f-measure; macro-averaging) for a Multinomial Logistic Regression classifier[1] (MLR), Random Forest Classifier[2] (RFC) and Support Vector Machine[3] (SVM). We classified UC and DC separately before combining them (UC & DC). In either case, the best results were obtained for RFC (Table 3). However, all classifiers performed remarkably well. The application of combined resampling (under-, oversampling, smote) and feature selection (cross-val. recursive elimination) methods did not significantly improve classification results.

Table 3: Classification results.

		acc	**prec**	**rec**	**f**
UC	MLR	0.774	0.770	0.762	0.762
	RFC	**0.979**	**0.982**	**0.974**	**0.977**
	SVM	0.815	0.805	0.793	0.793
DC	MLR	0.939	0.939	0.939	0.939
	RFC	**0.990**	**0.990**	**0.991**	**0.990**
	SVM	0.955	0.955	0.956	0.955
UC & DC	MLR	0.723	0.728	0.718	0.715
	RFC	**0.966**	**0.971**	**0.962**	**0.965**
	SVM	0.799	0.798	0.787	0.784

4 Conclusion and Future Work

In this paper we presented our results for the automated classification of both the user personality and the driving situation based on acoustic features. This investigation serves as a basis for the development of an in-vehicle SDS with user and situation-adaptive voice output. All of our statistical models achieved remarkable classification results. Thus a reliable identification of the driver personality in the driving context seems possible. For this purpose, acoustic features serve as a more than suitable source. They especially led to a precise DC categorization, presumably due to the unprocessed driving noises. They were surprisingly also able to successfully differentiate between our UCs. Due to our small and special data set, we will validate our findings in real driving situations in future work.

[1] C=80, penalty=l1, solver=liblinear

[2] n_estimators=500, bootstrap=false, max_features=log2

[3] C=0.1, degree=1, gamma=0.001, kernel=linear

References

Paul Boersma and David Weenink. 2020. Praat: Doing phonetics by computer [Computer program, version 6.1.09].

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

Jakob Landesberger, Ute Ehrlich, and Wolfgang Minker. 2020. Do the urgent things first! – Detecting urgency in spoken utterances based on acoustic features. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, New York, NY, USA. ACM. In press.

François Mairesse and Marilyn A. Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

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

Beatrice Rammstedt and Daniel Danner. 2016. Die Facettenstruktur des Big Five Inventory (BFI). *Diagnostica*.

Maxim Sidorov, Stefan Ultes, and Alexander Schmitt. 2014. Automatic recognition of personality traits: A multimodal approach. In *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*, pages 11–15.

Daniela Stier, Katherine Munro, Ulrich Heid, and Wolfgang Minker. 2020a. Personality traits, speech and adaptive in-vehicle voice output. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, New York, NY, USA. ACM. In press.

Daniela Stier, Katherine Munro, Ulrich Heid, and Wolfgang Minker. 2020b. Towards situation-adaptive in-vehicle voice output. In *2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. ACM. In press.

Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. 2016. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European Conference on Computer Vision*, pages 337–348. Springer.

A corpus of Swedish conversations with a librarian

Magdalena Sandahl[1]
magdalena.sandahl@kultur.goteborg.se

Simon Dobnik[2]
simon.dobnik@gu.se

Mehdi Ghanimifard[2]
mehdi.ghanimifard@gu.se

[1]Stadsbiblioteket Göteborg
Gothenburg City Library, Sweden

[2]CLASP and FLOV
University of Gothenburg, Sweden

Abstract

We describe a novel Swedish corpus of social and task related dialogues that has been collected from the interactions of library patrons with the library staff. The corpus is intended to aid the development of language technology tools, including a dialogue system, that can be incorporated in the existing chat solution to facilitate and improve interaction between patrons and library staff. We describe the properties of the corpus collected so far, its annotation with topics and evaluation of annotation.

1 Introduction

During a reference conversation a librarian may be asked to place a reservation for a book. The patron may go on and ask which book the librarian is reading right now. In this way, the conversation evolves from a simple request to a chitchat about personal genre preferences. A dialogue of this kind may occur when a user is seeking help at a reference desk in the library, via a phone call or via chat software. The goals of this project are to develop dialogue technology tools, including a dialogue system, to reduce the workload and to increase efficiency and quality of the responses. The idea of chat-bots in libraries in not new (Mohammed Ali, 2019), there are few projects in the field: for example, 'Bizzy' from University Library of Oklahoma (Young, 2019) or ANTswers from UC Irvine Libraries (Kane, 2016). An interesting study of the likely impact on human employment by AI/robotics, with a specific focus on library workers and users can be found in (Phillips, 2017), where the feasibility and acceptability are discussed. Interactions between patrons and library staff are often task-oriented but also social.

There are limited dialogue corpora for Swedish, especially in the area of task-oriented dialogue, and to the best of our knowledge there have been no previous corpora of conversations in this domain. However, several other language technology resources exist for Swedish. For example, in February of 2020, the National Library of Sweden released three pre-trained language models based on BERT and ALBERT. These models are trained on approximately 20GB of text, i.e. 200 million sentences from various sources aiming to provide a representative BERT model for Swedish text.[1] One of the aims of this project is also to try to utilise such external resources for development of new dialogue systems for Swedish.

Stadsbiblioteket Göteborg (SB, Gothenburg City Library, Sweden) uses LibraryH3lp as a virtual service software platform.[2] The software is designed as a chat system, allowing reference service and interaction between librarians and library users. One of the platform's features allows storage of chat transcripts, annotating the data, i.e. tagging the chats with descriptive categories as well as downloading metadata in CVS format for further analysis. The platform also allows integration of external tools in conversations with patrons through APIs. The goal of the ongoing corpus collection and annotation is to aid the development of language technology resources such as topic modeller and chatbot that will be incorporated in LibraryH3lp, save some resources and predominantly heighten the service quality and quantity. By developing this corpus, however, we are also proving a novel and interesting conversational resource for Swedish.

2 Data collection and annotation

Corpus collection and annotation started in January 2020. The data is still being collected and incrementally annotated. As of the middle of June 2020, the dataset comprises over 1600 conversations with

[1]https://github.com/Kungbib/swedish-bert-models
[2]https://libraryh3lp.com/

ca 15 chats added daily. The dataset is incrementally annotated for topics relevant in the domain of library conversations, for example: *book and subject search*, *app Libby* and *purchase requests* and up to now the annotation has converged on 30 topics which are incrementally revised. All annotation was carried out by the same librarian and was later validated by other librarians (Section 3). As our goal is development of tools that would assist communication between patrons and librarians, we opted for annotating entire conversations for topics. Depending on the detected topic, the tools would either answer the question or chitchat with the patron or rout the query to a librarian (Smith et al., 2011; Pulman et al., 2010).

Here is an example of a summarised conversation. A patron gets a written overdue library book notice. She contacts the library through chat to ask about the notice, claiming that she had already returned the book. After a short conversation, checking the patron's library account and explaining the rules for overdue fines, the chat evolves from a task-oriented to social-oriented when the patron comments that they enjoyed interacting with a librarian through chat: "Ok! Tack för hjälpen! Så bra att kunna chatta!" This chat is annotated with the following topics: (i) check out, extend loan and check in, (ii) overdue fines, and (iii) social chat. The tags refer to simple actions (*naxos*, *app PressReader*) or complex ones (check out, extend loan and check in) and related to and motivated by how the work of librarian is organised. In addition to the topics tags, a dialogue may also be flagged with the label *admin* (when the conversation is carried out between two librarians) or *split chat* (when either the librarian or the user does not respond or when the interaction gets split into two and continues as a new one due to technical problems).

3 Corpus validation

As mentioned above, the tag set is not predefined and can be expanded. Hence, after additional three weeks from the beginning of data collection, a poll was sent out to 21 librarians working with the LibraryH3lp chat service, in order to validate the corpus and verify the topics. The poll was anonymous and all librarians responded to it. They received 18 authentic anonymised chat interactions that were randomly selected from the saved data.

The chats were presented as two tasks. In the first task which consisted of 15 dialogues, each respondent had the option to choose one of the four topics that in their opinion best describes the conversation. They also had a possibility to create their own tag instead in case they would disagree with all four pre-given choices. Three of four topics were relevant to the conversation but to a different degree (the goal was to measure annotator preference, e.g. *app Libby* or *suggestions, opinions and questions*), the fourth one was chosen at random to check if the participants are providing useful answers. The results in Table 1 indicate that dialogues are not restricted to a single topic and must be assigned multiple classes which has implications for computational modelling. In the second task of the study, the librarians were presented with three chats which they were asked to tag freely. The results were compared with the original annotations and ten new additional topics were identified.

Dialogue	Majority topic	%
D1	låneregler	0.571
D2	app uTalk	0.952
D3	fysisk utrustning & verktyg	0.905
D4	bok- & ämnessökning	0.810
D5	bok- & ämnessökning	0.952
D6	reservationer	0.667
D7	TV-spel	0.667
D8	förseningsavgifter, krav & inkasso	0.857
D9	lån, omlån & återlämning	0.476
D10	programpunkter & kalendarium	0.952
D11	app PressReader	0.524
D12	fjärrlån	0.571
D13	app Legimus	0.952
D14	bibliotekskort & PIN-kod	1.000
D15	grupprum	1.000
D16	Utskrifter	0.095
D17	cineasterna	0.238
D18	bokcirkelkassar	0.143

Table 1: The agreement (in %) of domain experts on the most likely topic of each dialogue. The results indicate that several dialogues should be assigned more than one label as several topics are discussed.

4 Conclusions

Our future work will focus on (i) further extension of the corpus and refinement of the tags; (ii) application of the corpus in computational modelling; (iii) examining the relation between domain specific dialogue (specific tools and procedures at SB), task specific dialogue (conversations between patrons and librarians) and general social dialogue; (iv) releasing the corpus to the public domain so that it will be useful both for libraries and researchers in dialogue.

Acknowledgements

The research of Dobnik and Ghanimifard 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

Danielle A Kane. 2016. The role of chatbots in teaching and learning. *E-learning and the academic library : essays on innovative initiatives*, 131:1–26.

Salihin Mohammed Ali. 2019. Bots in libraries: They're coming for your jobs (or is it?). In *ALIA Information Online 2019*, Sydney, Australia. Research Collection Library, University of California, Berkeley School of Law.

David Phillips. 2017. Robots in the library: gauging attitudes towards robotics and AI, and the potential implications for library services. Master's thesis, City University of London, January.

Stephen G Pulman, Nigel Crook, Ramon Granell, Simon Dobnik, and Manjari Chandran-Ramesh. 2010. Deliverable D4.2.3: Report on topic detection with Dirichlet Process Priors. Technical report, Department of Computer Science, Oxford University, Oxford, United Kingdom.

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

Jeffrey R Young. 2019. Bots in the library? Colleges try AI to help researchers (but with caution). Technical report, EdSurge.

Detecting Urgency in Speech with Personalised Acoustic Features

Jakob Landesberger
Mercedes-Benz AG & University Ulm
jakob.landesberger@daimler.com

Ute Ehrlich
Mercedes-Benz AG
ute.ehrlich@daimler.com

Abstract

Finding an appropriate reaction to a spoken utterance with multiple intents is not easy for a spoken dialogue system. One way to simplify the dialogue is to prioritize urgent intents. For this purpose, the urgent part of an utterance must be reliably identified. This is possible purely with acoustic features directly from the audio signal. In this work, we consider the acoustic features for each speaker individually and show that the recognition can be improved by this personalization.

1 Introduction

Utterances in dialogue serve often more than one communicative function. Like giving feedback about the understanding of a question and answering the question in a single utterance. The ability of humans to easily process such multiple communicative functions and to react accordingly, allows for a swift and effective communication (Lemon et al., 2002). If an utterance contains two sequentially occurring intents, each relating to a different task or activity, it is often referred to as multi-intent (MI). Such utterances also occur in interactions with a machine (Shi et al., 2019). If there is a need for further clarification of all mentioned intents, it can be difficult for a system to find a suitable answer. Answering with a MI, too, can produce long utterances, which can be cognitively very demanding (Landesberger and Ehrlich, 2019a). To solve this problem, we proposed a model for prioritizing tasks when dealing with MI utterances (Landesberger and Ehrlich, 2019b). The task, referred by an intent, is prioritized according to various criteria and the remaining tasks are postponed. After solving the prioritized task, the postponed tasks can be resumed. One criterion for prioritisation is urgency. A task is urgent if the task has to be completed in a short amount of time, because if not, it loses relevance or other negative consequences occur. *"What do I do with the zucchini? Oh, the pan is hot. What is coming in now?"* Certainly, the speaker would be frustrated if the first mentioned task is considered before the second one. In order to realise a suitable dialogue behaviour for this problem, it must first be detected which part of the utterance is urgent.

2 Urgency Corpus

In order to develop a method for the detection of urgency in utterances we use data obtained with the help of the game "What is it?" (Landesberger et al., 2020). The players' task was to find a special symbol on a matrix. To exclude wrong symbols, questions that could be answered with yes or no had to be asked to a simulated system. This task was regularly interrupted by an easier but time-limited, urgent task. This resulted in a Corpus of spoken urgent and non-urgent utterances. Since the task types alternated quickly, many utterances contained both urgent and non-urgent intents, just like in the example in section 1. Although, the different utterances are not very comparable regarding the content, we decided on a semantically independent approach, to later tackle real world problems. Accordingly we tried to detect urgency based on acoustic features from the audio signal.

3 Acoustic Features

In our previous work (Landesberger et al., 2020) the analysis of 108 acoustic features like pitch, intensity or MFCCs, showed that depending on the speaker's phase it was more or less difficult to identify urgency. If the speaker switches from a non-urgent intent to an urgent intent, he is in the phase Transition (TRA). Conversely, if he changes from an urgent intent to a non-urgent one, he is in the Decline (DEC) phase. For each of these phases and a general distinction (ALL) in which all utterances were analysed, we first identified a feature subset that contains as few irrelevant and redundant (John et al., 1994) features as possible. For this purpose we used three different estimators with the recursive feature elimination method (Granitto et al., 2006): Logisitc Regression Estimator (LRE (Hosmer Jr et al., 2013)), Random Forest Estimator (RFE (Breiman, 2001)) and Gradient Boosting

		RFC	GBC	LRC	KNNC	MLPC	SVMC
ALL	LRE	.883 +1.20%	.795 +4.26%	.774 +7.27%	.814 +1.11%	.798 +10.28%	.888 +13.66%
	RFE	.883 +1.85%	.795 +5.12%	.774 +8.82%	.814 +1.71%	.801 +12.27%	.888 +14.00%
	GBE	.883 +1.93%	.796 +4.35%	.775 +8.20%	.814 +1.06%	.799 +11.67%	.887 +14.61%
TRA	LRE	.779 +4.35%	.769 +2.17%	.782 +8.32%	.712 +5.22%	.784 +13.70%	.767 +1.02%
	RFE	.774 +2.55%	.770 +2.44%	.782 +8.61%	.712 +6.63%	.787 +16.15%	.767 +2.58%
	GBE	.787 +4.40%	.779 +3.67%	.781 +8.42%	.687 +3.44%	.777 +8.95%	.784 +3.84%
DEC	LRE	.883 -1.38%	.891 -0.22%	.906 +3.33%	.878 +1.80%	.895 +1.25%	.891 -0.43%
	RFE	.889 -1.08%	.889 -1.08%	.895 +2.69%	.886 +2.95%	.886 +0.98%	.891 -1.63%
	GBE	.887 -0.92%	.890 -1.47%	.897 +1.97%	.877 +3.20%	.893 +1.22%	.901 +0.07%

Table 1: Accuracy and personalisation improvement in percent to detect urgent user requests over all the data (*ALL*), during the switch from non-urgent to urgent utterances (*TRA*) and during the switch from urgent to non-urgent utterances (*DEC*) for each classifier estimator combination

Estimator (GBE (Friedman, 2001)).

In order to automatically distinguish between urgent and non-urgent intents, we trained and tested six different classifiers with different supervised learning algorithms: Random Forest Classifier (RFC (Breiman, 2001)), Gradiant Boosting Classifier (GBC (Friedman, 2001)), Logistic Regression Classifier (LRC (Hosmer Jr et al., 2013)), K Nearest Neighbours Classifier (KNNC (Peterson, 2009), Multi-Layer Perceptron Classifier (MLPC (Pal and Mitra, 1992)) und Support Vector Machine Classifier (SVMC (Vapnik, 2013)).

We analysed the accuracies achieved by each estimator and classifier combination distinguishing between urgent and non-urgent intents. Each of these values was validated by a 10-fold stratified cross validation method. For the Transition phase, the SVMC in combination with LRE showed the best result with an accuracy of .759. During Decline, the best result was achieved with the SVMC and the RFE with an accuracy of .909. The LRE in combination with the RFC performed best with an accuracy value of .873 in the general distinction.

4 Personalisation

Acoustic features in speech, change strongly depending on the speaker. For example, men usually speak with a lower pitch than female speakers do. Accordingly, the absolute or normalized values of certain features are not best suited for the detection of urgency. If a male speaker speaks higher than he would normally speak during an urgency, this value can still be below the normal pitch of a female speaker. Accordingly, the consideration of individual differences and corresponding relative, personalised variations in acoustic features could improve the recognition of urgency. In the examined data set 40 participants uttered 10737 intents. With on average of 268.4 per test person, we adopted the min-max normalization formula to include the individual variance for each participant p and feature f:

$$x'_{pf} = |\frac{x_{pf} - min(x_{pf})}{max(x_{pf}) - min(x_{pf})}|$$

The new values x'_{pf} rank on a scale from 0 to 1. 0 corresponds to the lowest measured feature value of one participant in all his utterances and 1 to the highest. Based on the recalculated values, the already evaluated classifiers and estimators were retrained and re-evaluated. The results are shown in Table 1. In addition to the accuracy of the individual classifier estimator combinations, the percentage change compared to the non-personalized data is shown. It is noticeable that there is no improvement if the accuracy was already high before. During the Phase Decline (DEC) even minimal deteriorations occur. In the other comparisons (ALL and TRA) all values improved. Especially the SVMC and the MLPC benefited from the personalization.

5 Conclusion

In this paper we show that adaptation to the speaker can help to identify urgent intents in utterances more reliably. For this adaptation, however, the speaker must already be known to the system. Adaptation to more general characteristics, such as gender, might be practical, too.

To further improve the detection of urgency, we plan to consider other aspects of the spoken utterance and the context of the utterance, too. We assume that certain words such as deictic expressions, the task addressed, and the current situation of the speaker can all provide indications for urgency.

References

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. 2006. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*, volume 398. John Wiley & Sons.

George H John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier.

Jakob Landesberger and Ute Ehrlich. 2019a. Finding a meta-dialogue strategy for multi-intent spoken dialogue systems. In *Proceedings of the 3rd International Conference on Computer-Human Interaction Research and Applications - Volume 1: CHIRA,*, pages 171–176. INSTICC, SciTePress.

Jakob Landesberger and Ute Ehrlich. 2019b. Towards finding appropriate responses to multi-intents - spm: Sequential prioritisation model. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, London, United Kingdom. SEMDIAL.

Jakob Landesberger, Ute Ehrlich, and Wolfgang Minker. 2020. Do the urgent things first! detecting urgency in spoken utterances based on acoustic features. In *Proceedings of the 28th Conference on User Modeling, Adaptation and Personalization*, UMAP '20. Association for Computing Machinery.

Oliver Lemon, Alexander Gruenstein, Alexis Battle, and Stanley Peters. 2002. Multi-tasking and collaborative activities in dialogue systems. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-Volume 2*, pages 113–124. Association for Computational Linguistics.

Sankar K Pal and Sushmita Mitra. 1992. Multilayer perceptron, fuzzy sets, classifiaction.

Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Chen Shi, Qi Chen, Lei Sha, Hui Xue, Sujian Li, Lintao Zhang, and Houfeng Wang. 2019. We know what you will ask: A dialogue system for multi-intent switch and prediction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 93–104. Springer.

Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

Topic modelling for a virtual librarian assistant tool

Linnea Strand[1]
gusstrlip@student.gu.se

Robert Rhys Thomas[1]
gusroberth@student.gu.se

Simon Dobnik[2]
simon.dobnik@gu.se

Mehdi Ghanimifard[2]
mehdi.ghanimifard@gu.se

[1]FLoV
University of Gothenburg, Sweden

[2]CLASP and FLoV
University of Gothenburg, Sweden

Abstract

We outline a topic classifier for a virtual librarian assistant trained on a corpus of conversations between patrons and librarians. Because this conversation data is limited, we explore the possibility of training the system on a domain of artificial data generated from distilled versions of real chats and applying transfer learning techniques to augment learning on real data. By training the model on artificial dialogues and fine-tuning it on real data, we hope to achieve better overall performance for the classifier even when little data is available. This abstract describes the data generation (augmentation) process, the model architecture and presents some preliminary results.

1 Introduction

In the interaction between librarians and patrons there is an increasing need for language technology tools that would provide assistance to librarians, especially when the interaction is split between different modalities that include on-site personal conversations, conversations on the phone, email and chat as in Göteborgs stadsbibliotek (SB, Gothenburg City Library). One of the tools that was identified useful in this domain is a topic modelling system that detects a topic of the conversation in written chats and can then support a rule-based dialogue system or routes the conversation to a human librarian if the topic is better handled by them. A similar idea has been applied in dialogue systems as in (Smith et al., 2011) where topic modelling has be achieved through application of different data-driven techniques (Pulman et al., 2010) that supported a rule-based system.

The main aim of this project is to train a topic classification system that detects the topic of a patron conversation with a librarian. However, a major drawback of applying data-driven learning techniques is that they require a large number of domain-specific conversations. Within the project with SB a dataset of 1,600 conversations have been collected to date which were annotated with topics by a domain expert. While this is not a negligible amount of dialogue data for a novel dialogue domain, in terms of the data-driven techniques the size of the corpus is still limited. To address this limitation we introduce a novel way of expanding the dataset through artificial dialogue generation. To achieve this, we distil rules from the real data and generate additional samples from it. Effectively, this is an integration of a rule-based approach with a data-driven approach. The rules are translated into data on which the system is pretrained. This provides biases encoded in the rule to the system in the forms of distributions in the data upon which the system can then fine-tune real examples. If our experiment(s) are successful, this approach could increase the chances of companies and other organisations to bootstrap robust dialogue systems without having access to large amounts of domain-specific data.

2 Data Collection

The application which the current library chat function is built with is LibraryH3lp.[1] The software works as an instant messaging platform and contains functionality that allows librarians to tag each conversation. The tags were introduced and refined incrementally by a domain expert and senior librarian who identified 34 topics. However, further data collection and annotation may introduce new topics or lead to a refinement of the existing topics. It is important to note that each conversation may be assigned several tags, as a conversation may be about more than one topic, e.g. *app Libby* or *suggestions, opinions and questions*. These tags are what the topic classifier algorithm will be trained to predict in a multi-class classification task. So far, a total 1,600 chat transcripts have been collected over the

[1]https://libraryh3lp.com/

course of 6 months. Each chat is labelled with one or more of the 34 available topics that have been chosen by the domain expert.

3 Distilling and Generating

The process consists of two steps: (i) the real dialogues are first distilled into a set of templates, and then (ii) the templates are expanded back into several examples that represent our artificial dataset. (Larsson et al., 2000) proposes guidelines for applying distilling techniques for dialogue to templates. We ensure that the artificial transcripts remain true to the real samples but differ enough to provide features that may be important when the data is fed through the machine learning model. Distillation is achieved by splitting transcripts into utterances and rephrasing them. From the distilled templates we can generate multiple dialogues that follow the same basic structure as the real examples by varying certain slots in the templates (examples in Appendix A). The generations are limited to how many phrases are added to the templates and to the variations of the slots. We also apply simple data augmentation techniques such as Random Deletion and Synonym Substitution (Wei and Zou, 2019), (Zhang et al., 2015), (Wang and Yang, 2015). The augmentation techniques allow for up to around 10,000 unique examples per tag.

4 The Model

We employ a simple, two-step neural architecture to classify the topic(s). First, each dialogue – represented as a list of utterances – is encoded using the pre-trained Swedish BERT model[2] which is based on BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019) and is trained over 200 million sentences over various sources. Each embedded utterance is then passed through a max pooling layer before being concatenated with its speaker vector (1 for patron, 0 for librarian) and fed into an LSTM cell. Finally, each hidden state is fed into a linear classification layer, thus predicting the topic(s) at every time step[3] (see Figure 1).

5 Evaluation

We compare the performance of the model trained on real dialogues (169), generated dialogues (2908

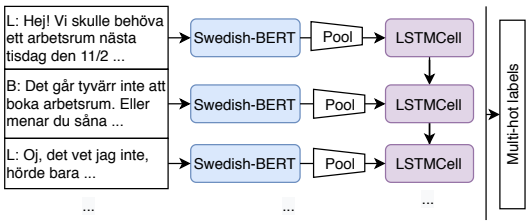


Figure 1: The model diagram.

dialogues, 3000 with duplicate generations removed) and mixed dialogues containing a combined bag of the two on three categories: *app Libby*, *app PressReader*, and *suggestions*. The generated dialogues have one target topic whereas real data may have several target topics (multi-class prediction). The held-out test set contains of 23 real dialogues. We train the model for 20 epochs. While the system can predict the topic incrementally for parts of dialogue, we evaluate the system for the entire dialogue by comparing *loss* on the test data for different models. We will report the figures in the final version of this paper.

6 Discussion and Conclusions

In this paper, a method for creating of training data has been introduced that is inspired by dialogue distillation techniques from dialogue system design coupled with data augmentation techniques. Additionally, we also investigate how we could benefit from transfer learning in contextual word embeddings such as BERT. We train a model of dialogue topic classification using this data and pre-trained knowledge and test it on real data. So far the experiments have given mixed results. We hope that the dialogue generation approach can be applicable in practically any domain, not just for a library chat bot. Furthermore experiments in future could explore different curriculum learning techniques on augmented data to be able to maximise the benefits of transferring knowledge from other domains. Future work includes improving our data generation method, evaluating the naturalness of the generated dialogues and including other forms of pre-trained knowledge.

Acknowledgements

The research of Dobnik and Ghanimifard 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.

[2] https://github.com/Kungbib/swedish-bert-models

[3] One variation of the system is designed for incremental topic classification, but evaluations are done on dialogue classification.

References

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*, arXiv:1909.11942 [cs.CL].

Staffan Larsson, Lena Santamarta, and Arne Jönsson. 2000. Using the process of distilling dialogues to understand dialogue systems. In *Proceedings of ICSLP 2000*, pages 374–377.

Stephen G Pulman, Nigel Crook, Ramon Granell, Simon Dobnik, and Manjari Chandran-Ramesh. 2010. Deliverable D4.2.3: Report on topic detection with Dirichlet Process Priors. Technical report, Department of Computer Science, Oxford University, Oxford, United Kingdom.

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

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason W Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv*, arXiv:1901.11196 [cs.CL].

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

A Appendices

A real transcript
L: hej! vad behöver du hjälp med? **U:** hej är studieplatserna fortfarande öppna så man kan sitta och plugga några timmar? **L:** nej tyvärr **U:** gäller det alla bibliotek i Göteborg eller bara stadsbiblioteket? **L:** det kan jag inte svara på jag tror att detsamma gäller alla bibliotek **U:** Tack
Artificially generated transcript
L: hej! vad behöver du hjälp med? **U:** hej! är studieplatserna fortfarande öppna så man kan sitta och plugga några timmar? **L:** det är dom tyvärr inte **U:** är det bara stadsbiblioteket eller alla bibliotek? **L:** det kan jag inte svara på jag tror att det är begränsat på de flesta bibliotek **U:** tack så mycket
A real transcript (translated to English)
L: hello! what do you need help with? **U:** hey, the study sites are still open so you can sit and study for a few hours? **L:** no unfortunately **U:** applies to all libraries in Gothenburg or just the city library? **L:** I can't answer I believe that the same applies to all libraries **U:** Thanks
An artificially generated transcript translated to English
L: hello! what do you need help with? **U:** hello! are the study sites still open so you can sit and study for a few hours? **L:** unfortunately they are not **U:** is it just the city library or all libraries? **L:** I can't answer that, I think it's limited in most libraries **U:** Thank you so much

Table 1: An example of real and artificial transcripts.

Sample 1		
Grammar	Template	Utterance
what_help	<librarian> hello! what do you need help with?	- <librarian> hello! what do you need help with?
user	- <user>	
hello	- hello - hi - hi! - hello!	
study_places	- are the study places still open to sit and study for a few hours? - is it possible to sit and work at the study places? - are the study places open? - do you still have study places?	<user> hello! are the study sites still open so you can sit and study for a few hours?
librarian	- <librarian>	
no	- no i'm afraid - unfortunately not - they are not unfortunately - i'm sorry but no - unfortunately they are not	<librarian> Unfortunately they are not
user	- <user>	
all_lib	- is it all libraries in Gothenburg or just the city library? - is this the same for all libraries? - is it just the city library or all libraries?	<user> is it just the city library or all libraries?
librarian	- <librarian>	
cant_say	- i'm afraid I can't say - I can not answer - I am not sure	
probably	- i think it is limited in most libraries - i believe it may be the same in all libraries - i think it is the same in all libraries	<librarian> I can't answer that, I think it's limited in most libraries
user	<user>	
thanks	- thank you so much - thank you - thanks a lot - thanks	<user> thank you so much

Table 2: Template generation

A Statistical and Machine Learning Investigation into the Effect of Demographic Context on Dialogue Phenomena

David Hopes and Julian Hough[1]
[1]Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK
d.hopes@se17.qmul.ac.uk, j.hough@qmul.ac.uk

Abstract

We describe a study into how disfluencies are measures of communication efficacy and can indicate a demographic identity between participants in dyadic dialogue. The relative distribution of disfluency types are calculated for ternary (FF, FM, MM) gender context. Tests for association show that certain disfluency types are dependent on the demographic context, classification models using these features score well, predicting the gender context from the provided disfluency rates. The model is compared to common text classifiers and performs with similar accuracy at a faster training speed and the disfluency features are added to these classifiers to improve performance.

1 Introduction

The literature shows there is potential value in looking into disfluencies as indicators of a difference in communication efficacy and hence demographic context. The analysis of disfluency and dialogue annotation distribution for participants of differing demographics has not yet taken place using the Switchboard data-set and doing so will provide insight into differing demographic context and its effect on communication.

2 Background & Related Work

2.1 Disfluency

Disfluency, an involuntary disruption in the flow of speech, can be one way that dialogue efficacy is affected. Suggestions were made that disfluency rates may be correlated with gender (Lickley, 1994) and further studies elaborated on this by showing that filled pause rates are indeed correlated by gender, specifically that filled pauses that are more prevalent in males than females (Shriberg, 1994). Disfluency rates were shown to be higher in task-oriented speech for males with 4.35±1.62 disfluencies per 100 words than females with 3.76±1.42 (Branigan et al., 1999). The Switchboard Dialogue Act Corpus (SwDA) (Jurafsky et al., 1997) is composed of transcripts, tagged with demographic metadata, and utterances, annotated with disfluency markers.

3 Approach

3.1 Disfluency & Gender Context

The relative frequencies of disfluency types per word for ternary gender categories are calculated and plotted as a heat-map in Figure 1 and appears to show a marked difference between the three groups.

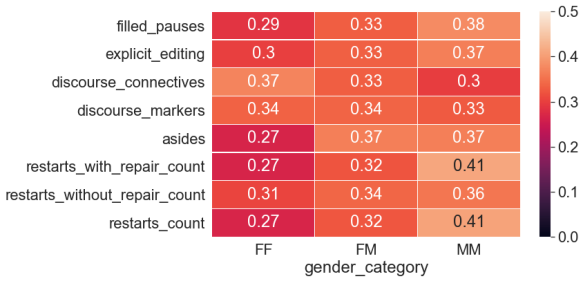


Figure 1: Heat-map showing relative frequency of disfluency per word rate for ternary gender context

4 Evaluation

4.1 Hypothesis Testing

Using chi-squared testing, the null hypothesis that there is no difference in the distribution of each disfluency type for female-only (FF), mixed (FM) or male-only (MM) gendered dialogue was rejected for all disfluency types, as shown in Figure 2.

4.2 Classification

A model was created to classify corpus utterances into the relevant labels. All classification was performed using 20-fold cross-validation on 80% of the corpus. The mean classification results for each

disfluency_types	chi_squared_score	p_value	null_rejected
restarts_with_repair_count	2191.327635	0.000000e+00	True
restarts_count	2190.436674	0.000000e+00	True
filled_pauses	1119.396647	8.435416e-244	True
explicit_editing	61.244858	5.021668e-14	True
discourse_connectives	59.624815	1.128849e-13	True
discourse_markers	57.108373	3.972595e-13	True
restarts_without_repair_count	37.288947	7.994810e-09	True
asides	13.653355	1.084455e-03	True

Figure 2: Table showing disfluency type and chi-squared testing result

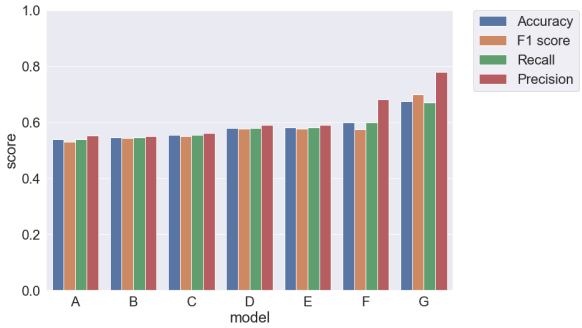


Figure 3: Chart showing classifier performance

fold were compared to standard text classifiers and the disfluency features were then combined to create more complex models.

4.2.1 Support Vector Classification

Using the disfluency types as features, a basic Support Vector Machine (SVM) classifier was trained resulting in an accuracy of 0.53±0.19 and F1 score of 0.524±0.18 (A). The parameters of the classifier were tuned with an L1 regularisation parameter with value of 1 resulting in an optimal accuracy of 0.538±0.18 and F1 score of 0.531±0.18 (B). The model training time was 384ms.

4.2.2 Comparison to Text Classifiers

To compare the performance of the disfluency classifier, standard text classification models were run, a bag-of-words classifier produced an accuracy of 0.55±0.12 and F1 score of 0.543±0.12 and took 7.03s (C). The TF-IDF classifier took 9.71s and produced a mean accuracy of 0.578±0.13 and F1 score of 0.576±0.13 (D).

4.2.3 Disfluency & Text Classification

The disfluency features were combined with the TF-IDF vector. The resulting classification took 10.5 seconds and produced an accuracy of 0.581±0.12 and F1 score of 0.576±0.12. This shows the accuracy of a standard TF-IDF classifier is improved upon by the inclusion of disfluency features (E).

An Adaboost classifier was implemented and the training time far exceeded the previous models at 8m16s, however the performance of the classifier increased with an accuracy of 0.600±0.11 and F1 score of 0.575±0.13 (F). The classifier was then trained on 80% of the corpus and tested on the remaining 20% providing an accuracy of 0.674 and F1 score of 0.700 (G).

4.3 Model Findings

The disfluency classifier worked well, achieving a mean accuracy of 53.8% on its own. Figure 3, shows the performance of the classifiers.

5 Conclusion

Studies of disfluency rates and context have been completed previously (Shriberg, 1996). However, they have only been analysed in the context of a single participant and their utterances, and not grouped participants and their utterances.

With the broader aim of investigating the relationship of how characteristics of dyads affect dialogue fluency and communicative efficacy, we found a dependency between participants' gender and disfluency rates. Previous research has been done into how to detect disfluency and similar phenomena, and models created for disfluency detection, but not the other way around, using disfluency to improve a language model. A number of machine learning models were successfully created as proof of concept that disfluency measures can be used to classify and improve classification of text. The model that used only disfluency features was much faster to train which could result in uses where training speed is valuable.

6 Limitations

Sparsity was a big limitation for the remaining forms of demographic context (age, dialect area and education) and getting reliable counts from a few transcripts that may be heavily influenced by a single caller reducing the ability to explore other types of context. The conclusions and test results from this analysis could differ when transferred to another data-set, therefore findings from this study could be SwDA specific.

References

Holly P Branigan, Robin J Lickley, and David McKelvie. 1999. Non-linguistic influences on rates of disfluency in spontaneous speech. *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 387–390.

Dan Jurafsky, E Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual.

Robin J Lickley. 1994. *Detecting Disfluency Spontaneous Speech*. Ph.D. thesis.

Elizabeth Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis.

Elizabeth Shriberg. 1996. Disfluencies in Switchboard. *Proceedings of International Conference on Spoken Language Processing*.

LDM: A Linear Dialogue Manager

Vladislav Maraev[⋆], **Jean-Philippe Bernardy**[⋆] **and Jonathan Ginzburg**[†]
[⋆]Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
{vladislav.maraev,jean-philippe.bernardy}@gu.se
[†]Laboratoire Linguistique Formelle (UMR 7110), Université de Paris
yonatan.ginzburg@univ-paris-diderot.fr

Abstract

For this system demonstration, we presents a modular dialog manager comprised of the following components. At the lowest level, we have implemented a proof-search engine based on linear logic. This engine features unification. In this engine we have implemented a suite of generic dialog-management rules, inspired from Ginzburg's KoS framework, such that firing a rule correspond to an information-state update of the agent. These generic rules can then be complemented by domain-specific rules. We show in particular how a question-answering agent, such as one operating an information kiosk, can be implemented.

1 Introduction

A key aspect of dialogue systems is the coherence of the system's responses. In this respect, a key component of a dialogue system is the dialogue manager, which selects appropriate system actions depending on the current state and the external context.

Two families of approaches to dialogue management can be considered: hand-crafted dialogue strategies (Allen et al., 1995; Larsson, 2002; Jokinen, 2009) and statistical modelling of dialogue (Rieser and Lemon, 2011; Young et al., 2010; Williams et al., 2017). Frameworks for hand-crafted strategies range from finite-state machines and form-filling to more complex dialogue planning and logical inference systems, such as Information State Update (ISU) (Larsson, 2002) that we employ here. Although there has been a lot of development in dialogue systems in recent years, only a few approaches reflect advancements in *dialogue theory*. Our aim is to closely integrate dialogue systems with work in theoretical semantics and pragmatics of dialogue.

We believe that it is crucial to use formal tools which are most appropriate for the task: one should be able to express the rules of various genres of dialogue in a concise way, free, to any possible extent, of irrelevant technical details. In the view of Dixon et al. (2009) this is best done by representing the information-state of the agents as updatable sets of propositions. Subsets of propositions in the information state can be treated independently, and, therefore, a suitable and flexible way to represent updates is as propositions in linear logic.

By using well-known techniques which correspond well with the intuition of information-state based dialogue management, we are able to provide a fully working prototype of the components of our framework:

1. a proof-search engine based on linear logic, modified to support inputs from external systems (representing inputs and outputs of the agent)

2. a set of rules which function as a core framework for dialogue management (in the style of KoS (Ginzburg, 2012))

3. several examples which use the above to construct potential applications of the system.

2 Linear logic as a Dialogue Management Framework

Typically, and in particular in the archetypal logic programming language prolog (Bratko, 2001), axioms and rules are expressed within the general framework of first order logic. However, several authors (Dixon et al., 2009; Martens, 2015) have proposed to use linear logic (Girard, 1995) instead. For our purpose, the crucial feature of linear logic is that hypotheses may be used *only once*.

In general, the linear arrow corresponds to *destructive state updates*. Thus, the hypotheses available for proof search correspond to the *state* of the

system. In our application they will correspond to the *information state* of the dialogue participant.

This way, firing a linear rule corresponds to triggering an *action* of an agent, and a complete proof corresponds to a *scenario*, i.e. a sequence of actions, possibly involving action from several agents. However, the information state (typically in the literature and in this paper as well), corresponds to the state of a *single* agent. Thus, a scenario is conceived as a sequence of actions and updates of the information state of a single agent a, even though such actions can be attributed to any other dialogue participant b. (That is, they are a's representation of actions of b.) Scenarios can be realised as a sequence of actual actions and updates. That is, an action can result in sending a message to the outside world (in the form of speech, movement, etc.). Conversely, events happening in the outside world can result in updates of the information state (through a model of the perceptory subsystem).

In our implementation, we treat the information state as a multiset of *linear hypotheses* that can be queried. Because they are linear, these hypotheses can also be removed from the state. In particular, we have a fixed set of rules (they remain available even after being used). Each such rule manipulates a part of the information state (captured by its premisses) and leaves everything else in the state alone.

Our DM models the information-state of only one participant. Regardless, this participant can record its own beliefs about the state of other participants.In general, the core of DM is comprised of a set of linear-logic rules which depend on the domain of application. However, many rules will be domain-independent (such as generic processing of answers). We show these generic rules here, and the demo will illustrate them within an example application.

Integrating moves from NLU and NLG:

$$
\begin{array}{l}
hearAndRemember: \\
\quad (m : DP \to DP \to Move) \to \\
\quad (x\ y : DP) \to (ms : List\ Move) \to \\
\quad Heard\ (m\ x\ y) \multimap \\
\quad Moves\ ms \multimap HasTurn\ x \multimap \\
\quad [_ :: Moves\ (Cons\ (m\ x\ y)\ ms); \\
\quad\ \ _ :: Pending\ (m\ x\ y); \\
\quad\ \ _ :: HasTurn\ y]; \\
utterAndRemember: \\
\quad (m : DP \to DP \to Move) \to \\
\quad (ms : List\ Move) \to (x\ y : DP) \to \\
\quad Agenda\ (m\ x\ y) \multimap Moves\ ms \multimap \\
\quad HasTurn\ x \multimap \\
\quad [_ :: Utter\ (m\ x\ y); \\
\quad\ \ _ :: Moves\ (Cons\ (m\ x\ y)\ ms); \\
\quad\ \ _ :: HasTurn\ y]; \\
pushQUD: \\
\quad (q : Question) \to (qs : List\ Question) \to \\
\quad (x\ y : DP) \to Pending\ (Ask\ q\ x\ y) \multimap \\
\quad QUD\ qs \multimap QUD\ (Cons\ q\ qs)
\end{array}
$$

Basic adjacency:

$$
\begin{array}{l}
counterGreeting: \\
\quad (x\ y : DP) \to HasTurn\ x \multimapinv \\
\quad Pending\ (Greet\ y\ x) \multimap \\
\quad Agenda\ (CounterGreet\ x\ y);
\end{array}
$$

Processing user replies:

$$
\begin{array}{l}
processAssert: \\
\quad (a : Type) \to (x : a) \to (p : Prop) \quad \to \\
\quad (qs : List\ Question) \to (dp\ dp1 : DP) \to \\
\quad Pending\ (Assert\ p\ dp1\ dp) \quad \multimap \\
\quad QUD\ (Cons\ (Question\ dp\ a\ x\ p)\ qs) \multimap \\
\qquad\quad [_ :: UserFact\ p; _ :: QUD\ qs]; \\
processShort: \\
\quad (a : Type) \to (x : a) \to (p : Prop) \quad \to \\
\quad (qs : List\ Question) \to (dp\ dp1 : DP) \to \\
\quad Pending\ (ShortAnswer\ a\ x\ dp1\ dp) \multimap \\
\quad QUD\ (Cons\ (Question\ dp\ a\ x\ p)\ qs) \multimap \\
\qquad\quad [_ :: UserFact\ p; _ :: QUD\ qs];
\end{array}
$$

Answering or clarifying:

$$
\begin{array}{l}
produceAnswer: \\
\quad (a : Type) \to (x : a) \to_! (p : Prop) \to \\
\quad (qs : List\ Question) \to \\
\quad QUD\ (Cons\ (Question\ U\ a\ x\ p)\ qs) \multimap \\
\quad p \multimapinv \\
\quad [\ _ :: Agenda\ (ShortAnswer\ a\ x\ S\ U); \\
\quad\ \ _ :: QUD\ qs; \\
\quad\ \ _ :: Answered\ (Question\ U\ a\ x\ p)]; \\
produceCR: \\
[\ \ a : Type; x : a; p : Prop; \\
\quad qs : List\ Question; \\
\quad _ :: QUD\ (Cons\ (Question\ U\ a\ x\ p)\ qs); \\
\quad _ :: p] \to_? CR;
\end{array}
$$

References

James F Allen, Lenhart K Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato,

Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, et al. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Ivan Bratko. 2001. *Prolog programming for artificial intelligence*. Pearson education.

Lucas Dixon, Alan Smaill, and Tracy Tsang. 2009. Plans, actions and dialogues using linear logic. *Journal of Logic, Language and Information*, 18(2):251–289.

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

J.-Y. Girard. 1995. *Linear Logic: its syntax and semantics*, London Mathematical Society Lecture Note Series, page 1–42. Cambridge University Press.

Kristiina Jokinen. 2009. *Constructive dialogue modelling: Speech interaction and rational agents*, volume 10. John Wiley & Sons.

Staffan Larsson. 2002. *Issue-based dialogue management*. Ph.D. thesis, University of Gothenburg.

Chris Martens. 2015. *Programming Interactive Worlds with Linear Logic*. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

From answer to question: Coherence analysis with Rezonator

John W. DuBois[1,3] **Terry DuBois**[2,3] **Georgio Klironomos**[2,3] **Brady Moore**[2,3]

University of California Santa Barbara[1], University of California Santa Cruz[2], Rezonator[3]
dubois@ucsb.edu, {terry|georgio|brady}@rezonator.com

Abstract

We present Rezonator as a tool for annotating and visualizing the structure of resonance relations between utterances, with the goal of identifying strategies used by conversational co-participants to create coherence in extended dialogue. To go beyond the predictable sequence of question-to-answer (Q2A), we foreground the reverse case: a progression from answer to question (A2Q). Crucially, A2Q sequences often contribute to building longer stretches of cohesion than those created by Q2A alone. Rezonator provides an intuitive graphical UI/UX for human-in-the-loop markup of structural-functional patterns of harmonic resonance. By providing annotators with tools that respond dynamically to their analytical actions, Rezonator seeks not only to efficient and reliable data collection, but help researchers visualize the strategies interlocutors use to collaboratively construct coherence in extended dialogue.

1 Introduction

How do conversational co-participants move from answer to question—coherently? A vast literature has targeted the familiar question-to-answer sequence (Q2A), but the opposite order is effectively invisible in current research. We ask what happens when the order of dialogue acts is (apparently) reversed: What are the consequences for coherence when an answer triggers a question (A2Q)? More generally, what happens when a declarative statement of any kind is followed by a question? In framing the matter in this way, we target an issue which has long challenged state-of-the-art AI and NLP: the hard problem of coherence in extended dialogue (Lai et al. 2018).

While awaiting a solution at the algorithmic level, we present a tool that supports human-in-the-loop annotation and visualization of structural cohesion between successive utterances in naturally occurring conversation. Specifically, we draw on the theory of dialogic syntax (DuBois 2014) to identify the structural patterns and functional strategies that participants use to build coherence. While Rezonator is designed as a general-purpose tool for annotating a broad array of discourse features (including coreference, rhetorical structure, dialogue acts, conversational sequences, etc.), in this paper we specifically target dialogic resonance in A2Q sequences. While analysis of the more familiar Q2A sequences can reveal strong cohesive bonds at the local level of the adjacency pair, with relatively predictable structure, to focus exclusively on such links may miss out on much of the coherence building strategies that users pursue. In contrast, A2Q sequences often appear to contribute to cohesive relationships over longer distances. We illustrate how Rezonator can be used to analyze these patterns, reflecting the overarching goal to analyze structural cohesion relations between utterances to shed light on coherence in extended dialogue.

2 Structural parallelism via resonance

Consider the following conversational exchange, representing a naturally occurring conversational exchange drawn from the Santa Barbara Corpus of Spoken American English (DuBois et al. 2000-2005):

(1) *A Book about Death* SBC005: 190.83-206.63)

128 DARRYL; I come up with my own ideas about that stuff.
129 PAMELA; And where do you get the ideas.
130 DARRYL; Thought.
131 PAMELA; And where do you get those thoughts?
132 DARRYL; Processing what goes on around me.
133 PAMELA; Well?
134 Is n't a book part of what goes on around you?

Figure 1: Stack annotation. A stack is a higher-level discourse unit of any kind, consisting of one or more lines (e.g. multiple intonation units). There is more than one way to stack a discourse, depending on the unit of interest. Here each stack corresponds to a region of maximal internal resonance; each is represented by its own arbitrary but distinctive color. Stacks may be continuous, as in Stack 1 (pink, lines 128-131), or discontinuous, as in Stack 2 (teal, lines 132, 134) and Stack 3 (violet, lines 133, 135). The user can quickly mark a large stack with a single stroke (e.g. dragging over lines 128-131), but also retains the freedom to mark more complex configurations (e.g. clicking on lines one-by-one to collect several of them, discontinuous or not, within a single stack).

Even an excerpt as short as this one can reveal a rich variety of complex discourse phenomena. For example, there are several instances of the Answer-to-Question (A2Q) sequence. Figures 1, 2 and 3 show an analysis of resonance relations in these utterances, as represented in Rezonator. The analysis reveals a high degree of resonance and parallelism in this brief exchange, through which the participants (especially Pamela) create affordances for coherence in extended dialogue.

How to represent such phenomena on the screen has been a driving force behind the design decisions that underlie Rezonator (DuBois 2019), which seeks to present intuitive, efficient, and responsive visualizations that shed light on the linguistic, cognitive, and interactional phenomena that shape the production of spoken dialogue.

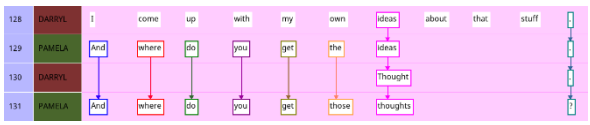


Figure 2: Insistent interrogation as a source of as a self-resonance (self-priming), combining an A2Q move (130→131) and question-to-question parallelism (129→131).

Figure 3: *"Processing what goes on around me"*. In this 2-line excerpt from the conversation in example (1), the diagraph notation shows one strategy for creating coherence, in which the answer to a prior question triggers a new question (A2Q, 132→134), yielding a harmonic resonance array: *what goes on around me : what goes on around you.* Here Rezonator deploys two innovative features designed to support quick capture of the harmonic resonance: (1) So-called right-justification positions the words of each line so that there is vertical alignment of the last word, penultimate word, etc., working backward from the end; (2) With a single stroke, the annotator can capture the alignment of all resonating elements shown above, by dragging the cursor diagonally down from *me* in the first line to *what* in the second line.

Rezonator was designed from the ground up with spoken dialogue in mind, so it is natural that it incorporates structural principles designed to represent the basic units of spoken language, such as intonation units and turns. These units of discourse are accorded conventions which are iconic, familiar, or otherwise readily interpretable. By default, the "unit" displayed on a single line of text is interpreted as representing the intonation unit. But Rezonator is in principle agnostic as to what a line of text is used to represent. It could as well denote a turn-constructional unit, clause. or sentence.

3 Stacks: Discourse beyond the IU

While the intonation unit is surely a key unit of discourse, it is not the only one that matters. In fact, discourse is sometimes defined as language beyond the sentence, which naturally implies the existence of units larger than the intonation unit, too. Figure 1 illustrates the discourse-level of the stack, comprised of more than one intonation unit.

3.1 Parallelism: Left and right periphery

Figures 2 and 3 both illustrate a version of the A2Q conversational move, but each represents a distinct structural pattern of parallelism, or harmonic resonance. In Figure 2, the speaker resonates with herself, reproducing 5 of 7 words within the the exact same syntactic structure, as she insistently interrogates her partner. Note that the region of maximum resonance begins in the left periphery. In contrast, Figure 3 shows a region

of maximum resonance aligned with the right periphery. Rezonator gives the user tools to easily switch views between highlighting the left or the right periphery, depending on which provides the most effective visualization of the discourse strategy at hand.

4 Availability and future development

Plans for future development of Rezonator include the addition of support for annotation of higher discourse-level structures such as Rhetorical Structure Theory, and support for inter-annotator reliability via integration with crowd-source worker platforms such as Mechanical Turk and Prolific.ai.

Rezonator is open source software, distributed under the MIT license. It is currently available in beta form for both PC and Mac at the Rezonator website https://rezonator.com/. Source code and extensive documentation of the project goals are available on GitHub at https://github.com/johnwdubois/rezonator.

5 Conclusions

In this paper we have explored how Rezonator can be used to learn how participants exploit the structural affordances of their language to create resonance for coherence in dialogue. While the familiar Question-to-Answer (Q2A) sequence provides local coherence, a One-Off Question may, due to its very success in achieving early closure, fail to build global coherence. In contrast, when an Answer serves as the basis for a new Question (A2Q), speakers may need to rely more on resonance with the prior dialogue, even as they open up new directions for coherence in the emerging discourse.

To gain a deeper understanding of how participants create ad hoc coherence and engagement in dialogue, researchers need tools that respond dynamically to their analytical decisions, supporting the real-time visualization of structural resonance between utterances. We have presented Rezonator as a tool for annotating and visualizing structural parallelism between adjacent utterances, using it to reveal strategies that interlocutors invoke to create coherence in extended dialogue. We suggest that by providing annotators with a rich, resonance-oriented toolkit that responds dynamically to patterns in the data, as mediated by their analytical choices. By gamifying and crowd-sourcing the collection of such data on a large scale in future research, we hope to show how Rezonator can facilitate efficient, reliable, and reproducible results in the collection of human-in-the-loop training data. Just as important, by creatively implementing visualization principles, Rezonator can help researchers gain insight into the specific structural and functional strategies that interlocutors consistently exploit in their collaborative construction of coherence in extended dialogue.

References

Branigan, H. P., & Pickering, M. J. (2016). An experimental approach to linguistic representation. Behavioral and Brain Sciences, 1-61. doi:10.1017/S0140525X16002028

Cieri, C., Fiumara, J., Liberman, M., Callison-Burch, C., & Wright, J. (2018, May 7-12, 2018). Introducing NIEUW: Novel incentives and workflows for eliciting linguistic data. Paper presented at the Language Resources and Evaluation Conference (LREC 2018), 11th Edition, Miyazaki, May 7-12.

Du Bois, J. W. (2014). Towards a dialogic syntax. Cognitive Linguistics, 25(3), 359–410. doi:10.1515/cog-2014-0024

Du Bois, John W. (2019). Rezonator: Visualizing resonance for coherence in dialogue. Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue. London: SEMDIAL.

Du Bois, John W., Chafe, Wallace L., Meyer, Charles, Thompson, Sandra A., Englebretson, Robert & Martey, Nii (2000-2005). Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitiv3e Sciences, 8(1), 8-11.

Ginzburg, J. (2012). The interactive stance: Meaning for conversation. Oxford: Oxford University Press.

Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., & Gurevych, I. (2017, 19 July 2017). Argotario: Computational argumentation meets serious games. Paper presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language ….

Jurgens, D., & Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. Paper presented at the Transactions of the Association for Computational Linguistics.

Lai, A., & Tetreault, J. R. (2018, July 2018). Discourse coherence in the wild: A dataset, evaluation and methods. Paper presented at the Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue.

Li, J., Monroe, W., Ritter, A., & Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. EMNLP.

Moscoso del Prado Martín, F., & Du Bois, J. W. (2015). Syntactic alignment is an index of affective alignment: An information-theoretical study of natural dialogue. Paper presented at the Proceedings of the 37th Annual Conference of the Cognitive Science Society, San Jose, California.

Osborn, Joseph Carter, Samuel, Benjamin & Mateas, Michael (2018). Visualizing the strategic landscape of arbitrary games. Information Visualization.

Plass, Jan L., Mayer, Richard E. & Homer, Bruce D. (Eds.) (2019). Handbook of game-based learning. Cambridge, MA: MIT Press.

Poesio, M., Chamberlain, J., Paun, S., Yu, J., Uma, A., & Kruschwitz, U. (2019, June 2019). A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. Paper presented at the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis.

Shneiderman, Ben (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings., IEEE Symposium on Visual Languages. Harvard: IEEE. 336-343.

Tufte, Edward R. (2006). Beautiful evidence. Graphics Press.

Ware, Colin (2013). Information visualization: Perception for design. Elsevier.