SEMDIAL 2016

JerSem

Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue

Julie Hunter, Mandy Simons, and Matthew Stone (eds.)
New Brunswick, NJ, 16–18 July 2016

ISSN 2308-2275
Serial title: Proceedings (SemDial)

SemDial Workshop Series

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

JerSem Website

http://semantics.rutgers.edu/jersem/

Preface

JerSem marks the twentieth year of the annual SemDial Workshop on the Semantics and Pragmatics of Dialogue! We are delighted to hold the meeting in New Brunswick, New Jersey, and to follow up 2011's Los Angelogue by bringing the SemDial meeting to North America for only the second time. This year, we have collocated SemDial with the North American Summer School in Logic, Language and Information (http://nasslli2016.rutgers.edu/). Both events are known for bringing together disciplines from across cognitive science and we hope the opportunity for extended participation in both events strengthens this cross-fertilization. To capitalize on overlaps with NASSLLI on questions under discussion and the problems of dialogue, we organized the first day of SemDial as a special session on "Questions under Discussion", to focus on the role of discourse purposes in utterance interpretation and dialogue structure, and their reflection in utterance form.

We received a total of twenty full paper submissions. Ten of those papers will be presented at JerSem, after selection based on a round of anonymous peer review that secured written evaluations from three experts on each submission. We are extremely grateful to the Program Committee members for their very detailed and helpful reviews. The poster session hosts seven additional contributions that came in response to a call for late-breaking posters and demonstrations. All accepted full papers and poster abstracts are included in this volume. We are pleased that the mix of papers continues to reflect the diverse range of methods available to dialogue research, including experimental studies, corpus studies, and formal and computational models.

The JerSem program features four keynote presentations, by Jonathan Ginzburg, Kordula de Kuthy, Nigel Ward and Elisabeth Camp (in order of appearance). We thank them for participating in SemDial and are honored to have them at the workshop. Abstracts of their contributions are also included in this volume.

JerSem has received generous financial support from the Rutgers Center for Cognitive Science (http://ruccs.rutgers.edu). Partial funding for the Special Session "Questions Under Discussion" has been provided by NSF support of project number 1452674 "What's the question? A cross-linguistic investigation into compositional and pragmatic constraints on the question under discussion." We are very grateful for this sponsorship. We have also been given the endorsement of SIGdial, the special interest group on discourse and dialogue of the Association for Computational Linguistics and the International Speech Communication Association.

Last but not least we would like to thank everyone else who helped with Page i of 5the organisation, particularly acting RuCCS directors Ernie Lepore and Gretchen Chapman, RuCCs staff members Sue Cosentino and Jo'Ann Meli, and our student helpers.

Julie Hunter, Mandy Simons and Matthew Stone
New Brunswick, NJ
July 2016

Program Committee

Julie Hunter (co-chair), University of Toulouse III - Paul Sabatier
Mandy Simons (co-chair for special session on QUD), Carnegie Mellon University
Matthew Stone (co-chair), Rutgers University

Nicholas Asher, Université Paul Sabatier
David Beaver, University of Texas
Claire Beyssade, Université de Paris 8
Ellen Breitholtz, University of Gothenburg
Eve Clark, Stanford University
Judith Degen, Stanford University
Valeria de Paiva, Nuance
Paul Dekker, University of Amsterdam
David DeVault, University of Southern California
Simon Dobnik, University of Gothenburg
Raquel Fernandez, University of Amsterdam
Kallirroi Georgila, University of Southern California
Jonathan Ginzburg, Université Paris-Diderot
Eleni Gregoromichelaki, King's College London
Pat Healey, Queen Mary University London
Julian Hough, University of Bielefeld
Chris Howes, University of Gothenburg
Amy Isard, University of Edinburgh
Andrew Kehler, University of California San Diego
Ruth Kempson, King's College London
Staffan Larsson, University of Gothenburg
Alex Lascarides, University of Edinburgh
Oliver Lemon, Heriot-Watt University
Pierre Lison, University of Oslo
Marie-Catherine de Marneffe, The Ohio State University
Gregory Mills, University of Groningen
Chris Potts, Stanford University
Matthew Purver, Queen Mary University London
Kyle Rawlins, Johns Hopkins University
Hannes Reiser, University of Bielefeld
Craige Roberts, The Ohio State University
David Schlangen, University of Bielefeld
Amanda Stent, Yahoo!
Judith Tonhauser, The Ohio State University
David Traum, University of Southern California

Contents

Invited Speakers

QUD: Past, Present, and Future
Jonathan Ginzburg 1

Annotating Questions under Discussions in Authentic Data
Kordula De Kuthy 2

Prosody, Action, and Coordination in Real-time Gameplay
Nigel G. Ward 3

Sarcasm in Conversational Action
Elisabeth Camp 4

Full Papers

Why?
Julian J. Schlöder, Ellen Breitholtz, and Raquel Fernández 5

Joint Information Structure and Discourse Structure Analysis in an Underspecified DRT Framework
Uwe Reyle and Arndt Riester 15

A Model for Attention-Driven Judgements in Type Theory with Records
Simon Dobnik and John D. Kelleher 25

Grounding, Justification, Adaptation: Towards Machines That Mean What They Say
David Schlangen 35

Comparing Dialogue Strategies for Learning Grounded Language from Human Tutors
Yanchao Yu, Oliver Lemon, and Arash Eshghi 44

Evaluating Conversational Success: Weighted Message Exchange Games
Nicholas Asher and Soumya Paul 55

Disagreement with Evidentials: A Call for Subjectivity
Natalia Korotkova 65

Prominence Shifts in English and Spanish Parallel Constructions
Jeffrey Klassen, Michael Wagner, Annie Tremblay, and Heather Goad 76

Impotent Speech Acts and Awareness
Phil Crone 86

Multi-layered analysis of laughter
Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg 97

Short Papers

Extralinguistic State Localization in Service of Turn Generation in Task-Oriented Dialogue
Petr Babkin 108

Structural focus and discourse structure in Hungarian narratives
Kata Balogh 110

Are you mocking me or are you laughing with me?
Ellen Breitholtz and Kristina Lundholm Fors 112

PentoRob: A Puzzle-Playing Robot for Dialogue Experiments
Julian Hough and David Schlangen 114

Generating Surplus Content in a Q/A-Setting
Sebastian Reuße, Ralf Klabunde, Jon Stevens, and Anton Benz...............................116

Aligning Intentions: Acceptance and Rejection in Dialogue
Julian J. Schlöder, Antoine Venant, and Nicholas Asher 118

An Incremental Dialogue System for Learning Visually Grounded Language (demonstration system)
Yanchao Yu, Arash Eshghi, and Oliver Lemon..120

QUD: Past, Present, and Future

Jonathan Ginzburg
CLILLAC-ARP (EA 3967) & Laboratoire d'Excellence (LabEx)—EFL
Université Paris-Diderot, Paris, France
`yonatan.ginzburg@univ-paris-diderot.fr`

In this talk I will start by considering some past motivations for positing QUD, a repository of questions that conversational participants exploit in, arguably, just about any form of interaction. A number of distinct QUD theories are possible, with parameters including how shared its elements are and the nature of its ordering. I will show how QUD enables a theory of interaction to accommodate and sharpen insights concerning domain dependence (from AI) and other-repair (from conversational analysis), and indeed to significantly change our view of grammar by integrating self-repair. I will conclude by discussing how appraisal theories utilised in cognitive theories of emotion can be integrated in interaction theories, along with more speculative comments on gesture and music.

Annotating Questions under Discussions in Authentic Data

Kordula De Kuthy
University of Tübingen
kdk@sfs.uni-tuebingen.de

The information structure of sentences is receiving increased interest in linguistics as the attention has shifted from the analysis of isolated sentences to the question how information is packaged in sentences analyzed in context. In order to connect the information structure of sentences to the overall structure of the discourse, an analysis in terms of *Questions under Discussion* is proving to be a useful tool.

According to Roberts' (2012) account, natural discourse in general serves to answer hierarchically ordered Questions under Discussion (QUDs). These implicit QUDs can be used to account for the information structure of utterances in context: the part of a sentence contained in the formulation of the current question is called the *background*, while the part which provides the actual answer is the *focus*.

The notion of implicit QUDs has also been referenced in corpus-based research attempting to analyze the information structure of naturally occurring, authentic data (e.g., Ritz et al., 2008; Calhoun et al., 2010). Yet these approaches were only rewarded with limited success in terms of achieving agreemen, arguably because the task of identifying QUDs was not made explicit.

In this talk reporting joint work with Arndt Riester, we introduce our methodology for a combined analysis of data in terms of both discourse and information structure, integrating an explicitly spelled out notion of QUDs. We identify the necessary steps of an analysis procedure based on QUDs and demonstrate the method on authentic data taken from a German spoken-language corpus. We formulate pragmatic principles that allow us to analyze the discourse structure, formulate adequate QUDs, and analyze the information structure of individual utterances in the discourse. Based on the authenic data analysis, we illustrate that the formulation of QUDs can be successfully guided by the formulated principle and that QUDs play a crucial role in accounting for discourse structural configurations. At the same time, they also provide an objective means to determine the information structure, including both the focus-background divide as well as not-at-issue content.

References

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. 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*, 387–419.

Ritz, J., S. Dipper, and M. Götze (2008). Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, pp. 2137–2142.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics 5*(6), 1–69.

Prosody, Action, and Coordination in Real-time Gameplay

Nigel G. Ward
Department of Computer Science
University of Texas El Paso
nigel@utep.edu

Language is invaluable for real-time coordination of action. We are studying this in a corpus of fast-paced games in which pairs of players cooperate to solve a maze with obstacles and puzzles. The players routinely perform astounding feats of communication, rapidly producing utterances which simultaneously convey multiple dimensions of semantic and pragmatic information, and which are adapted continuously as the game state and dialog state change. To investigate, we applied unsupervised methods to discover the most common patterns, considering both prosodic features and game-action features. We found superimposable behavior patterns that involve both language acts and domain actions, and that are comprised of synchronized contributions by both players. These phenomena and patterns pose challenges for many current theories, models, and technologies.

Sarcasm in Conversational Action

Elisabeth Camp
Department of Philosophy
Rutgers, the State University of New Jersey
`elisabeth.camp@rutgers.edu`

Sarcasm is a paradigmatically pragmatic phenomenon: the most obvious case of meaning something other than what you say. It is also a pragmatically important and puzzlingly complex phenomenon. It is a key tool for joking, nudging and shaming one another into conversational alignment, but it often seems to underwrite a kind of communicative bluff: it enables speakers to make a conversational move without shouldering commensurate conversational liability. Further, its communicative effects range from stiletto-sharp clarity ('You sure know a lot.') to infuriatingly amorphous evocation ('How old did you say you were?'). I argue that we can analyze a wide range of sarcastic utterances in traditional terms of 'meaning inversion' if we take a broader view of meaning than philosophers and linguists have traditionally done. Different species of sarcasm take different aspects of meaning as targets for inversion, producing systematically distinct conversational effects.

Why?

Julian J. Schlöder[†], **Ellen Breitholtz**[‡] and **Raquel Fernández**[†]
[†]Institute for Logic, Language and Computation
University of Amsterdam
[‡]Department of Philosophy, Linguistics and Theory of Science
Gothenburg University
`julian.schloeder@gmail.com, ellen@ling.gu.se, raquel.fernandez@uva.nl`

Abstract

Even casual dialogue contains instances of reasoning. A paradigmatic case is the usage of *Why*-questions that intuitively elicit a *reason* for something. We present a thorough analysis of *Why*-questions in dialogue from a rhetorical perspective. We specify the semantics of *Why*-questions, *i.e.*, we define what the space of acceptable answers is, how this acceptability is itself up for further negotiation, and discuss some context-sensitive aspects of bare *Why?*. We formalise our model in a type-theoretical framework.

1 Introduction

Participating in a dialogue requires the interlocutors to *reason* about certain propositions and circumstances. On one hand, interlocutors are generally expected to back up the assertions they make with arguments, should this be required. On the other hand, the notion of *relevance* of an utterance is linked to reasoning: a relevant utterance is made for a reason, *e.g.*, to provide or inquire about information pertinent to the purpose of the dialogue.

Such reasons are not always explicated by the interlocutors, but can be elicited by clarification questions (Jackson and Jacobs, 1980; Breitholtz, 2010; Schlöder and Fernández, 2015). The paradigmatic examples are *Why*-questions. We are interested in what constitutes the space of possible answers to such questions and how they are interpreted in a discourse. The following examples retrieved from the British National Corpus (BNC) (Burnard, 2000) exemplify the basic phenomenon.

(1) a. B: He's in hospital.
b. C: Why?
c. B: Because he's not very well
(BNC, file KBF, lines 3394–3396)

(2) a. G: Do you want mum to come to Argos with me tomorrow morning?
(three lines omitted)
b. R: Why are you asking me?
c. G: Cos you said you'd come to Argos with me. (BNC, file KC8, lines 191–196)

In (1), B makes an assertion and C asks for a reason that backs the truth of the proposition expressed in (1a); note that this need not entail that C is doubting the content of B's assertion. We contrast this with what is happening in (2). There, G asks a question and R inquires about G's reason for doing so. In both cases, the initial speaker then supplies a *reason* that is marked with the particle 'because'. Here, we use the concept 'reason' intuitively—only if we can *define* what makes a reason, we can define what makes an answer to a *Why*-question.

A first observation is that the arguments expressed by the first and third utterances in (1) and (2) are logically incomplete: they indicate that the third utterance *is* a reason for the first, but not what warrants the inference. In classical rhetoric, such arguments are called *enthymemes*. An enthymeme is an argument of the form 'p hence q' which requires the listener to supply one or more underpinning premises. It has been observed that enthymematic reasoning is widespread in natural dialogue, and has been linked to clarification and cognitive load management (Jackson and Jacobs, 1980; Breitholtz and Villing, 2008). Therefore, we will analyse different types of *Why*-questions in terms of enthymematic reasoning to find out what the correlation is between rhetorical structure and different types of *Why*-questions.

The paper is structured as follows: In the next section, we will give an overview of existing work in discourse modelling related to reasoning, *Why*-questions and enthymemes. Afterwards, in section 3 we will further elucidate the dynamics of enthymematic reasoning with natural dialogue ex-

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 5-14.

amples. We will describe a formal treatment of our analysis in section 4.

2 Reasoning in Dialogue

Many conversational phenomena like disagreement, misunderstanding, and clarification can be linked to enthymematic reasoning (Breitholtz, 2014a). Consider the example in (3).

(3) A: Let's walk along Walnut Street
It's shorter (cited from Walker (1996))

This excerpt is uttered in the context of two colleagues on their way to work, where several routes are possible. Speaker A suggests to take one of them and provides a reason supporting this. The two propositions convey an enthymeme: an argument that relies on generally recognised facts and notions regarding how it is acceptable to reason. Enthymemes consist of two parts, a premise and a conclusion, as in the case of our example:[1]

(4) $\dfrac{\text{It (Walnut Street) is shorter}}{\therefore \text{Let's walk along Walnut Street}}$

In this case the speaker counts on the interlocutor being able to supply something that underpins (3). That is, something that warrants its interpretation as an argument while simultaneously validating it.

These kinds of underpinnings are often referred to as *topoi* in the literature on rhetoric and argumentation. Some topoi may be applied to various subjects, while others are specific to a particular subject. Ducrot (1980; 1988) and Anscombre (1995) talk about topoi as links between propositions that are necessary for the propositions to cohere in discourse. A topos that could be drawn upon to validate the argument in (3) could be something like *'if a route is shorter (than other options), choose that route'*.

We refer to the topoi that are available to an individual as that individual's *rhetorical resources*. On this view, speakers have access to a vast set of topoi which to a great extent mirrors the experiences they have had. Another important aspect of this view is that the topoi accessible to one individual do not constitute a monolithic logical system. In contrast to, for example, a representation of world knowledge, a set of topoi may contain contradictions or principles of inference which lead to contradictions.

These phenomena have also been discussed from the perspective of *discourse relations*; most notably in the SDRT framework (Asher and Lascarides, 2003). SDRT includes the discourse relations *Explanation*(α, β) and *Result*(β, α). These relations are assigned to α and β only if it is *true* in the underlying world model that β can be a cause for α. Because these inferences are done in a defeasible logic, SDRT can also account for the fact that sometimes β does not explain α in spite of 'β, hence α' being a valid form of inference.

SDRT also includes meta-discursive versions of these relations. These model the fact that sometimes speakers give reasons for *making* certain speech acts, as *e.g.*, in example (3) where the speaker gives a reason for making a suggestion. A relation particularly interesting to us is *Q-Elab*(α, β) that applies when β asks a question pertinent to the goal that the speaker of α wants to achieve by uttering α. To our understanding, *Why*-questions broadly fall under this umbrella, but no such account of *Why?* has yet been elaborated.

We prefer the rhetorical approach over the discourse relations model for the following reason. As our analysis will show, inference patterns are dynamic in that they can be presupposed, accommodated, elicited and themselves be discussed. The SDRT account, as far as we understand it, is not amenable to such flexibility. In particular, the semantics of *Q-Elab*(α, β) requires that the space of possible answers to β is fixed and known (Asher and Lascarides, 2003, Sec. 9.3.3). If β is a *Why*-question, we do not believe this to be the case. This is because the answer set to a *Why*-question depends on the available topoi. Since topoi are dynamic, so must be these answer sets.

An important consequence of this, as we see it, is that the acceptability of a given reason does not depend on an inferential relationship being *correct* (in some objective sense, *e.g.*, in a model), but merely on it being subjectively *acceptable* to the interlocutors. Acceptability, in turn, depends on the rhetorical resources of individual speakers.[2]

3 Analysing Reasons

We now describe how we model what counts as a *reason*, *i.e.*, what counts as an answer to a *Why*-

---

[1] This distinguishes them from logical arguments or syllogisms which typically have *three* parts: A premise, a conclusion, and a rule sanctioning the inference.

[2] This also means that our interest in *Why*-questions differs from analyses that seek to elucidate what *explanations* are in philosophy of science (Bromberger, 1992; van Fraassen, 1980). Our *reasons* are dialogical phenomena, whereas their *explanations* are, roughly, about natural or physical laws.

question. We then discuss by way of examples how these questions are used and answered in dialogue and how they contribute to grounding. Then, we summarise our findings and present some interesting cases that fall outside our analysis.

3.1 Reasons

Certainly, the answers given by B and G in our initial examples (1) and (2) are not arbitrary. Not *any* utterance would be an acceptable answer to the *Why*-questions in these examples. Similarly, not every utterance that expresses 'p because q' is immediately acceptable to its addressee. We stipulate that q is a *reason* for p if there is a topos that validates the enthymeme $q \therefore p$. Stating that 'p because q', 'if q, then p' or answering 'q' to 'Why p?' expresses that q is a reason for p. Hence, such utterances presuppose that there is such a topos. Thus, addressees can either retrieve an appropriate topos from their set of rhetorical resources or infer and accommodate a new one.

The following examples provide evidence for this conception.

(5) a. J: I roasted it and we couldn't eat it on the Sunday and
 b. A: Could not? Why could you not eat it?
 c. J: That was bull beef.
 d. A: Oh right.
 e. H: our second class beef, you see.
 f. J: Then I, I put it in a saucepan and I stewed it the next day
 (BNC, file K65, lines 284–299; some backchannel utterances omitted)

In (5c), J gives an answer to a *Why*-question, *i.e.*, J gives what she construes to be a reason for *'being unable to eat the roast'*. Speaker A indicates that he accepts this as an answer, but H still elaborates in (5e). The addtional information in (5e,f) suggests the following enthymeme:

(6) x is bull beef
 ――――――――――――――――
 ∴ J could not eat roasted x
 Topos: *one cannot roast bull beef*
 (but ought to stew it)

This dialogue offers evidence for our claim that what *makes* (5c) an answer to (5b) is the more general statement indicated in (5e,f), *i.e.*, the topos of (6). To an interlocutor that is unaware of this information, answering (5c) to (5b) would seem like a *non sequitur*. The following example is an explicit case in point. The second speaker explicitly mentions the principle that he takes to back the conditional statement in (7a).[3]

(7) a. D: I'm self-funding my campaign, I tell the truth.
 b. J: 'I'm rich, therefore I tell the truth' has [. . .] no cause and effect between the two.
 (from *Last Week Tonight*, Feb. 29th, 2016)

The explication of the topos in (7b) suggests to us that J has interpreted D's utterance as (8).[4]

(8) D self-funds his campaign
 ――――――――――――――――
 ∴ D tells the truth
 Topos: *rich people tell the truth*

3.2 Contextual dependence of *Why*-questions

Based on this definition of what makes a reason, we now look into the context of *Why*-questions. We propose that the reasons elicited by these questions are dependent on (i) the current issue under discussion and (ii) the form of the question itself, *i.e.*, its sentential or non-sentential character. The elicited reasons can be either factive ('why p?', given some proposition p under discussion) or meta-discursive ('why are you saying p / asking q / suggesting r?', given some salient dialogue act).[5] While any type of reason can be queried with a sentential *Why*-question, only a restricted set of possible reasons can be elicited by bare *Why?*.

3.2.1 Factive reasons

In contexts where the current issue under discussion has arisen from an asserted proposition, *Why*-questions typically ask for a reason justifying the asserted content. For instance, example (5b) is a sentential form of such a *factive Why*-question and example (1b) from the Introduction is a bare factive *Why?*. We consider the following to be an appropriate reading of (1a,c).

---

[3] This example is from the TV show *Last Week Tonight*. (7a) is an excerpt from a speech by Donald Trump; (7b) is John Oliver's commentary.

[4] Though (8) is not the *only* possible interpretation of (7a).

[5] We call *Why*-questions 'factive' if they inquire about a *claim* and contrast them with those inquiring about an *act*. We do not claim that such questions are factive in the sense of factive verbs like *know*. One can pose a factive *Why*-question without presupposing the truth of the claim, *e.g.*, *'Why would this be true?'*. Some prior work, *e.g.*, Hempel (1965) or Hintikka and Halonen (1995), claims that a *Why*-question carries its core proposition as a presupposition (sometimes, *e.g.*, Bromberger (1992), with the restriction that the content is in indicative mood, excluding '*Why would...*' cases). The observations we make seem to cast doubt on this. Apparently, one can ask '*Why* p?' without accepting p. An example is (25) below; see our discussion there.

(9) $$\frac{x \text{ is not very well}}{\therefore\ x \text{ is in the hospital}}$$
Topos: *unwell people go to the hospital*

An interesting special case arises when the issue under discussion stems from a conditional statement. Then, *Why?* elicits a backing for an already stated premise–conclusion pair (*i.e.*, it elicits a reason for the enthymeme *itself*). Simply put, asking for the grounds of a conditional statement elicits an underpinning premise, but the application of this premise is *itself* enthymematic.

(10) a. D: If you feel cold you'd be dead.
b. C: Why?
c. D: You just are. Part of you being alive is that you're warm.
(BNC, file KBW, lines 11065–11068)

The utterance (10a) states an inferential relationship without giving grounds for the relationship. The *Why?* in (10b) asks for a reason for this relationship, *i.e.*, for the premise in (11).

(11) $$\frac{-}{\therefore\ (x \text{ is cold} \therefore x \text{ is dead})}$$

Put differently, (10b) asks for a reason why *x is cold* is a reason for *x is dead.* Thus, we represent the content of (10a) as the enthymeme in (12) and (10b) as asking about the *topos* of (12).

(12) $$\frac{x \text{ is cold}}{\therefore\ x \text{ is dead}}$$
Topos: –

Then, the utterance (10c) supplies such a topos, so the *nested* enthymeme in (13) is a representation of what is under discussion after (10c).

(13) $$\frac{\text{Living things are warm.}}{\therefore\ \dfrac{x \text{ is cold}}{\therefore\ x \text{ is dead}}}$$
Topos: *contraposition.*

This in particular serves to illustrate the fact that enthymemes can be nested: In principle, this situation would now license the elicitation of a backing to support the enthymeme in (13) again (and so on). Already Lewis Carroll (1895) observed that one can always ask for what licenses an inference, then ask for what licenses the license *etc. ad infinitum*. Therefore, an adequate model needs to *always* assume that there is a topos in the context that the interlocutors do not explicate, but implicitly accommodate. The difference between (12) and (13) is that the topos implicit in (12) is explicated in (13)—but the explication again presupposes a new implicit inference pattern.

Similarly, in the next example, the asker of the *Why*-question is *not* able to accommodate the answer as a reason. So he questions the relevance of the answer with *So what?* (we will further discuss *So what?* in the next subsection).

(14) a. P: I was with Nanna and Adrian.
b. R: No Daddy said you should be with Michelle and Mutty.
c. P: ⟨unclear⟩⟨pause⟩ Why do I ⟨unclear⟩ Nan and Adrian?
d. R: Well cos ⟨pause⟩ erm some of the questions are sort of, English questions.
e. P: So what?
f. R: Well Michelle's not English.
(BNC, file,KD0, lines 3624–3629)

The dialogue (14) is about assigning groups in some (not further specified) game. In (14e), speaker P indicates that he does not see what makes (14d) an answer to (14c). Then, in (14f), R supplies an additional premise that supports the following nested enthymeme.

(15) $$\frac{\text{Michelle is not English}}{\therefore\ \dfrac{\text{there are some English questions}}{\therefore\ \text{P should be with Michelle and Mutty}}}$$
Topos: *Non-English people need help with English questions*

The enthymeme in (15) particularly exemplifies the notion described above: Once elicited, a backing becomes a premise in a superordinate enthymeme that again requires an implicit topos to be interpreted.

3.2.2 Meta-discursive reasons

The utterance (2b) in the Introduction is an example of a *Why*-question asking for a reason justifying a *linguistic* fact. Such a meta-discursive interpretation is the only one available to bare *Why?* when the active issue under discussion does not stem from asserted content. In example (16) the active issue is a question and in (18) it is a suggestion.[6] Here, rather than prompting a reason to justify a contextually provided proposition, the *Why?*'s can be glossed as *'Why are you saying this?'*. The answers in (16b) and (18b) raise the enthymemes in (17) and (19), respectively.

---

[6] (16) is between a child (A) and its minder (B). A wants something, and B is wise to A's attempt at manipulation.

(16) a. A: Do you love me ⟨unclear⟩?
b. B: Why?
c. A: ⟨unclear⟩ I love you so much.
(BNC, file KCM, lines 1057–1060)

(17) A loves B
∴ A wants to know if B loves A
Topos: *one wants to know if love is requited*

(18) a. D: Oh I should keep the strawberries if I were you.
b. C: Why?
c. D: Strawberries are delicious.
(BNC, file KBW, lines 9848–9850)

(19) D thinks strawberries are delicious
∴ D suggests to keep strawberries
Topos: *one should keep delicious things.*

When replying to assertions, bare *Why?* does not have this effect, as it is interpreted to ask for a reason for the asserted proposition being factual, as in (9). Instead, a sentential *Why*-question is needed to elicit a meta-discursive reason, *e.g.*, *'Why are you telling me this?'*. Interestingly, meta-discursive reasons can be queried in these cases with non-sentential *So (what)?*, as exemplified below:

(20) a. C: Who are you going to snog on Saturday? (two lines omitted)
b. K: I don't know.
c. C: Snog Phil.
d. K: No I've done him already ⟨laugh⟩.
e. C: So?
f. K: done it, been there, got the T-shirt.
(BNC, file KPH, lines 1582–1588)

In (20e), C questions the relevance of *'having done him already'* to the issue of *'not snogging Phil'*. We model this as the enthymeme in (21): C recognises that K is giving a reason for her rejection of the proposal in (20c), but cannot supply or infer a topos to validate the inference. The topos K supplies (by conventional implicature) in (20f) seems to be *'repeated experiences are boring'*.[7]

(21) K has snogged Phil already
∴ K will not snog Phil
Topos: –

A *Why?* in place of (20e) would ask for a reason why K has already snogged Phil, *i.e.*, it would ask for the missing premise in (22) (like in 1b).

(22) –
∴ K has snogged Phil already

So (what)? is meta-discursive in particular when replying to an answer to an earlier question. That is, asking *'Why are you saying this?'* of an answer is asking *'How does this answer my question?'*. This explains the function of *'So what?'* in (14).

3.3 Reasons and grounding

As mentioned before, sometimes *Why*-questions function as clarification questions. We draw the conclusion that the dynamics of reasons we just discussed can be related to the grounding process. We begin by observing that sometimes the rejection of a premise in an enthymeme can leave the *conclusion* ungrounded, *i.e.*, not mutually accepted by the interlocutors. The dialogue in (23) is a case in point.

(23) a. M: You're not having bacon till Monday. ⟨pause⟩ (three lines omitted)
b. M: You're working, so you don't need bacon.
c. J: I'm not working Monday.
d. M: Well you can go and get it.
(BNC, file KCL, lines 405–411)

Here, M makes a proposal in (23a) and backs it with the enthymeme (24) in (23b).[8] J in (23c) denies the premise of (24). M in (23d) concedes that therefore the conclusion (23a) is defeated.

(24) J is working on Monday
∴ J does not need bacon on Monday

Also, loosely following the distinction between *intention recognition* and *intention adoption* of Schlöder and Fernández (2015), we observe that one can *recognise* a topos that validates an enthymeme without *accepting* the topos as valid (*i.e.*, without adopting the topos in one's private set of available topoi). This is shown in example (7), where J cites a topos that would support the enthymeme, but denies that it is valid.

With these preliminaries in place, we can consider an example where a *Why?* is asked before accepting an assertion.

---

[7] A variety of online dictionaries (Urban Dictionary, Wiktionary, and The Free Dictionary) agree that *'been there, done that, got the T-shirt'* conventionally means that the speaker is familiar with an activity to the point of boredom.

[8] The topos licensing the enthymeme is not clear to us.

(25) a. C: Got the junior tap and the senior tap.
b. B: Yeah but you'll get that next year again.
c. C: Why?
d. B: Because you got honours didn't you? In grade three. ⟨pause⟩
e. C: No cos junior tap was for grade three.
f. B: Have you done grade four tap?
(BNC, file KBF, lines 12258–12264)

We analyse this as follows. B makes an assertion in (25b) that is not immediately acceptable to C, so she asks for a reason in (25c). B supplies a reason in (25d), completing the enthymeme in (26).

(26) $$\frac{\text{C got grade three honors}}{\therefore \text{ C will get junior and senior tap}}$$

Then, in (25e), C denies that this is a valid inference. Apparently B concedes this: instead of arguing the point of (26) she is looking for a *different* premise that would allow her to infer the conclusion of (26). This evinces that the proposition asserted in (25b) is still not accepted by C, *i.e.*, it is left ungrounded.

3.4 Summary of findings

Based on the evidence analysed in the preceding subsections, we summarise our findings on the dialogue dynamics of *Why*-questions as follows. We also include the question *So (what)?*, which, as we have seen, serves to elicit reasons not available to bare *Why?*. Our (informal) model goes like this:

(i) *Why*-questions, including bare *Why?*, can have factive and meta-discursive readings.

(ii) The availability of these readings depends on context. In the case of a propositional antecedent, the meta-discursive reading is not immediately available to bare *Why?*, but it can instead be obtained with *So (what)?*

(iii) A *reason*, *i.e.*, an acceptable answer to a *Why*-question, is a proposition that connects enthymematically to the question's antecedent.

(iv) In interpreting such an answer, the listener can either apply an available topos, accommodate the presupposition that there *is* such a topos, or elicit another tacit premise. The last case can again be modelled as asking for a *reason* for why the enthymeme *itself* is valid.

(v) To understand an enthymeme—or that something is given as a reason—it is not required to consider the underpinning topos valid.

3.5 Special cases

Our main interest in this paper is the elicitation, interpretation, and accommodation of *reasons* as a dialogical phenomenon. We note that while the interpretation of bare *Why?* is of interest to us, we cannot claim to model the phenomenon exhaustively. A particularly striking example is Ginzburg's much discussed *turn-taking puzzle* (Ginzburg, 2012, Ex. 23, here as 27).

(27) A: Which members of the audience own a parakeet?
a. A: Why? [Why own a parakeet?]
b. B: Why? [Why are you asking?]
c. A: Why am I asking this question?

Our account of what it means to give a reason, *i.e.*, to answer a *Why*-question, straightforwardly accounts for all three cases in (27), but our informal discussion of bare *Why?* only accounts for (27b). As (27c) shows, the meta-discursive reading is available in the context of (27a), but, still, the bare *Why?* there has a factive reading. Modelling these differences would require a more sophisticated analysis of what is under discussion than we can provide here.

We used SCoRE (Purver, 2001) to systematically search for further counterexamples.[9] The following two examples show further functions of *Why* that our analysis does not cover.

(28) a. D: You know why they can't put more carriages on a train?
b. G: Why? (BNC, file KCA, lines 1912–1913)

(29) a. U: Andy, do you want a cup of tea?
b. A: Er er, yeah. Cheers.
c. M: Do you want one Nick?
d. N: Why not? (BNC, file KPR, lines 95–99)

The *Why?* of (28b) is a reprise fragment of its antecedent and cannot be glossed as *'Why are you asking?'*. This example indicates to us that the interpretation of bare *Why?* is at least sometimes elliptical. An elliptical account of *Why?* would also serve to disentangle the turn-taking puzzle (27). Such an account would be complementary to our discussion in that it would help to determine the proposition that a *Why*-question is about.[10] From

---

[9] According to our search, the dialogue section of the BNC contains 2256 *Why*-questions, 858 (38%) of these bare *Why?*. We manually surveyed a random selection of about 200.

[10] It seems possible that a grammar for elliptical *Why?* can also consider *'are you asking?'* as elliptical content and thereby predict meta-discursive readings as well. It seems unlikely, however, that *'so (what)?'* is elliptical (we thank an anonymous reviewer for pointing this out to us).

that point onward, our account of what makes a reason would apply. In addition, the antecedent (28a) is also an *embedded Why*-question that does not prompt G to provide a reason. Here, we need to leave the embedding behaviour of *Why*-questions, elliptical interpretations and the relation to reprise fragments to further work.

In (29d), the speaker N seems to use *Why not?* to indicate that he would like some tea, *i.e.*, as an agreement move.[11] We believe that this function of *Why not?* is related to the function of *Why?* as a clarification question (see subsection 3.3). In the account of clarifying *Why?* of Schlöder and Fernández (2014), an addressee is assumed to accept a proposal if they have no reason not to. Hence, we interpret *Why not?* in contexts like (29) to mean that the speaker cannot think of a reason not to. Thereby, it implicates acceptance. A strikingly explicit example for this is (30).

(30) A: Do you agree with that?
G: I have no reason to disagree. Yes.
(BNC, file FMN, lines 492–493)

We note however that *Why not?* also can have the factive function we discussed in subsection 3.2.1 as long as its antecedent has negative polarity. Example (31) is a typical case.

(31) T: I'm not going to sleep.
C: Why not? (BNC, file KBH, lines 4408–4409)

4 Formal Modelling

In this section we will use a Dialogue Game Board (DGB) semantics cast in Type Theory with Records (TTR) to formalise the notions discussed in the previous section. We will take as our point of departure the model for analysing rhetorical reasoning in dialogue developed by Breitholtz (2014a). This account of enthymematic reasoning builds on the formal work on dialogue modelling by Cooper and Ginzburg (2012; 2015). The leading idea of this approach is that a theory of dialogue should be cognitively plausible as well as computationally feasible. TTR is put forward as a framework that is just that.[12] We intend to show that our observations are structured and precise enough to be embedded in such a framework and be integrated in a well-developed dialogue semantics. Since the rhetorical model we employ makes frequent reference to both cognition and computation, *e.g.*, when it comes down to the availabilty and retrieval of certain topoi by individual interlocutors, the TTR framework seems appropriate.[13]

The semantics of Cooper and Ginzburg models the information states ('game boards') of individual speakers and their changes as the dialogue progresses. A full dialogue semantics, *e.g.*, KoS (Ginzburg, 2012), might make use of a large set of features in these information states. Here we will only explicitly mention a minimal subset that is sufficient to model enthymematic reasoning. A gameboard is modelled as a *record type*, *i.e.*, a structured type featuring multiple labelled fields of certain types; 'l : T' expresses that whatever is associated with label 'l' ought to be of type T (Cooper, 2005). A DGB has two major fields called 'shared' and 'private'. Shared information is information which is in some way necessary for a dialogue contribution to be interpreted in a relevant way. This includes 'moves', the list of *moves* in the dialogue, , 'l-m', the *Latest Move*, and 'qud', the *questions under discussion*.

Breitholtz (2014a) adds two addtional 'shared' fields: 'eud', *enthymemes under discussion* and 'topoi', a *list of topoi* required to interpret the dialogue. An enthymeme being *under discussion* on a speaker's game board means that this speaker acknowledges the enthymeme to be an argument put forward in relation to some issue raised in the dialogue. There may be several enthymemes simultaneously under discussion. Note that recognising an enthymeme as being under discussion is not the same as accepting it as valid.[14] Arguably, speakers are aware of many topoi, some of which they do not agree with, and use them to recognise rhetorical structure.[15]

Finally, the field 'private' contains information private to one interlocutor; this includes an 'agenda' and another field 'topoi' that records the

---

[11] Similarly, questions like *'Why don't you come in?'* are conventionally read as suggestions: the space of possible answers includes *'Thank you'*. However, if the suggestion is *not* followed, the literal Why-question can be answered by giving a reason, *e.g.*, *'I don't want to impose.'*

[12] One particular advantage attributed to TTR is that it allows one to model natural language without appealing to (sets of) possible worlds. Possible world models are criticised for having both cognitive and computational problems; see for example Ranta (1994) or Chatzikyriakidis and Luo (2014).

[13] It is noteworthy, however, that the model we apply bears a strong connection to rather more conventional logics of default inference (Breitholtz, 2014b).

[14] In most cases we discuss here, recognising a pair of utterances as forming an enthymeme is a given, as they are rhetorically connected by a *Why*-question.

[15] Breitholtz mentions political examples like *we love freedom – we are against taxes* that can be recognised even by people who do not support the argument themselves.

rhetorical resources, *i.e.*, the topoi acceptable to this particular interlocutor. The record type (38) in the next subsection is an example for a DGB.

Now we can formalise enthymemes and topoi. Following Breitholtz (2014a), we model enthymemes and topoi in the same way: as functions from *situations* to *situation types*. An enthymeme $A \therefore B$ expresses that in a situation satisfying A, B holds, *i.e.*, if s is an A-situation, then s is also a B-situation. Hence, the enthymeme can be represented as a mapping of situations in which A holds to a type corresponding to B. If such a function can be computed from an available topos, then the enthymeme is acceptable. Thus we also represent topoi as such functions and say that a topos licenses an enthymeme if from the function representing the topos we can compute the function representing the enthymeme. In the typical case of a topos being a general principle, this computation would be to *restrict* the topos to the situations in which the enthymeme is supposed to apply.

Note in particular that this means that the formal representation of enthymemes and topoi is the same: functions on situations of certain types. The difference between the two concepts lies in their dialogue dynamics: an enthymeme under discussion *claims* that there is such a function and an available topos says that there, in fact, *is* one.

4.1 A formal account of reasons

In section 3.1 we stipulated that q is a reason for p if there is a topos that validates $q \therefore p$. The examples there also show that what is taken by one dialogue participant as an acceptable validation of an argument may be unacceptable—even unrecognisable—to another. Consider the example in (7), where J points to a topos that seems to be a possible backing for the enthymeme conveyed by D's utterance—and then rejects the enthymeme. Let us consider the enthymeme conveyed by D in *'I'm self-funding my campaign, I tell the truth'*, here formalised as $\mathcal{E}_1$ in (32).

(32) $$\mathcal{E}_1 = \lambda \text{r}{:}\begin{bmatrix} x = \text{SELF} : \textit{Ind} \\ c_{\textit{self\_fund}} : \text{self\_fund(x)} \end{bmatrix} . \begin{bmatrix} c_{\textit{truth}} : \text{tell\_truth(r.x)} \end{bmatrix}$$

J points out that he considers *'rich people tell the truth'* to be the topos that underpins D's statement. We formalise this as the topos $\mathcal{T}_1$ in (33).

(33) $$\mathcal{T}_1 = \lambda \text{r}{:}\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\textit{rich}} : \text{rich(x)} \end{bmatrix} . \begin{bmatrix} c_{\textit{truth}} : \text{tell\_truth(r.x)} \end{bmatrix}$$

Now, to see that (33) justifies (32), we need to derive the function $\mathcal{E}_1$ from the function $\mathcal{T}_1$. First, as we discussed, the application of a topos can require further tacit premises and topoi. Here, it seems reasonable to assume that J counts *someone who self-funds their campaign is rich* among his rhetorical resources. This is the topos $\mathcal{T}_2$ in (34).

(34) $$\mathcal{T}_2 = \lambda \text{r}{:}\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\textit{self\_fund}} : \text{self\_fund(x)} \end{bmatrix} . \begin{bmatrix} c_{\textit{Rich}} : \text{rich(r.x)} \end{bmatrix}$$

Intuitively, to justify (32), one needs to apply (34) and (33) in succession. That is, we can compute $\mathcal{E}_1$ by composing $\mathcal{T}_1 \circ \mathcal{T}_2$ and instantiating the individual x as the person D. Note, however, that $\mathcal{T}_1$ is probably not acceptable to most people, and it also seems likely that D had a different topos in mind for underpinning his statement.

4.2 Factive reasons

Let us return to example (1), repeated here as (35).

(35) a. B: He's in hospital.
b. C: Why?
c. B: Because he's not very well

After B uttered (35a), C and B updated the latest move '1-m' on their DGBs to include the claim that X *is in the hospital*, where X is the anaphoric resolution of 'he'. In uttering *Why?*, C inquires about the reason why this is the case. To answer, B searches her rhetorical resources for a topos that can underpin an inference $\varphi \therefore \psi$ satisfying these properties: ψ can be used to conclude that someone is in the hospital, and φ applies to X in this context. Such a topos has the form of $\mathcal{T}_1$ in (36).

(36) $$\mathcal{T}_1 = \lambda \text{r}{:}\begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\textit{unwell}} : \text{unwell(x)} \end{bmatrix} . \begin{bmatrix} \text{c}_{\textit{hospital}} : \text{in\_hospital(r.x)} \end{bmatrix}$$

This is, intuitively, a generally acceptable pattern of inference. So it is plausible that B can retrieve $\mathcal{T}_1$ from her rhetorical resources. Informed by this topos, B then utters (35c), expressing the enthymeme in (37).

(37) $$\mathcal{E}_1 = \lambda \text{r}{:}\begin{bmatrix} \text{x} = \text{X} : \textit{Ind} \\ \text{c}_{\textit{male}} : \text{male(x)} \\ \text{c}_{\textit{unwell}} : \text{unwell(x)} \end{bmatrix} . \begin{bmatrix} \text{c}_{\textit{hospital}} : \text{in\_hospital(r.x)} \end{bmatrix}$$

Now, the other speaker C, upon interpreting (35c), updates his game board to include $\mathcal{E}_1$ as the *enthymeme under discussion* ('eud'), as seen on C's game board in (38).

(38) $$\text{DGB}_C = \begin{bmatrix} \text{private} : \begin{bmatrix} \text{agenda} : \text{list}(\textit{RecType}) \\ \text{topoi} : \text{list}(\textit{Rec} \rightarrow \textit{RecType}) \end{bmatrix} \\ \text{shared} : \begin{bmatrix} \text{l-m} : \begin{bmatrix} \text{e} : \text{Assert(B,C)} \\ \text{ctnt} : T_{\textit{unwell}} \end{bmatrix} \\ \text{qud} : \text{list}(\textit{Question}) \\ \text{moves} : \text{list}(\textit{Illoc}) \\ \text{eud} = [\mathcal{E}_1] : \text{list}(\textit{Rec} \rightarrow \textit{RecType}) \\ \text{topoi} : \text{list}(\textit{Rec} \rightarrow \textit{RecType}) \end{bmatrix} \end{bmatrix}$$

Then, C searches his resources for a topos of which $\mathcal{E}_1$ is an instantiation. Let us assume here that he can retrieve the topos $\mathcal{T}_1$ as well. Note that the domain of $\mathcal{T}_1$ is a more general type than the domain of $\mathcal{E}_1$. Thus, $\mathcal{T}_1$ can be restricted to the function $\mathcal{E}_1$, underpinning the enthymeme in (37).

4.3 Meta-discursive reasons

We also observed that *Why*-questions can have *meta-discursive* readings, *i.e.*, asking for the justification of a *linguistic* fact. We model this as follows: factive *Why*-questions ask about the latest move's *content*, whereas meta-discursive questions ask about the move *itself*.

In a game board semantics, the contents of the DGB are modified via *update rules* that link the progression of the dialogue (as recorded in 'shared') to the interlocutors' beliefs and plans (as recorded in 'private'). The function $\mathcal{U}_{\text{why}_f}$ in (39) is an update rule for factive *Why*-questions: the interlocutor asks about a premise for an enthymeme justifying the *content* of 'l-m'. We use $p \therefore q$ to abbreviate the type of the enthymeme 'p hence q' and the notation $l = \langle x \mid .\rangle$ to say that x is the first element of the list l.

(39)

$$\mathcal{U}_{\text{why}_f} = \lambda\text{r:} \left[\text{shared}: \left[\text{l-m}: \left[\begin{array}{l}\text{e}: \text{Assert(SELF, OTHER)}\\ \text{ctnt}: T_c\end{array}\right]\right]\right].$$

$$\left[\begin{array}{ll}\text{shared:} & \left[\begin{array}{l}\text{qud} = \langle \lambda\text{p.(p} \therefore \text{r.shared.l-m.ctnt)}|.\rangle : \text{list}(\textit{Question})\\ \text{l-m}: \left[\begin{array}{l}\text{e}: \text{Ask(OTHER,SELF)}\\ \text{ctnt}: \lambda\text{p.(p} \therefore \text{r.shared.l-m.ctnt)}\end{array}\right]\end{array}\right]\end{array}\right]$$

Note that updating 'l-m' tacitly also updates 'moves'. Now, for meta-discursive *Why*-questions, the interlocutor inquires about the move *itself*. For the case of a question in the antecedent, this is the function $\mathcal{U}_{\text{why}_m}$ in (40).

(40)

$$\mathcal{U}_{\text{why}_m} = \lambda\text{r:} \left[\text{shared}: \left[\text{l-m}: \left[\begin{array}{l}\text{e}: \text{Ask(SELF,OTHER)}\\ \text{ctnt}: T_c\end{array}\right]\right]\right].$$

$$\left[\begin{array}{ll}\text{shared:} & \left[\begin{array}{l}\text{qud} = \langle \lambda\text{p.(p} \therefore \text{r.shared.l-m)}|.\rangle : \text{list}(\textit{Question})\\ \text{l-m}: \left[\begin{array}{l}\text{e}: \text{Ask(OTHER,SELF)}\\ \text{ctnt}: \lambda\text{p.(p} \therefore \text{r.shared.l-m)}\end{array}\right]\end{array}\right]\end{array}\right]$$

Now we can formalise (16), repeated here as (41).

(41) a. A: Do you love me ⟨unclear⟩?
b. B: Why?
c. A: ⟨unclear⟩ I love you so much.

The *Why*-question in (41b) aims at eliciting a reason for asking, not a motivation for the content being true. On a certain level of abstraction, this can be modelled in much the same way as a factive *Why?*. That is, we can represent the *fact* that one speaker has asked a question as a situation type. So, the 'eud' after (41c) can be put as (42).

(42)

$$\mathcal{E}_1 = \lambda\text{r:} \left[\begin{array}{l}\text{x} = \text{A}: \textit{Ind}\\ \text{y} = \text{B}: \textit{Ind}\\ \text{c}_{\textit{love}}: \text{love(x,y)}\end{array}\right]. \left[\text{c}_{\textit{asked}}: \text{asked(r.x,?love(r.y, r.x))}\right]$$

Thus, after (41c), $\mathcal{E}_1$ is now under discussion. That is, B has to evaluate whether he can accommodate *A loving B* is a reason for asking (41a). We attributed the topos *'one wants to know if love is requited'* to this example in (17). This can be formalised as $\mathcal{T}_1$ in (43).

(43)

$$\mathcal{T}_1 {=} \lambda\text{r:} \left[\begin{array}{l}\text{x}: \textit{Ind}\\ \text{y}: \textit{Ind}\\ \text{c}_1: \text{love(x,y)}\end{array}\right]. \left[\text{c}_{\textit{wtk}}: \text{want\_to\_know(r.x,?love(r.y,r.x))}\right]$$

Again we need to assume a tacit background topos. In this case, *'if someone desires to know something, this is a reason for asking for it'*.[16] This is the topos $\mathcal{T}_2$ in (44).

(44)

$$\mathcal{T}_2 = \lambda\text{r:} \left[\begin{array}{l}\text{x}: \textit{Ind}\\ \text{y}: \textit{Question}\\ \text{c}_{\textit{wtk}}: \text{want\_to\_know(x,y)}\end{array}\right]. \left[\text{c}_{\textit{asked}}: \text{asked(r.x,r.y)}\right]$$

As before, computing $\mathcal{T}_2 \circ \mathcal{T}_1$ and restricting the domain to the proposition *love*(y, x) for the individuals $x = A$ and $y = B$ yields $\mathcal{E}_1$.

As said, this is on a certain level of abstraction. The constraint $\text{c}_{\textit{ask}}$ differs from $\text{c}_{\textit{hospital}}$ in (36) in that the former specifies a *linguistic* situation. The DGB allows us to be more precise about what such linguistic situations are. We may represent 'asked(A,?love(B,A))' as the type in (45).

(45)

$$\left[\text{shared:} \left[\text{moves} {=} \left\langle . \middle| \left[\begin{array}{l}\text{e}: \text{Ask(A,B)}\\ \text{ctnt}: \text{?love(B,A)}\end{array}\right] \middle| . \right\rangle : \text{list}(\textit{Illoc})\right]\right]$$

5 Conclusion

We have conducted an analysis of the functions that *Why*-questions can have in dialogue and explained them from the perspective of enthymemes and topoi. Our discussion covers the phenomenon broadly, but there remain open questions related to embedded *Why*-questions and elliptical *Why?*. The cornerstone of our analysis is a definition of what counts as a *reason*, *i.e.*, as an answer to a *Why*-question. We have formalised that notion in a TTR framework and formally described two examples for the major functions we have attributed to *Why*-questions in the informal analysis.

---

[16] Note that this is an example of a topos that appears generally reasonable, but fails to apply in many situations. *E.g.*, when asking would be embarrassing or socially dispreferred.

Acknowledgements

The research presented in this paper has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 567652 *ESSENCE: Evolution of Shared Semantics in Computational Environments* (`http://www.essence-network.com/`) and by the Netherlands Organisation for Scientific Research (NWO) under VIDI grant n. 276-89-008, *Asymmetry in Conversation*.

References

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

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

Ellen Breitholtz and Jessica Villing. 2008. Can aristotelian enthymemes decrease the cognitive load of a dialogue system user? In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2008, 'LonDial')*, pages 94–100.

Ellen Breitholtz. 2010. Clarification requests as enthymeme elicitors. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2010, 'PozDial')*, pages 135–138.

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.

Sylvain Bromberger. 1992. *On what we know we don't know*. University of Chicago Press.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Lewis Carroll. 1895. What the tortoise said to achilles. *Mind*, 4(14):278–280.

Stergios Chatzikyriakidis and Zhaohui Luo. 2014. Assistant technology: Rich typing and beyond. In *Proceedings of EACL 2014 TTNLS Workshop*, pages 37–45.

Robin Cooper and Jonathan Ginzburg. 2012. Negative inquisitiveness and based negation. In *Logic, Language and Meaning*, pages 32–41. Springer.

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. Wiley-Blackwell.

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

Oswald Ducrot. 1980. *Les échelles argumentatives*.

Oswald Ducrot. 1988. Topoï et formes topique. *Bulletin d'études de la linguistique française*, 22:1–14.

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

Carl Gustav Hempel. 1965. *Aspects of Scientific Explanation*. The Free Press.

Jaakko Hintikka and Ilpo Halonen. 1995. Semantics and pragmatics for why-questions. *The Journal of Philosophy*, 92(12):636–657.

Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*, 66(3):251–265.

Matthew Purver. 2001. SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Department of Computer Science, King's College London.

Aarne Ranta. 1994. *Type-theoretical Grammar*. Oxford science publications. Clarendon press.

Julian J. Schlöder and Raquel Fernández. 2014. Clarification requests on the level of uptake. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2014, 'DialWatt')*, pages 237–239.

Julian J. Schlöder and Raquel Fernández. 2015. Clarifying intentions in dialogue: A corpus study. In *Proceedings of the 11th Conference on Computational Semantics*, pages 46–51.

Bas C. van Fraassen. 1980. *The scientific image*. Oxford University Press.

Marilyn A. Walker. 1996. The effect of resource limits and task complexity on collaborative planning in dialogue. *Artificial Intelligence*, 85(1):181–243.

Joint information structure and discourse structure analysis in an Underspecified DRT framework

Uwe Reyle and Arndt Riester
Institute for Natural Language Processing
University of Stuttgart, Germany
Pfaffenwaldring 5b, 70569 Stuttgart
`{uwe.reyle,arndt.riester}@ims.uni-stuttgart.de`

Abstract

This paper presents the methodology and semantics of a general procedure for the joint analysis of textual data in terms of discourse structure and information structure, which makes use of Questions under Discussion (QUDs). We define a number of pragmatic principles that govern the reconstruction of implicit QUDs.

1 Introduction

This paper introduces major aspects of a method for the analysis of natural language in terms of information structure and discourse structure using *Questions under Discussion (QUDs)*, which will be demonstrated on a short constructed discourse.[1] The main purpose of the paper is to introduce a number of principles that determine the formulation of QUDs, as well as a semantic implementation of the procedure in Underspecified Discourse Representation Theory (UDRT) (Reyle, 1993).

By the term *information structure*, we are referring to a division of clauses into an alternative-evoking *focus* and a *background* (plus some optional, so-called *not-at-issue*, material), largely following the paradigm of *Alternative Semantics*, established by Rooth (1985; 1992) and developed further, for instance, in Büring (2003; 2008; in press), Beaver and Clark (2008), Krifka (2008) or Wagner (2012). In order to determine the information structure of a clause, it is usually necessary to consider the discourse context in which it is uttered, although some aspects of its information structure will be reflected – to a language-specific degree – in its morphosyntactic properties or, when spoken, in its prosodic realization. In line with assumptions made in Klein and von Stutterheim (1987), Ginzburg (1996) and Roberts (2012), we are assuming that discourse not only consists of the overt spoken or written material but, in addition, contains implicit *Questions under Discussion* that provide the background against which the actual assertions are made. The *focus* of any clause uttered in its respective discourse context can, therefore, be defined as the answer to its current QUD. In the following section, we present a number of principles that will help us reconstruct the implicit Questions under Discussion of a text.

The term *discourse structure* is generally understood to explain the organization of a text into smaller sections and subsections, down to the level of atomic assertions. We assume that a well-formed text can be represented in the form of a single discourse tree. In contrast to various established theories of discourse structure, e.g. Mann and Thompson (1988), Taboada and Mann (2006), or Asher and Lascarides (2003), the current proposal does not depend on the identification of *discourse relations (rhetorical relations)* but assumes that the structure of discourse can be reconstructed with the help of Questions under Discussion, which are supposed to constitute an essential part of discourse trees.

2 Constraints on the construction of implicit Questions under Discussion and discourse trees

A fundamental, and probably uncontroversial, constraint on the formulation of a QUD is that a QUD that immediately dominates some assertion must be congruent with it.

First QUD Constraint (Q-A-Congruence)
QUDs must be answerable by the assertion(s) that they immediately dominate.

In the absence of context, (A)ssertion A_2 in (3) can be the answer to any of the (Q)uestions in (1a)-(1d) but not to question (2).

---

[1] But see Riester (2015), Riester and Piontek (2015) for first analyses of real corpus data.

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 15-24.

(1) a. Q: {What happened?}
b. Q: {What did they do?}
c. Q: {Who worked hard?}
d. Q: {Did they work hard?}

(2) Q: {Who bought a bicycle?}

(3) A_2: They worked hard.[2]

If more context is introduced, as in (4), it becomes clear that the questions in (1a-d) are not all equally good.

(4) A_1: John and Mary are really proud.

(3) A_2: They worked hard.

It seems intuitively clear that question (1c) does not fit in between assertions A_1 and A_2. The apparent reason is that, in the context of A_1, Question (1c) would introduce the phrase *worked hard* as new information, which seems to be dispreferred. Likewise, assuming the polarity question (1d) as the implicit QUD would force us to treat *worked hard* as given information at the level of the answer, and to interpret A_2 in the sense of *Yes, they DID work hard*, which seems odd in the current context. Note that apparently there is an important difference between explicit and implicit questions. While explicit questions can be used to introduce new information without causing any problems, the role of implicit questions is confined to enabling a smooth transition between two assertions, without the option of introducing any new material by themselves and thereby changing the actual discourse. We formulate this in a second constraint.

Second QUD Constraint (Q-Givenness)
Implicit QUDs can only consist of given (or, at least, highly salient)[3] *material.*

The principle of Q-GIVENNESS directly follows from the GIVENNESS principle by Schwarzschild (1999), which, in effect, says that discourse-new information is necessarily focused. Since in a question-answer pair the focus of the answer typically corresponds to a *wh*-pronoun in the question while only the background occurs in both of them, we conclude that discourse-new material is banned from implicit QUDs. This explains why the Questions (1a) or (1b) represent better transitions from A_1 to A_2 than do (1c) or (1d)[4] – the latter ones violate Q-GIVENNESS.

(4) A_1: John and Mary are really proud.

(1) a. Q: {What happened?}
b. Q: {What did they do?}
c. #Q: {Who worked hard?}
d. #Q: {Did they work hard?}

(3) A_2: They worked hard.

But should we prefer question (1a) or (1b)? (1a) evokes a broad sentence focus while (1b) contains an anaphoric pronoun *(they)* and asks for a predicate in focus. The question that contains the anaphoric pronoun creates a higher degree of textual cohesion (Halliday and Hasan, 1976) and is, therefore, preferable. This has been expressed in various principles in the literature which all demand, in some sense, that sentences should be maximally anaphoric or given and, therefore, have a minimal focus; for instance, the principles MAXIMIZE PRESUPPOSITION (Heim, 1991), AVOIDF (Schwarzschild, 1999) or MAXIMIZE ANAPHORICITY (Büring, 2008). Applying this idea to QUDs, we define a third constraint.

Third QUD Constraint (Maximize Q-Anaphoricity)
Implicit QUDs should contain as much given or salient material as possible.

Now, since (1a) violates MAXIMIZE Q-ANAPHORICITY, (1b) is chosen as the actual QUD Q_2, in the respective context (indicating question-answer congruence by means of identical subscripts.) Concerning the discourse structure of the example, we assume that answers must be subordinated to their question. Furthermore, questions which make reference to previously mentioned material must be subordinated to the clause containing this antecedent material, as shown in Figure 1.

We take the three principles mentioned above to be hard constraints, which must be fulfilled

---

[2] We choose a simple past form in A_2 for the sake of having simple representations.

[3] We assume that function words (determiners, pronouns, prepositions etc.), as well as very general concepts like *to happen* are always salient, even if they are not literally given in the discourse context.

[4] Another question that comes to mind is *Why are they proud?*, asking for an explanation, which denotes a proposition rather than a predicate. While this is indeed a likely question in the given context, it is at odds with the intuition that the subject pronoun should be excluded from the focus of A_2. An ad-hoc solution to this recurring problem with explanations is to allow for a nesting of two questions: *Why are they proud?* > *What did they do?* and to let A_2 function as the simultaneous answer to both of them.

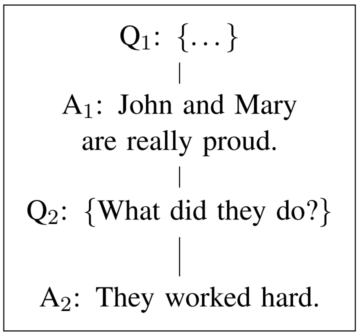

Q$_1$: {...}
A$_1$: John and Mary are really proud.
Q$_2$: {What did they do?}
A$_2$: They worked hard.

Figure 1: Discourse tree

at any time (with one important exception to Q-GIVENNESS discussed below), i.e. there will be no ranking between them, and it is precisely the universality of the constraints that makes them useful in the analysis of text.

3 Information structure

With the principles defined in the previous Section, we are now able to account for the information structure of our discourse. Definitions are provided in Table 3.

Category	Definition
Focus (F)	Answer to the current QUD
Focus Domain ($\sim$) (Rooth, 1992) Büring (2008)	Stretch of discourse with the same background as the QUD (possibly recurring elsewhere with a different focus)
Background (BG)	Material given in the QUD
Aboutness Topic (T)	Referential entity in the background
Contrastive Topic (CT) (Büring, 2003)	Focused topic, signals a discourse strategy (explanation below)

Table 1: Information structure inventory

As noted above, the QUD determines the focus-background divide of its answer. The information structure of A$_2$ is, therefore, the one shown in (5).[5]

(5) Q$_2$: {What did they do?}
$>$ A$_2$: [[They$_T$]$_{BG}$ [worked hard]$_F$.]$\sim$

Following Rooth (1992) and Büring (2008), we adopt a holistic approach, i.e. we are not only interested in the position of the focus itself but in the entire combination of focus and background taken together, called a focus domain ($\sim$). In addition, we suggest a definition of *aboutness topics* as backgrounded referring expressions. This means that aboutness topics are necessarily in the background but not all backgrounded information qualifies as a topic, as shown in (6).

(6) Q$_{10}$: {What is John going to eat?}
$>$ A$_{10}$: [[John$_T$ is going to eat]$_{BG}$ spinach$_F$.]$\sim$

Again, we see a background-focus divide but only the referring expression *John* counts as topic.[6]

The next issue in this informal discussion is *parallelism*. Again, focus domains will play a crucial role. We discuss two types of parallelism: a simple one with only one focus per assertion, and a complex one that contains pairs consisting of a focus and a contrastive topic. Explicit parallelisms, like the one in (7)[7], are rare in natural discourse, since they will typically occur in elliptical form and be rendered as simple co-ordinations. In (8)[8], the elided material has been recovered, which is indicated by means of strikethrough text.

(7) Q$_{50.1}$: {Whom can you wire-tap?}
$>$ A$_{50.1'}$: [[You$_T$ can wire-tap]$_{BG}$ [the President of the United States]$_F$]$\sim$,
$>$ A$_{50.1''}$: [[you$_T$ can wire-tap]$_{BG}$ [a Federal Judge]$_F$]$\sim$.

(8) Q$_{60}$: {What will the bill prescribe?}
$>$ A$_{60'}$: [[[The bill]$_T$ will prescribe]$_{BG}$ [having windows in staff kitchens]$_F$]$\sim$
$>$ A$_{60''}$: and [[~~it$_T$ will~~ also ~~prescribe~~]$_{BG}$ [the brightness of the home workplace]$_F$]$\sim$.

It seems reasonable to assume that, indeed, most co-ordinations in assertions can be analyzed as remnants of elided parallel statements. In

---

[5] For reasons of space, we represent subordination in a tree by means of a $>$.

[6] This is in line with Krifka (2008), who assumes a *topic-comment* distinction that need not be coextensive with BG-F. Our definition makes no use of the *comment* notion. It remains to be sorted out whether one wants to allow for several aboutness topics in one utterance or whether backgrounded referring expressions should compete for topichood according to grammatical and thematic role, animacy etc., cf. Reinhart (1981), Givón (1983), Brunetti (2009).

[7] Quote: Edward Snowden in an interview with German TV (ARD), Jan. 26, 2014.

[8] Ex. translated from *Stuttgart SFB 732 Silver Standard Corpus* (German radio interviews).

information-structural terms, the coordinated elements are (contrastive) foci. The two parallel assertions, whether overtly present in the text or partly reconstructed, function as two partial answers to a common QUD, with whom they share the same background (and, therefore, a structurally identical focus domain). We indicate this by using subscripts of the form $\mathrm{A}_{1'}$, $\mathrm{A}_{1''}$. Examples (7) and (8) show that parallelisms provide us with a second way of identifying Questions under Discussion. QUDs can simply be determined by collecting the parallel material of two (or more) subsequent clauses, and by replacing the variable – i.e. focal – material by a *wh*-pronoun. We define a fourth constraint.

Parallelism constraint
The background of a QUD with two or more parallel answers consists of the (semantically) common material of the answers.

The PARALLELISM constraint will sometimes collide with, and override, the principle of Q-GIVENNESS defined above, since the parallel, backgrounded material need not always be salient already. This means that a parallelism may sometimes force the interpreter to accommodate a more specific (sub-) question – $\mathrm{A}_{50.1}$ in (9) – than the one that would be licensed from the previous discourse alone (A_{50}). The notation is meant to indicate that A_{50} and $\mathrm{A}_{50.1}$ stand in an entailment relation, cf. Groenendijk and Stokhof (1984, 16); Roberts (2012, 6f.)

(9) *Context: When you are on the inside and you go into work everyday and you sit down at the desk and then you realize the power you have.*
$>$ Q_{50}: {What power do you have?}
$>>$ $\mathrm{Q}_{50.1}$: {Whom can you wire-tap?}
$>>>$ $\mathrm{A}_{50.1'}$: [[You$_T$ can wire-tap]$_{BG}$ [the President of the United States]$_F$]$\sim$,
$>>>$ $\mathrm{A}_{50.1''}$: [[you$_T$ can wire-tap]$_{BG}$ [a Federal Judge]$_F$]$\sim$.

We now turn to the issue of complex parallelisms, i.e. two subsequent assertions that differ with respect to two syntactic positions.[9] Like in the case of simple parallelisms, it is again possible to define a common QUD, albeit one containing two *wh*-pronouns (or, at least, a question that expresses variability in two positions). Among the two variable – i.e. focal – positions, one must take precedence over the other. Following Büring (2003), we will call this primary position the *contrastive topic*, the other one the *focus*. Furthermore, each contrastive topic introduces a more specific subquestion. An example is given in (10), in which the subquestions of the main question Q_3 are indicated as $\mathrm{Q}_{3.1}$ and $\mathrm{Q}_{3.2}$.

(10) Q_3: {Who did what?}
$>$ $\mathrm{Q}_{3.1}$: {What did John do?}
$>>$ $\mathrm{A}_{3.1}$: [John$_{CT}$ [painted a self-portrait]$_F$]$\sim$
$>$ $\mathrm{Q}_{3.2}$: {What did Mary do?}
$>>$ $\mathrm{A}_{3.2}$: and [Mary$_{CT}$ [rehearsed a piano sonata]$_F$]$\sim$.

For the sake of completeness – although it will not play a role in the rest of the paper – we briefly sketch our treatment of *not-at-issue* material (more precisely, triggers of *conventional implicatures*), including evidentials, appositions, parentheses, speaker-oriented adverbs and others, cf. Potts (2005), Simons et al. (2010). Generally, we declare some expression to be *not-at-issue* with respect to the current QUD *iff* deleting the expression has no influence on the interpretability and the truth-conditions of the main assertion. As an example, take the evidential phrase *Paul said that* in (11), marked in gray.

(11) Q_{11}: {What is John going to eat?}
$>$ A_{11}: Paul said that [[John$_T$ is going to eat]$_{BG}$ spinach$_F$]$\sim$.

It is crucial to keep in mind that calling an expression *not-at-issue* is merely a statement about its relation to the current QUD. It should not be misunderstood as a negative rating of its relevance to the discourse as a whole. Structurally, we treat not-at-issue content as forming an answer (A_{12}) to a (non-entailed) subquestion, which comes with its own information structure (Riester and Baumann 2013, 221), as shown in (12).

---

[9] Researchers working in the SDRT framework might not want to call an example involving CT and F a *parallelism*, since the transition between the two utterances does not license the discourse connector *too*. Instead, at least for some cases, though probably not for all, a CONTRAST relation seems appropriate. What matters for us in this regard is that the two (or more) assertions containing CT and F might still share some backgrounded linguistic material: *[Fred$_{CT}$ ate$_{BG}$ beans$_F$]$\sim$; [Carl$_{CT}$ ate$_{BG}$ peas$_F$]$\sim$*, which is why we keep using the *parallelism* notion in a broad sense.

(12) Q_{11}: {What is John going to eat?}
$>$ A_{11}: [John$_T$ is going to eat spinach$_F$]$\sim$.
$>>$ Q_{12}: {Where does this information come from?}
$>>>$ A_{12}: [[Paul said]$_F$ ~~it$_T$~~.]$\sim$

4 Construction of QUDs and background-focus structures

Following Kamp (1998),[10] we represent BG-F structures by means of pairs of DRSs (of λ-DRT, Kohlhase et al. 1996), as shown in the lower half of the preliminary DRS for the answer in (3), Figure 2.

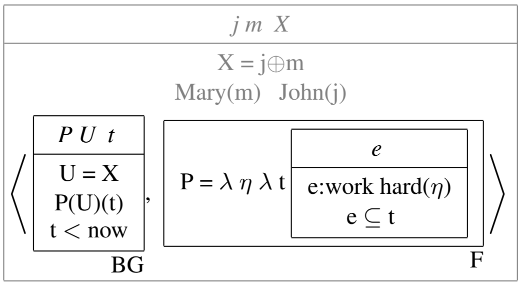


Figure 2: DRS for A_2: *[They$_{BG/T}$ [worked hard]$_F$.]$\sim$* (within context)

The first member of this pair represents the background (BG) and the second one the focus (F). The variable P is called the focus-variable and the λ-DRS the focus value. Unification of the two components gives the ordinary meaning of the sentence A_2, and different values for the focus variable P determine its focus alternatives with respect to the first component. The focus alternatives are usually – following Rooth (1992) – claimed to be restricted by a condition C that has to be retrieved from the context in such a way that the alternatives contain at least one other proposition. We have not made this presupposition of focus explicit here, because in cases of parallelism it will be automatically fulfilled and for the other cases we discuss here it wouldn't contribute anything substantial. As the first component of BG-F pairs determines a set of alternatives it also determines the representation of the (implicit) QUD the sentence answers. We only need to let the focus variable be bound by a question operator $\mathcal{Q}$. So, the implicit QUD that A_2 answers will be represented as in Figure 3.[11]

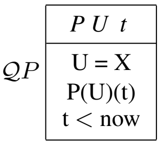


Figure 3: DRS for implicit QUD Q_2: {*What did they do?*}

Our construction of QUDs will, however, be based not directly on DRS-representations but on UDRS-representations (Reyle, 1993). This is necessary, because we need to have access to the different syntactic components of a sentence, which are explicitly present in UDRS representations. Consider (13).

(13) A_2: They worked hard.

The UDRS for (13) is given by the components K_{TENSE}, K_{SUBJ}[12] and K_{VERB}, as specified in Figure 4, with the partial order given in Figure 5. (A more complex example is presented below for sentence (14) in Figures 6 and 7.) The order between the components is such that if a discourse referent occurs free in a component K, then the component in which this discourse referent is declared must dominate K. Temporal information dominates all other components of a clause. A UDRS is turned into a DRS by recursively unifying components bottom-up that immediately dominate each other. As long as there are no scope bearing elements involved the order doesn't matter.[13]

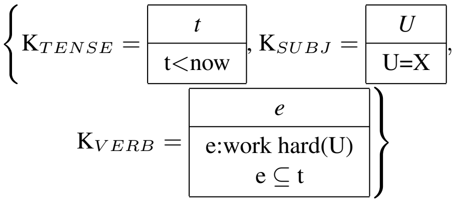


Figure 4: Components of A_2

Givenness

Having constructed the UDRS for A_2 in its context we now look for the maximal set of UDRS-components that are given, i.e. plausibly derivable

---

[10] BG-F structures go back to *Structured Meanings Theory* (von Stechow, 1982; Krifka, 1992).

[11] For yes-no questions we may assume that P is a polarity operator, i.e. P = λKK or P = λK$\neg$K.

[12] The pronoun *they* is taken to refer to the contextually given X representing *John and Mary*.

[13] The original motivation for UDRT is to have representations that leave the relative scope of quantifiers and other operators underspecified. For details see Reyle (1993).

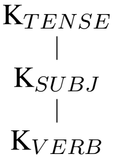


Figure 5: Partial order on the components

from the current context. This is required by Q-GIVENNESS and MAXIMIZE Q-ANAPHORICITY. Following Asher (1993, 305) we will call this set the *maximal common theme* between the sentence under consideration and its context. We see that K_{SUBJ} is trivially derivable from the context because the referent X is declared in it. K_{TENSE} is derivable too, because simple past presupposes a temporal location time in the past. But no other component may be shown to be given. Hence the relation determined by the complement set of the maximal common theme, i.e. $\lambda\eta\lambda t.K_{VERB}$[14], represents the discourse-new material of the second sentence and determines the focus variable P as provided by the second member of the BG-F pair in Figure 2. The first member of Figure 2 is determined by the merger of the common theme components K_{SUBJ} and K_{TENSE} together with the condition P(U)(t), stating that the focus value is applied to the referential argument U of the subject component and to the time period t.

Parallelism

Suppose we are at step i in the construction of the QUD for discourse $A_1, \ldots, A_n$. Then there are always two options. The first option is to integrate A_i only with respect the previous discourse A_1,..., A_{i-1}, as we did just above. However, with this givenness-based method we run the risk of determining too broad a focus, as we already showed in Example (9).[15] The second option is to look ahead and see if there is a parallelism between A_i and A_{i+1}. Let us look at a case of a simple parallelism first.

(14) $A_{3.1'}$: John painted a self-portrait
$A_{3.1''}$: and ~~he painted~~ a landscape.

The identification of simple parallel sentences as in (14) boils down to finding a non-empty common theme between the two sentences, here $K_{TENSE} \uplus K_{SUBJ} \uplus K_{VERB}$, where the UDRSs for the two conjuncts of (14) are given in Figure 6, and their order in Figure 7.

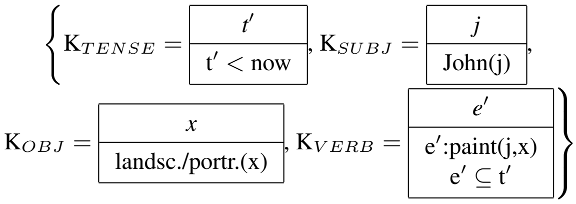


Figure 6: Components of $A_{3.1'}/A_{3.1''}$

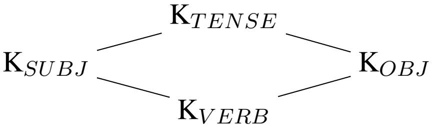


Figure 7: Partial order on the components of $A_{3.1'}/A_{3.1''}$

$K_{TENSE} \uplus K_{SUBJ} \uplus K_{VERB}$, the maximal common theme, thus forms the BG and K_{OBJ} specifies the value of a focus-variable x that is newly introduced into the universe of $K_{TENSE} \uplus K_{SUBJ} \uplus K_{VERB}$ and replaces all free occurrences of z in its condition set. For for the second conjunct this gives us the BG-F representation in Figure 8.

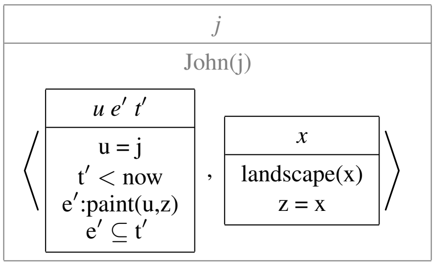


Figure 8: DRS for $A_{3.1''}$: *[[he$_T$ painted]$_{BG}$ [a landscape]$_F$]*$\sim$ (within context)

The representation of the first conjunct of (14) is identical to the one in Figure 8, except for the focus variable, which denotes a self-portait and not a landscape. Note that both sentences are now represented as answers to the same question, namely the question that is represented by their first (and identical) components, i.e. *What did John paint?*

We now turn to the case of complex parallelisms. As we said, our notion of parallelism in these cases is not to be confused with the rhetorical relation PARALLELISM as used e.g. in SDRT.

---

[14] The λ-bound variable U has been replaced by η.

[15] From the perspective of speech production, this means that we might sometimes predict the wrong pitch-accent placement.

Consider the sequence in Example (15), the information structure of which will be represented by Figure 9.

(15) A_2: John and Mary worked hard last week.
$A_{3.1}$: John painted a self-portrait
$A_{3.2}$: and Mary rehearsed a piano sonata.

The occurrences of *John* and *Mary* in $A_{3.1}$ and $A_{3.2}$ are topics. This means on the one hand that we may assume the existence of *subquestions* about the two persons. On the other hand they are contrastive, which means that each of them requires an alternative to be present. In (15) the set of alternatives for the contrastive topics is explicitly given by the group of John and Mary introduced in A_2, but there are also cases where this type of antecedent has to be reconstructed on the bases of two constituents that have been identified as contrastive topics. Our procedure will thus first look for the existence of a structurally isomophic split of the two sentences into the two parts indexed by X and Y in (16).

(16) $A_{3.1}$: [John]$_X$ [painted a self-portrait]$_Y$.
$A_{3.2}$: [Mary]$_X$ [rehearsed a piano sonata]$_Y$.

We start with the maximal common theme – the BG – of the two sentences; in this case merely the tense information. Furthermore, the BG provides the basis for relating the variable meaning components of X and Y. Let us assume that X is chosen as the sortal key (the contrastive topic) (Büring, 2003) of the two answers $A_{3.1}$ and $A_{3.2}$, and Y functions as the focus. Then, BG is first constrained by identifying the referent of X in $A_{3.1}$ with *John*. This identification is represented in Figure 9 by the CT component. The merger of BG and CT will then result in the BG-F representation of $A_{3.1}$, viz. <BG$\uplus$CT,F>. This structure also identifies the (sub-)question to which $A_{3.1}$ is an answer, namely $\mathcal{Q}$P.BG$\uplus$CT, the question *What did John do?* This is done in an analogous way for $A_{3.2}$. Finally, the super-question *Who did what?* is determined by BG alone and has, in our case, the form $\mathcal{Q}$z$\mathcal{Q}$P.BG.[16]

The identification of parallel structures in text is a relatively easy task for a human interpreter. However, we need to say more about how the informal procedure can be made a bit more precise in an algorithmic form. For the construction of the representation in Figure 9 we proceed as follows. We first build the UDRSs for $A_{3.1}$ and $A_{3.2}$. We will assume that the UDRS representation for the first is given in the form of the already familiar UDRS in Figure 6 and partial order as in Figure 7. The second conjunct has a completely identical structure, with the components shown in Figure 10.

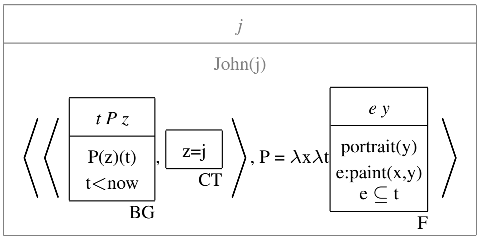


Figure 9: Information structure of $A_{3.1}$

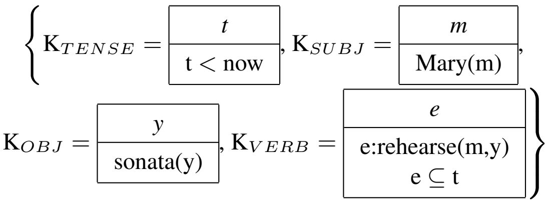


Figure 10: Components of $A_{3.2}$

Then, after having determined K$_{TENSE}$ as the maximal common theme of the two hypothetically parallel sentences, we will *split* the rest of each UDRS, i.e. the set of non-backgrounded components, into two parts, one of which will later represent the focus and the other the contrastive topic. After the split all components in each part are unified (by $\uplus$). The options for splitting are the following (remember that K$_{TENSE}$ is in the background): { K$_{SUBJ}$, K$_{OBJ}$$\uplus$K$_{VERB}$ }, { K$_{OBJ}$, K$_{SUBJ}$$\uplus$K$_{VERB}$ }, { K$_{VERB}$, K$_{SUBJ}$$\uplus$K$_{OBJ}$ }. Note that the splitting must be such that it results in two isomorphic orderings of the resulting UDRSs of the two parallel sentences.[17]

Each element of the split will now be turned into a pair that indicates alternatives to the given meaning, i.e. we form structured representations by introducing a variable that ranges over the semantic type of the component. In Figure 11,

---

[16] If, instead, Y is chosen as the sortal key the super-question is defined in the reverse order of variables $\mathcal{Q}$P$\mathcal{Q}$z.BG.

[17] For cases as, e.g., *Mary and John went to have a picnic at the seaside. John prepared the lunch. Mary swam.* there cannot be a split of the form { K$_{OBJ}$, K$_{SUBJ}$$\uplus$K$_{VERB}$ } for the second and { K$_{SUBJ}$, K$_{VERB}$ } for the third sentence. Only 'parallel' splits are possible, i.e. { K$_{SUBJ}$, K$_{OBJ}$$\uplus$K$_{VERB}$ } for the second { K$_{SUBJ}$, K$_{VERB}$ } for the third.

K_{SUBJ} is split into a variable z ranging over individuals, and the lexical content of *John*; for $K_{OBJ} \uplus K_{VERB}$ we get the type by abstraction over the free variables. After renaming, we thus get $P = \lambda x \lambda t'.K_{OBJ} \uplus K_{VERB}$. The alternatives of P are constrained by applying P to z and t, declared in the other components of the URDS.

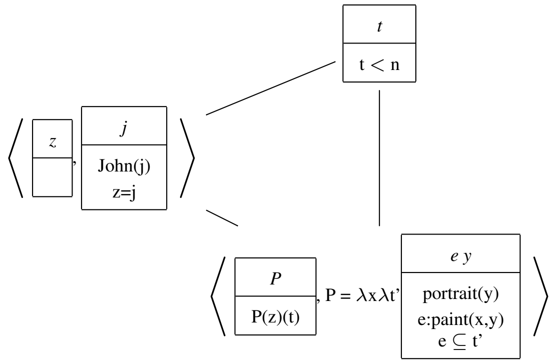


Figure 11: URDS of $A_{3.1}$ after splitting into $\{ K_{SUBJ}, K_{OBJ} \uplus K_{VERB} \}$ and structuring the non-backgrounded components

Suppose now, we decide to take the SUBJ to be the contrastive topic and the combination of VERB and OBJ to be the focus, then we will get the final BG-F representation in Figure 9 in the following way. The BG is obtained by unifying the background K_{TENSE} of Figure 11 with the first components of the two structured UDRS components. This is then paired with the second component of the subject, and the result is grouped together with the second component of the VERB-OBJ complex as the final <<BG, CT>, F> representation.

In the final example, (17), we apply the procedure from above to the issue of polarity contrast.

(17) A_5: Yesterday, I talked to John's mother.
$A_{6.1}$ She will praise him.
$A_{6.2}$: I won't ~~praise him~~.

The UDRS structure for $A_{6.2}$ is shown in Figure 12. As above K_{TENSE} can be put into the background. Furthermore, the property of *praising John* is in the background, too. This is represented in the bottom component of Figure 12 by the fact that the variable z ranges over all individuals that the first component of the subject representation may be mapped to. The figure also indicates the structuring of, on the one hand, the subject component (sp representing the speaker) and, on the other hand, of the polarity component.

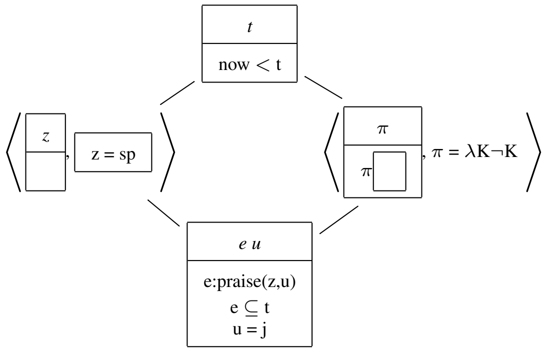


Figure 12: UDRS for $A_{6.2}$

If we want to have a split representation for $A_{6.1}$ in (17), which is structurally similar to $A_{6.2}$, we may introduce a node of the form λK.K (i.e. an identity condition) between its K_{TENSE} and $K_{OBJ} \uplus K_{VERB}$. This will not change the truth conditions of the representation and just serves to make the polarity contrast explicit. If again we take the subject to be the sortal key (i.e. the contrastive topic) we get the following two information-structural representations, in which x is the discourse referent introduced in the first sentence for John's mother.

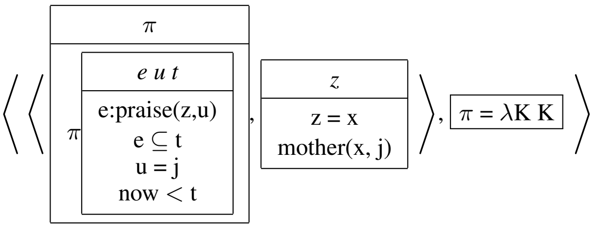


Figure 13: Information structure <<BG,CT>,F> of $A_{6.1}$

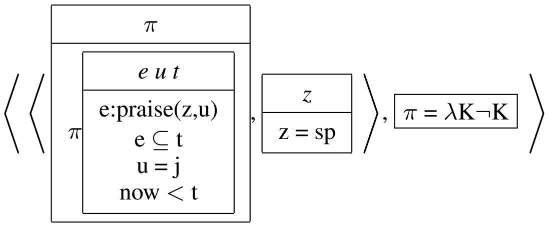


Figure 14: Information structure of $A_{6.2}$

5 Assembling the QUD tree

Let us now have a look at the discourse as a whole, repeated in (18).

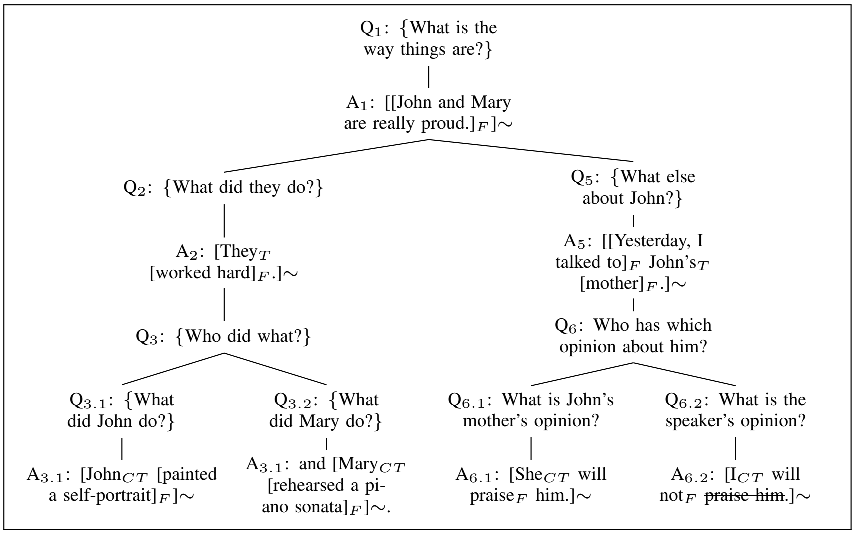

Q_1: {What is the way things are?}
A_1: [[John and Mary are really proud.]$_F$]~
Q_2: {What did they do?}
A_2: [They$_T$ [worked hard]$_F$.]~
Q_3: {Who did what?}
$Q_{3.1}$: {What did John do?}
$A_{3.1}$: [John$_{CT}$ [painted a self-portrait]$_F$]~
$Q_{3.2}$: {What did Mary do?}
$A_{3.1}$: and [Mary$_{CT}$ [rehearsed a piano sonata]$_F$]~.
Q_5: {What else about John?}
A_5: [[Yesterday, I talked to]$_F$ John's$_T$ [mother]$_F$.]~
Q_6: Who has which opinion about him?
$Q_{6.1}$: What is John's mother's opinion?
$A_{6.1}$: [She$_{CT}$ will praise$_F$ him.]~
$Q_{6.2}$: What is the speaker's opinion?
$A_{6.2}$: [I$_{CT}$ will not$_F$ ~~praise him~~.]~

Figure 15: Final discourse tree with QUDs

(18) A_1: John and Mary are really proud. A_2: They worked hard. A_3: John painted a self-portrait A_4: and Mary rehearsed a piano sonata. A_5: Yesterday, I talked to John's mother. $A_{6.1}$ She will praise him. $A_{6.2}$: I won't ~~praise him~~.

As we said, each new assertion A_i is either processed against the existing discourse context (thereby determining the background as its given material, and its QUD Q_i as a congruent question which shares with A_i the same background) or might, alternatively, be processed in a forward-looking manner against some following assertion. In the latter case, the QUD and the background constituent are identified as the maximal common material of two parallel assertions.

After it has been determined, each Q_i is inserted as a node in the tree right above A_i; in the parallel case, the two (or sometimes more) parallel assertions $A_{i'}$ and $A_{i''}$ will become sibling nodes of the QUD node. Attachment is only possible at the *Right Frontier* (Asher and Lascarides, 2003), i.e. below any of the nodes at the right edge of the existing tree. The exact attachment site is determined based on the given information within Q_i.

This means, in particular, that any content in the discourse context that is not at the right edge does not count as given in the information-structural sense. The corresponding constraint is the following one:

Attachment constraint (Back-to-the-Roots)
A QUD (and its answers) must attach below any antecedent of its given content, and otherwise as high as possible.

The final tree analysis is shown in Figure 15.

6 Conclusions

In this paper, we have argued that QUDs constitute a vital part of discourse trees that allow us to jointly analyze the information structure and discourse structure of text. The procedure does not rely on the use of discourse relations. It remains to be seen whether the outcome of our analyses is generally comparable to the analyses from other approaches to discourse structure such as SDRT but, in any case, we think that a discourse should have precisely one discourse structure. Finally, by only referring to the semantic content – and not to particular morpho-syntactic or prosodic properties – of discourse, we argue that the procedure will also be applicable in a cross-linguistic setting.

Acknowledgements

This work was developed in the Stuttgart Sonderforschungsbereich 732 *Incremental Specification in Context*, Project A6 *Encoding of Information Structure in German and French*, funded by the German Science Foundation (DFG). We would like to thank Lisa Brunetti, Kordula De Kuthy and Cristel Portes for very helpful discussions and three anonymous reviewers for their criticism and suggestions.

References

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

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

David Beaver and Brady Clark. 2008. *Sense and Sensitivity. How Focus Determines Meaning*. Wiley & Sons, Chichester.

Lisa Brunetti. 2009. On the semantic and contextual factors that determine topic selection in italian and spanish. *The Linguistic Review*, 26(2-3):261–289.

Daniel Büring. 2003. On D-Trees, Beans, and B-Accents. *Linguistics & Philosophy*, 26(5):511–545.

Daniel Büring. 2008. What's New (and What's Given) in the Theory of Focus? In *Berkeley Linguistics Society*, pages 403–424.

Daniel Büring. in press. *Intonation and Meaning*. Oxford University Press.

Jonathan Ginzburg. 1996. Interrogatives: questions, facts and dialogue. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell, Oxford.

Talmy Givón. 1983. *Topic continuity in discourse: A quantitative cross-language study*. John Benjamins Publishing, Amsterdam.

Jeroen Groenendijk and Martin Stokhof. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Universiteit van Amsterdam.

Michael Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Irene Heim. 1991. Artikel und Definitheit. In A. von Stechow and D. Wunderlich, editors, *Semantik: ein internationales Handbuch der zeitgenössischen Forschung*, pages 487–535. de Gruyter, Berlin.

Hans Kamp. 1998. A DRT-based treatment of the focus-frame focus division and its presuppositions. Technical report, Universität Stuttgart. manuscript.

Wolfgang Klein and Christiane von Stutterheim. 1987. Quaestio und referentielle Bewegung in Erzählungen. *Linguistische Berichte*, 109:163–183.

Michael Kohlhase, Susanna Kuschert, and Manfred Pinkal. 1996. A Type-Theoretic Semantics for λ-DRT. In P. Dekker and M. Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*. University of Amsterdam.

Manfred Krifka. 1992. A Compositional Semantics for Multiple Focus Constructions. In J. Jacobs, editor, *Informationsstruktur und Grammatik*, pages 17–53. Westdeutscher Verlag, Opladen.

Manfred Krifka. 2008. Basic Notions of Information Structure. *Acta Linguistica Hungarica*, 55(3-4):243–276.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford University Press.

Tanya Reinhart. 1981. Pragmatics and linguistics: an analysis of sentence topics in pragmatics and philosophy I. *Philosophica*, 27(1):53–94.

Uwe Reyle. 1993. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10:123–179.

Arndt Riester and Stefan Baumann. 2013. Focus triggers and focus types from a corpus perspective. *Dialogue and Discourse*, 4(2).

Arndt Riester and Jörn Piontek. 2015. Anarchy in the NP. When new nouns get deaccented and given nouns don't. *Lingua*, 165(B):230–253.

Arndt Riester. 2015. Analyzing Questions under Discussion and information structure in a Balinese narrative. In *Proceedings of the 2nd Workshop on Information Structure of Austronesian Languages*, pages 1–26, Tokyo.

Craige Roberts. 2012. Information structure: towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69. Earlier ver. (1996) in OSU Working Papers in Linguistics, Vol. 49.

Mats Rooth. 1985. *Association with Focus*. Ph.D. thesis, University of Massachusetts, Amherst.

Mats Rooth. 1992. A Theory of Focus Interpretation. *Natural Language Semantics*, 1(1):75–116.

Roger Schwarzschild. 1999. GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What Projects and Why. In *Proceedings of SALT 20*, pages 309–327, Vancouver.

Maite Taboada and William Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8:423–459.

Arnim von Stechow. 1982. Structured Propositions. Technical report, Universität Konstanz. Arbeitspapiere des SFB 99.

Michael Wagner. 2012. Focus and givenness: a unified approach. In I. Kučerová and A. Neeleman, editors, *Contrasts and Positions in Information Structure*, pages 102–148. Cambridge University Press.

A Model for Attention-Driven Judgements in Type Theory with Records

Simon Dobnik
CLASP
University of Gothenburg, Sweden
simon.dobnik@gu.se

John D. Kelleher[*]
School of Computing
Dublin Institute of Technology, Ireland
john.d.kelleher@dit.ie

Abstract

This paper makes three contributions to the discussion on the applicability of Type Theory with Records (TTR) to embodied dialogue agents. First, it highlights the problem of type assignment or judgements in practical implementations which is resource intensive. Second, it presents a judgement control mechanism, which consists of grouping of types into clusters or states by their thematic relations and selection of types following two mechanisms inspired by the Load Theory of selective attention and cognitive control (Lavie et al., 2004), that addresses this problem. Third, it presents a computational framework, based on Bayesian inference, that offers a basis for future practical experimentation on the feasibility of the proposed approach.

1 Type Theory with Records

One of the central challenges for multi-modal dialogue systems is information fusion or how such a system can represent information from different domains, compare it, compose it, and reason about it. Typically, a situated agent will have to deal with information that comes from its perceptual sensors and will be represented as real-valued vectors and conceptual categories (some of which correspond to words in language) that are formed through cognitive processes in the brain. When situated agents are implemented practically one typically adopts a layered approach starting at the scene geometry and finishing at the level of the agent's knowledge about the objects and their interactions (Kruijff et al., 2006). Although, this approach may be good for practical reasons, for example there are pre-existing systems which may be organised in a pipeline this way, this also assumes that representations and operations are distinct at each level and one needs to design interfaces that would mediate between these levels.

Type Theory with Records (TTR) (Cooper, 2005; Cooper, 2012; Cooper et al., 2015) provides a theory of natural language semantics which views meaning and reference assignment being in the domain of an individual agent who can make *judgements* about situations (or invariances in the world) of being of types (written as $a : T$). The type inventory of an agent is not static but is continuously refined through agent's interaction with its physical environment and with other agents through dialogue interaction which provides instances and feedback on what strategies to adopt to learn from these instances. The reason why agent's meaning representations or type inventories converge to an approximately identical inventory is that agents are situated in the identical or sufficiently similar physical environment and have grounded conversations with other agents; see for example the work of (Steels and Belpaeme, 2005) and (Larsson, 2013) for an approach in TTR. Having the capability to adjust the type representations they can adapt to new physical environments and new conversational exchanges. Such view is not novel to mobile robotics (Dissanayake et al., 2001) nor to approaches to semantic and pragmatics of dialogue (Clark, 1996), but it is novel to formal semantics (Dowty et al., 1981; Blackburn and Bos, 2005) which represents important body of work on how meaning is constructed compositionally and reasoned about. Overall, we see TTR as a highly fitting framework for modelling cognitive situated agents as it connects perception and high level semantics of natural language and vice versa.

The type system in TTR is rich in comparison to that found in traditional formal semantics (entities, truth values and function types constructed from these and other function types). In addition types

---

[*] Both authors contributed equally.

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 25-34.

are used to model meaning in a proof-theoretic way rather than constraining model theoretic interpretation. Types in TTR can be either basic types such as *Ind* or *Real* or record types. Record types are represented as matrices containing label-value pairs where labels are constants and values can be either basic types, ptypes which act as type constructors and record types. The corresponding proof-objects of record types are records. These may be thought of as *iconic representations* of (Harnad, 1990) or sensory readings that an agent perceives as *sensory projections* of *objects* or *situations* in the world. The example below shows a judgement that a record (a matrix with = as a delimiter) containing a sensory reading is of a type (with : as a delimiter). The traditional distinction between symbolic and sub-symbolic knowledge is not maintained in this framework as both can be assigned appropriate types.

$$\begin{bmatrix} \text{a} & = & \text{ind}_{26} \\ \text{sr} & = & [[34,24],[56,78]\ldots] \\ \text{loc} & = & [45,78,0.34] \end{bmatrix} :$$
$$\begin{bmatrix} \text{a} & : & \text{Ind} \\ \text{sr} & : & \text{list(list(Real))} \\ \text{loc} & : & \text{list(Real)} \end{bmatrix}$$

An important notion of TTR is that types are *intensional* which means that a given situation in the world may be assigned more than one type. For example, a sensory reading of a particular situation in the world involving spatial arrangement of objects captured as records of types shown in the previous example may be assigned several record types of spatial relations simultaneously, for example *Left*, *Near*, *At*, *Behind*, etc. Another important notion of TTR is *sub-typing* which allows comparison of types. In addition to the type in the previous example, let's call it T_1, the record can also be assigned the following two types T_2 and T_3 whereby the following sub-typing relation between them holds: $T_1 \sqsubseteq T_2 \sqsubseteq T_3$ where $\sqsubseteq$ reads as "is a subtype of".

$$T_2 = \begin{bmatrix} \text{a} & : & \text{Ind} \\ \text{sr} & : & \text{list(list(Real))} \end{bmatrix}$$
$$T_3 = \begin{bmatrix} \text{a} & : & \text{Ind} \end{bmatrix}$$

Thirdly, types may be component types of other types, for example p(redicate)-type list(Real) is a component of the larger record type shown above. Finally, record types may also be dependent on other record types. A record type representing a geometric relation between two objects, for example *Left* is dependent on at least two record types shown in the previous example representing perceptual objects. The notion of dependent types is stronger than that of component types and is related to representing linguistic meaning. With missing information matching a component type an agent could still judge (with some error) that a situation is of that type whereas a judgement without a previous judgement of the dependent type would be impossible.

The rich type system of TTR and the relations between types give us a lot of flexibility in modelling natural language semantics in embodied dialogue agents. However, one practical problem that an application of TTR faces is how such an agent will cope with the an increasing number of types that it continuously acquires through learning and assign them effectively to every new situation it encounters given that such agent has limited processing resources. Since each type assignment involves a judgement (a probabilistic belief that something is of a type T) for each record of a situation an agent having an inventory of n types would make n judgements, a large proportion of which would yield very low or even zero probabilities as they will be irrelevant or very-little relevant for the current perceptual and conversational contexts. This is because due to the regularities in the world certain types would never be assigned or are very unlikely to be assigned in certain contexts.

(Hough and Purver, 2014) present a model where types are ordered in a lattice by $\sqsubseteq$ which drives incremental type checking for the purposes of resolution of incremental linguistic input or output which in itself is a different task to ours. The approach captures *taxonomic or categorial relations* encoded in types. As humans do not necessarily judge situations from most general to most specific or vice versa, the benefit of reducing judgements following taxonomic organisation of types would vary depending on the situation judged. Such knowledge would allow exclusion of judgements of sub-types of an incompatible type but agent's judgements could be further reduced if it were primed what to expect in its current state and its perceptual and conversational contexts by its knowledge about the world and the linguistic behaviour of its interlocutor captured in a model of *thematic relations*, that is spatial, temporal, causal or functional relations between individuals occuring in the same situations (Lin and Murphy, 2001;

Estes et al., 2011). Similarly, (Cooper, 2008) argue that agents organise their type inventory into resources that are employed and modified in different activities. If this is so, in addition to a reasoning mechanism on subtype relations humans must also rely on processes by which bundles of types are primed for in particular situational contexts. As a consequence agents will not need to check each situation (sensory reading in the form of a record) for every type in their inventory but only those that they are primed for. A property that such priming mechanism must take into account is that the more accurately an agent is primed by its contexts, the lower the uncertainty and hence the smaller the set of the types it is primed for.

In this paper we focus on the mechanisms that drive the discovery of thematic relations and propose a computational model how such relations are applied in interaction to prime an agent. The basic premise of the paper is that the mechanisms underpinning attention are fundamental to the control and priming of judgements in TTR. In Section 2, we introduce Load Theory of Attention. In Section 3, we present an account, based on Load Theory of Attention, of how two different kinds of TTR judgements can be controlled and primed in an agent. In Section 4, we introduce a mathematical framework that illustrates how an agent can maintain probability distributions over its cognitive states and types and use them in the priming process. In Section 5 we give a worked example of this framework priming an agent for judgements. Section 6 gives some remarks about its usability and presents our future work.

2 Load Theory of Attention

One of the major contributions of 20th century psychology has been the study and improved understanding of perceptual attention in humans. There is more than one type of attention mechanism. In particular, a distinction can be made between bottom-up attention and top-down attention processes. Bottom-up attention is automatic, task independent, not under conscious control and is attracted towards salient entities in the environment (e.g., moving object, singleton red objects, etc.). Top-down attention can be consciously directed by an agent and is dependent on the task they are carrying out as tasks will have different complexities. Sometimes top-down attention is described in terms of an agent being primed to respond to a mental-set of perceptual stimuli that are relevant to the task they are consciously carrying out.

Early research on attention was based on the concept of a structural single channel bottleneck in perceptual processing (Broadbent, 1958; Welford, 1967). The early orthodoxy of attention as a bottleneck within a single channel has been challenged by several researchers (e.g., (Allport et al., 1972)) and more recent models have viewed attention as a shared resource or capacity that can be spread across multiple tasks simultaneously. For example, in the (Kahneman, 1973) theory of attention and effort the attention capacity can be focused on an individual task or shared across multiple tasks and the more difficult a task is the more attention is required by that task. Furthermore, the allocation of attention across tasks can be flexibly updated as the agent changes their attention policy from one moment to the next.

An enduring question within attention research has been to understand the conditions under which the perception of task irrelevant distractors is prevented. Most of this research in the 60s, 70s, and 80s was framed in terms of the early-late debate which focused on whether the structural bottleneck that excluded distractors occurred early or late in perceptual processing. Some researchers argued that attention could exclude early perceptual processing of distractors (e.g., (Treisman, 1969)) while others argued that distractor objects were perceptually processed and attention only affected post perceptual processing – such as working memory and response selection (e.g., (Duncan, 1980)). The reason for such a protracted debate was that there was a lot of evidence to support both views. Results from some studies indicated that unattended information went unnoticed (supporting an early filter) and other studies indicated that distractors were perceptually processed and interfered with task response (supporting a late filter).

A well regarded recent model of attention is perceptual load theory (Lavie et al., 2004). The concept of perceptual load is difficult to define but can be characterised in terms of the number of items that are perceptually available (the more items, the higher the load) and the demands of the perceptual task (e.g., selecting an object based on type and colour is more demanding then selecting an object based solely on type). Perceptual load also involves defining what constitutes an item in a display: (Lavie et al., 2004) give the example that a string of letters can be considered one item (a word) or several items (letters) depending on the task. Perceptual load theory attempts to resolve the early-late debate using a model of attention that distinguishes between two mechanisms of selective attention: *perceptual selection* and *cognitive control*. Perceptual selection is a mechanism that excludes the perception of task irrelevant distractors under situations of high perceptual load; however, in situations of low perceptual load any spare capacity will spill over to the perception of distractor objects. The cognitive control mechanism is an active process that reduces the interference from perceived distractors on task response. It does so by actively maintaining the processing prioritisation of task relevant stimuli within the set of perceived stimuli.

3 Load Theory and type judgements

Agents learn types all the time by making generalisations of invariances in the world and information communicated to them through conversation (direct transferal of knowledge). However, in order to access the knowledge quickly and efficiently, they organise it in a certain way in memory. We propose a method of how an agent (i) organises its type inventory in memory and (ii) applies this type inventory using a model of attention that avoids the exponential problem of judgements it would have to make without prioritising its type checking. We turn to the second notion first, the priming of type judgements using a model of attention. Within this attention based account a distinction can be made between two types of judgements: (i) pre-attentive and (ii) task induced and context induced judgements.

3.1 Attention-driven judgements

Pre-attentive judgements are controlled by the perceptual selection mechanism of Load Theory. The result of a pre-attentive judgement is the introduction of a type into the working memory or information state in a dialogue model (Ginzburg and Fernández, 2010). Basic representations of visual environment (Ullman, 1984) such as segmentation of a visual scene into entities and background is an example of a pre-attentive judgement. At the very basic level these will be the iconic representations captured by agent's sensors (Harnad, 1990). *Task induced* and *context induced judgements* require conscious attention. As such, they are controlled by the cognitive control mechanisms of Load Theory. These judgements are applied to types that are in working memory and result in new types being introduced to working memory. Task induced and context induced judgements are primed by the types associated via memory with the current activities that the agent is currently engaged in and their physical location. For example, if an agent is making a cup of tea there are a default set of objects relevant to that task that the agent will carry out a visual search for and purposefully recognise (the kettle, tea bags, cups, etc.). The definition of a set of relevant types corresponding to these objects can be understood as priming a set of task induced judgements related to the recognition of these objects. Finally, context induced judgements can be understood as task related judgements that are not by default related with the task but that are extensions to this set and are caused by the agent's interactions with other agents and the physical context of the task. For example, while an agent is making a cup of tea another agent warns them to take care because the plate beside the kettle is very hot or the agent may inadvertently touch the plate and sense the heat on its own. The judgement relating to the interpretation of the utterance "the plate beside the kettle" or the sensing of and predicting the desired reaction to a hot surface can be understood as a context induced judgement. The utterance or the hot plate is not a part of the task but is introduced in the context in which the task is taking place.

This raises a question of what mechanisms define these classes of type judgements. Pre-attentive type judgements are the judgements that are fundamental to the agent's basic operation and the agent is continually making them in order to be able to cope with its internal states and the external environment. The types involved in these judgements are intimately linked to agent's biology and embodiment as they are the types of basic representations generated by the sensors of the agent. As such, there is a finite set of these types. The assignment of other types is governed by the attention model of Load Theory. Attention can be either introduced by the task (or agent internally) or the context (agent externally). In terms of knowledge representation there is no difference between the types of the activity of tea making and the types associated with handling of dangerously hot objects. The set of task and context induced types for which an agent is primed at any moment is defined by current pre-attentive judgements and the sequence of tasks and contexts the agent has been engaged so far. For example, given that the agent has previously been in the corridor coupled with new pre-attentive judgements could prime the agent to be attentive to types one typically judges in a kitchen. An agent learns through experience the types that are relevant in a particular task and context. Practically, this amounts to finding associations between types in agent's memory and their evolution over time.

3.2 Cognitive states

Thematic relations are relations between objects, events, people and other entities that co-occur or interact together in space and time (Lin and Murphy, 2001; Estes et al., 2011). Inspired by the concept of a thematic relation we propose that an agent's type inventory is organised as a set of cognitive states, where each state defines a set of types that are related by a thematic relationship. A cognitive state may be the cognitive correlate of the agent intentionally performing a task but may also be a non-explicit cognitive state of an agent generally being in a situation or having a disposition. Importantly, we don't believe that an agent has conscious access to all its cognitive states nor can all states be directly mapped to concrete activities. Rather a cognitive state can be understood as a sensitivity towards certain types of objects, events, and situations where this sensitivity mapping between states and entities has been learned from experience. For example, there may be a cognitive state associated with the agent's basic existence and its wish to continue existing, or of being a parent, or of being in a concert hall, or of being involved in a conversation about playing a trumpet, or of making a cup of tea. The commonality across this disparate set is the fact that it is possible to list a set of types that are relevant to each state which represents agent's resources in terms of (Cooper, 2008). For example, the very fact of an agent's existence makes it sensitive to entities in the environment that endanger it (large things moving towards it at speed) or help its existence (food nearby). The cognitive state related to being in a concert hall might prime the agent to make judgements about the music, the instruments and the conductor. The state of participating in a conversation about playing a trumpet prime judgements relating to the body language of your interlocutor or the relationship you have with your interlocutor (are they an experienced player or an observer). Finally, the state of making tea could prime an agent to make judgements relating to kettles and cups and their arrangement in space.

It might appear that our approach simply pushes the intractability of judgements over the set of combinatorially exploding types onto the intractability over a set of cognitive states. We argue, however, that not withstanding the apparent complexity of human inner life there are in fact relatively restricted number of cognitive states that a human or an agent trying to live like a human needs to maintain in the course of an average day. While theoretically there could be as many states as the number of type judgements discussed earlier it is important to note that these states are built by an agent bottom up when an agent discovers new situations. Since an agent will be constrained by the environment in which it operates and since it can only discover a finite set of situations in its life, and since it is equipped with learning mechanisms with a strong bias to make generalisations it will only build a subset of these states that can be managed by its memory.

Important features of states and types include: (a) an agent may be in several states at the same time (they may be making tea and talking about music), and (b) a type may be associated with more then one state. While an agent is in a state or states performing any additional type judgements associated with one of the states incrementally reduces its ambiguity of being in several states.

4 A computational model

There are three requirements for our computational model: (i) agents clusters types according to thematic relations into several states, (ii) types are associated with each state with a certain probability, and (iii) a particular type may be associated with more than one state. Thematic relations between types are expressed by the co-occurrence of types in states. There are several computational mechanisms that could be used to automatically create states (clusters of types) with the above properties from data. For example, all three requirements are satisfied by *Latent Dirichlet Allocation (LDA)* which is a popular approach to topic modelling (Blei et al., 2003), the analogy between topic modelling and our scenario being that a topic is similar to a state and the association of a word with a topic is equivalent to the association of a type with a state. A drawback with LDA is that the number of topics or states must be known a priori. However, *Hierarchical Dirichlet process* (Teh et al., 2006) is an extension of the LDA where the number of topics is also learned.

Given that an agent has learned thematic relations between types in the form of their association to states, the control problem which it is facing is that it cannot know which state it is in and consequently it cannot decide what is the optimal collection of types to be primed for in making judgements. As a result the agent must try to infer the best sets of types to prime for by estimating:

1. a posterior distribution over the possible states (and, updating this distribution as it receives observations from the world and makes judgements about the world)
2. make a decision regarding which judgements to be primed to make based on the updated probability distribution.

The posterior probability of being in a particular state at time t is dependent on the previous states at time $t-1$ (i.e., the Markov property holds: conditioned on the present the future is independent of the past), the task and context judgements the agent has made following the priming in the previous $t-n$ states where n is the length of history an agent keeps, and the new pre-attentive judgements which may reflect perceptual change in its world. So, the posterior probability of each of the cognitive states of the agent can be computed as follows:

$$\begin{aligned}
&P(s_t|Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1}) = \\
&\quad \eta \times P(Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1}|s_t) \\
&\quad \times P(s_t)
\end{aligned}$$

where s_t is a state at time t, AS_{t-1} is the set of *active states*[1] at time $t-1$, Pre_t is the set of new pre-attentive judgements the agent has just made, $Task_{t-1}$ is the set of task relevant judgements the agent has made following previous priming, and *Cont* is the set of contextual judgements the agent has made following previous priming, and η denotes a normalisation process that ensures that the total probability mass of the posterior distribution sums to 1. We argue that the probabilities on the right hand side of this equation can be learned from experience. This learning process can be simplified if we assume conditional independence between Pre_t, $Task_{t-1}$, $Cont_{t-1}$ and AS_{t-1} given s_t, essentially adopting the same formulation for calculating poster probabilities as is used by a standard naive Bayes' classifier:

$$\begin{aligned}
&P(s_t|Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1}) = \\
&\quad \eta \times P(Pre_t|s_t) \times P(Task_{t-1}|s_t) \\
&\quad \times P(Cont_{t-1}|s_t) \times P(AS_{t-1}|s_t) \\
&\quad \times P(s_t)
\end{aligned} \tag{1}$$

Once we have computed a posterior probability over the set of states an agent has we need a mechanism that explains how this distribution informs the process of priming types. The simplest mechanism would be to select the state with the *maximum a posteriori* probability and then load into working memory the set of types that are associated with this state. This approach has the advantage of being computationally simple. However, it has the disadvantage that the agent assumes that they are only ever in one state, and, furthermore, if two or more states have a high posterior probability there is the possibility that the agent will keep switching between these states from one moment to the next. An alternative approach that is less susceptible to switching between states is to:

1. use the posterior probability over the states to rank and prune the state set, (the states that are not pruned are the *active states*)
2. renormalise the probability distribution over the set of active states,
3. compute a posterior probability over the set of types associated with active states using a *Bayes optimal classifier*,
4. using the posterior probability over types, rank and prune the set of types and load the set of unpruned types into working memory.

In order to rank and prune the state set we simply order the states based on their posterior probabilities and remove all the states that have a probability below a predefined threshold. This rank and pruning approach essentially implements a process whereby an agent can recognise what is not relevant to the current situation. Renormalising the probability distribution over the remaining

---

[1] We will define the set of active states later.

states is a simple process of summing the probability mass of the unpruned states and then dividing the probability mass of each state by this sum. The posterior probability of a type at time t is calculated using a Bayes optimal classifier as follows:

$$\begin{aligned} & P(type_t|Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1}) = \\ & \sum_{s \in AS_t} P(type|s) \times \\ & P(s|Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1}) \end{aligned} \tag{2}$$

where $type_t$ denotes a type at time t, AS_t denotes the set of unpruned (active) states at time t, $P(s|Pre, Task, Cont, AS_{t-1})$ denotes the probability of an active state s after the state set has been pruned and the posterior probability over the active states has been renormalised, and Pre_t, $Task_{t-1}$, $Cont_{t-1}$, and AS_{t-1} have the same meanings as defined above. Using a Bayes' optimal approach to calculating the posterior distribution over the types associated with the active states is computationally expensive because it includes a summation across the set of active states. However, the size of this set can be restricted based on the pruning criteria used so the computational cost of this summation operation can be minimised. Some of the benefits, however, of using a Bayes' optimal formulation are that: (a) this process explicitly recognises the fact that more then one state may be active at one point, (b) it also recognises the fact that a type may be associated with more then one state and that the strength of association between the type and a state is probabilistic ($P(type|s)$), and (c) this formulation is robust to small variations in the posterior distribution over states (i.e., when the state with the *maximum a posteriori* probability changes the system is stable—in terms of the types that are loaded into memory—if the changes across the distribution are stable). Once the posterior distribution over the types has been calculated the types can be ranked and pruned in a similar fashion to the states. This means that we need two thresholds for pruning, one for pruning the states and one for pruning the types. The ranking and pruning across the states and the types both reflect the attention based approach we have taken to this work modelled by Load Theory. When the cognitive load on the agent is low the pruning of states and types can be relaxed and when the cognitive load from the perceptual selection is high the pruning can become more severe.

5 Worked example

In this section we present a worked example that illustrates how an agent interacting in and moving around an environment can use the proposed models to prime the set of types judgements it has loaded in its memory. This example assumes that the agent has already learned a number of types and has already associated these types with the cognitive states it has constructed over the course of its lifetime.

To begin we will assume our agent has three cognitive states: $\mathcal{S}1$, $\mathcal{S}2$, $\mathcal{S}3$ and the prior probabilities of these states are $< 0.4, 0.3, 0.3 >$ respectively. Furthermore, the state transition matrix is a right stochastic matrix with i rows and j columns where each cell defines the probability of going from state i to state j in one time step (i.e., each cell defines $P(\mathcal{S}j|\mathcal{S}i)$) and is defined as follows:

$$\begin{array}{cccc} & \mathcal{S}1 & \mathcal{S}2 & \mathcal{S}3 \\ \mathcal{S}1 & 0.7 & 0.2 & 0.1 \\ \mathcal{S}2 & 0.3 & 0.4 & 0.3 \\ \mathcal{S}3 & 0.1 & 0.2 & 0.7 \end{array}$$

We will assume that there are 3 different preattentive types that the agent can assign to low-level perceptual features. For labelling convenience let us assume that these features are biased to particular locations in a building so that we can name these types after these locations, namely: OFFICE, CORRIDOR, and KITCHEN (cf. *semantic labelling of places* (Martínez Mozos et al., 2007)). We will also assume that the agent knows three task/contextual types[2]. We are interested in constructing agents that can participate in dialogues so we have decided that these types include types assigned to utterances in dialogue that the agent can engage in; for example, this agent can take part in dialogues relating to WEATHER, MACHINE-LEARNING, or the general WELL-BEING of someone. According to our model the agent should have learnt a probabilistic relationship between each of these types and its own cognitive states. The following right stochastic matrix defines the probabilistic relationship between each of the preattentive types and the state (i.e., each cell defines $P(type|\mathcal{S}i)$):

---

[2] For the purposes of the example the distinction between task and contextual types is moot.

	OFFICE	CORRIDOR	KITCHEN
$\mathcal{S}1$	0.7	0.2	0.1
$\mathcal{S}2$	0.1	0.8	0.1
$\mathcal{S}3$	0.05	0.15	0.8

And, the following matrix defines the probabilistic relationships between each of the task/context types and the states:

	WEATHER	MACH.-LEARN.	WELL-BEING
$\mathcal{S}1$	0.1	0.7	0.2
$\mathcal{S}2$	0.4	0.1	0.5
$\mathcal{S}3$	0.6	0.3	0.1

We also need to define two attention thresholds: one threshold is used to define the set of active states and the other is used to define the set of active types. Unlike the probabilities defined above (which are relatively fixed and are updated via a separate learning process) these attention thresholds may change from moment to moment and are dependent on the cognitive load the agent is experiencing: high load and the thresholds are low, low load and the threshold are high. For this example, we will assume that the agent is under a moderately high load and that both of these thresholds are set to: 0.3.

To begin calculating the set of types that the agent is primed for at time step t we need information relating to: (a) the set of active states at $t-1$ (AS_{t-1}); (b) the set of task and context type judgements the agent made at time $t-1$ (T_{t-1}); and, (c) the set of pre-attentive judgements the agent has just made at time t (Pre_t). For this example we will assume the following: $AS_{t-1} = \{\mathcal{S}1, \mathcal{S}2, \mathcal{S}3\}$, $T_{t-1} = \{\text{MACHINE-LEARNING}, \text{WEATHER}\}$, and $Pre_t = \{\text{OFFICE}, \text{CORRIDOR}\}$.

Our first step is to calculate the probability distribution over the states at time t. We do this using Equation 1. Before we apply Equation 1 we need to calculate the probability distribution for $P(AS_{t-1}|\mathcal{S}_t)$. We can calculate these probabilities using the prior probabilities for the states and the transition matrix $P(\mathcal{S}_j|\mathcal{S}_i)$ and applying Bayes' Theorem. The resulting probabilities (rounded to 4 places) are as follows:

	$\mathcal{AS}1_{t-1}$	$\mathcal{AS}2_{t-1}$	$\mathcal{AS}3_{t-1}$
$\mathcal{S}1_t$	0.7000	0.2250	0.0750
$\mathcal{S}2_t$	0.3077	0.4615	0.2308
$\mathcal{S}3_t$	0.1176	0.2647	0.6176

Once we have these probabilities it is a relatively straightforward process to calculate $P(\mathcal{S}_t|Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1})$ using Equation 1. One technical point is that for each factor on the right hand-side of the equation ($P(Pre_t|s_t)$, $P(Task_{t-1}|s_t)$, $P(Cont_{t-1}|s_t)$ and $P(AS_{t-1}|s_t)$) we assume conditional independence between the conditioned events given the evidence. For example, for each $\mathcal{S}i_t$ we calculate $P(\mathcal{AS}_{t-1}|\mathcal{S}i_t)$ as:

$$P(\mathcal{AS}_{t-1}|\mathcal{S}i_t) = P(\mathcal{AS}1_{t-1}|\mathcal{S}i_t) \\ \times P(\mathcal{AS}2_{t-1}|\mathcal{S}i_t) \times P(\mathcal{AS}3_{t-1}|\mathcal{S}i_t)$$

In this context the posterior probability over the states (rounded to 4 places of decimal) is:

$$\mathbf{P}(\mathcal{S}_t|Pre_t, Task_{t-1}, Cont_{t-1}, AS_{t-1}) = \\ < 0.5412, 0.3677, 0.0911 >$$

Applying the attention threshold to this distribution over the states the set of active states at time t is then $\{\mathcal{S}1, \mathcal{S}2\}$ and renormalising the probability mass over these states gives us a probability distribution (again rounded to 4 places of decimal) of $< 0.5954, 0.4046 >$. We can now use Equation 2 to calculate the posterior probabilities over the task and contextual types. We only do this calculation for types that are associated (i.e., $P(type|\mathcal{AS}i) > 0$) with at least 1 of the active states. In this instance all three of the task/context types (WEATHER, MACHINE-LEARNING, and WELL-BEING) are associated with at least 1 of the active states so we calculate the posterior probability for all three of these types. The posterior probabilities over the types are $< 0.2214, 0.4573, 0.3214 >$. Applying the type attention threshold of 0.3 to this distribution there are two types that are active MACHINE-LEARNING, and WELL-BEING and the agent will be primed to make judgements of these types at time t.

In this example, we only pruned one of the task/context types from the primed list. However, as the number of types grows (remember that types will represent concepts at different levels of abstraction) and the number of states also grows then the number of types that are pruned will also grow.

6 Conclusion and future work

In this paper we present a computational mechanism for attention-driven type judgements in an interacting agent that is inspired by cognitive processes in humans such as discovery of thematic relations and sharing of cognitive resources between perceptual selection and cognitive control as proposed in Load Theory. It is important to note that the problem of multiple type assignments or judgements is not exclusive to TTR but is a general problem where a cognitive agent has to make numerous classifications based on limited computational resources. In robotics this task is known as *visual search* (Sjöö, 2011; Kunze et al., 2014). The proposed application of TTR allows us to formulate a cognitively-inspired computational model for visual search. The approach is also relevant to computational modelling of situated dialogue. Being primed for particular types would disambiguate interlocutors utterances based on the previous type judgements and perceptual observations. In dialogue generation it allows priming of the agent to particular topics and therefore can be used for topic modelling of a dialogue system.

The model proposes that an agent has a set of cognitive states that they have learned from past experience. An agent may be in more than one cognitive state at any one time. There are a set of types associated with each cognitive state of an agent. When a cognitive state is active (unpruned) an agent is primed to make judgements relating to the types associated with the state. This is why our account links judgement in TTR and attention. The difficulty with this account is that because more than one cognitive state is active at any one time the agent must decide which of the active cognitive states it should prime for observation. The solution is that the agent should maintain a distribution over its cognitive states and prime its observation relative to the types associated with the cognitive states with high probability. Following Load Theory the agent will actually perceive as many of these primed types as it can before its perceptual capacity is exhausted and it will then select a subset of these primed types for further cognitive processing.

In our forthcoming work we are working towards a computational application of the model to situated dialogue. We are particularly interested in evaluating the benefits of an agent being primed this way in comparison to when it has no priming at all. The model introduces several parameters, for example the number of states, the number of types, the size of memory for pre-attentive judgements and task and context related judgements whose effects on system performance will also be investigated.

References

D. Alan Allport, Barbara Antonis, and Patricia Reynolds. 1972. On the division of attention: A disproof of the single channel hypothesis. *The Quarterly journal of experimental psychology*, 24(2):225–235.

Patrick Blackburn and Johan Bos. 2005. *Representation and inference for natural language. A first course in computational semantics*. CSLI Publications.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning research (JMLR)*, 3:993–1022.

D.E. Broadbent. 1958. *Perception and Communication*. Pergamon Press.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

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, November.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2):333–362.

Robin Cooper. 2008. Type theory with records and unification-based grammar. In Fritz Hamm and Stephan Kepser, editors, *Logics for Linguistic Structures. Festschrift for Uwe Mönnich*, volume 201, page 9. Walter de Gruyter.

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 of *General editors: Dov M Gabbay, Paul Thagard and John Woods*. Elsevier BV.

M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

David R Dowty, Robert Eugene Wall, and Stanley Peters. 1981. *Introduction to Montague semantics*. D. Reidel Pub. Co., Dordrecht, Holland.

John Duncan. 1980. The locus of interference in the perception of simultaneous stimuli. *Psychological review*, 87(3):272–300.

Zachary Estes, Sabrina Golonka, and Lara L Jones. 2011. Thematic thinking: The apprehension and consequences of thematic relations. In Brian Ross, editor, *The Psychology of Learning and Motivation*, volume 54, pages 249–294. Burlington: Academic Press.

Jonathan Ginzburg and Raquel Fernández. 2010. Computational models of dialogue. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The handbook of computational linguistics and natural language processing*, Blackwell handbooks in linguistics, pages 429–481. Wiley-Blackwell, Chichester, United Kingdom.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1–3):335–346, June.

Julian Hough and Matthew Purver. 2014. Probabilistic type theory for incremental dialogue processing. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88, Gothenburg, Sweden, April. Association for Computational Linguistics.

Daniel Kahneman. 1973. *Attention and effort*. Prentice-Hall, Englewood Cliffs, N.J.

Geert-Jan M. Kruijff, John D. Kelleher, and Nick Hawes. 2006. Information fusion for visual reference resolution in dynamic situated dialogue. In *Perception and Interactive Technologies*. Springer.

Lars Kunze, Chris Burbridge, and Nick Hawes. 2014. Bootstrapping probabilistic models of qualitative spatial relations for active visual object search. In *AAAI Spring Symposium 2014 on Qualitative Representations for Robots*, Stanford University in Palo Alto, California, US, March, 24–26.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, online:1–35, December 18.

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

Emilie L. Lin and Gregory L. Murphy. 2001. Thematic relations in adults' concepts. *Journal of experimental psychology: General*, 130(1):3–28.

Óscar Martínez Mozos, Rudolph Triebel, Patric Jensfelt, Axel Rottmann, and Wolfram Burgard. 2007. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems*, 55(5):391–402, 5.

Kristoffer Sjöö. 2011. *Functional understanding of space: Representing spatial knowledge using concepts grounded in an agent's purpose*. Ph.D. thesis, KTH, Computer Vision and Active Perception (CVAP), Centre for Autonomous Systems (CAS), Stockholm, Sweden.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Anne M. Treisman. 1969. Strategies and models of selective attention. *Psychological review*, 76(3):282–299.

Shimon Ullman. 1984. Visual routines. *Cognition*, 18(1–3):97–159.

A.T. Welford. 1967. Single-channel operation in the brain. *Acta Psychologica*, 27:5–22.

Grounding, Justification, Adaptation: Towards Machines That Mean What They Say

David Schlangen
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
`david.schlangen@uni-bielefeld.de`

Abstract

Meaningful language use rests on the *grounding* of the language in the non-linguistic world and in the practices of language users. This grounding is built up and maintained in interaction, through *Conversational Grounding*, which is the interactive process with which interlocutors build mutual understanding; *Justification*, the ability to explain and provide reasons for one's language use; and *Adaptation*, the ability to accept corrections and adapt future language use accordingly. We outline a model of grounded semantics that combines perceptual knowledge (how to visually identify potential referents of terms; realised as classifiers taking visual information as input) and taxonomic knowledge (covering lexical relations such as hyponymy and hypernymy), and we sketch a proof-of-concept implementation of a dialogue system that realises the interactional skills that ground this knowledge.

1 Introduction

Computer systems that process natural language input and produce natural language output are becoming ever more common and ever more capable. So-called "intelligent personal assistants" built into mobile phones are already serving real customer needs (e.g., providing verbal access to the user's calendar), and current research systems show impressive results on tasks like image captioning (given an image, produce a textual description of its content). And yet, there is a strong sense in which these system do not *mean* anything with their use of natural language. Why is that so?

We propose that meaningful language use rests on the *grounding* of the language: in the non-linguistic world; in itself, among the parts of the language; and in the practices of the community of language users. These are, at a least to a certain degree, complementary aspects, as Hilary Putnam (1973) pointed out with the claim that someone who (like him) cannot reliably tell an elm from a beech tree would still mean the same with *elm* as someone who can. Marconi (1997) uses this observation to motivate a model of what he calls *lexical competence* that separates *referential competence*—the competence to identify actual referents, which Putnam claims to lack with respect to elms—and *inferential competence*, which uses semantic knowledge to place meanings in relation to other meanings (here, for example, the relation of hyponymy between *elm* and *tree*).

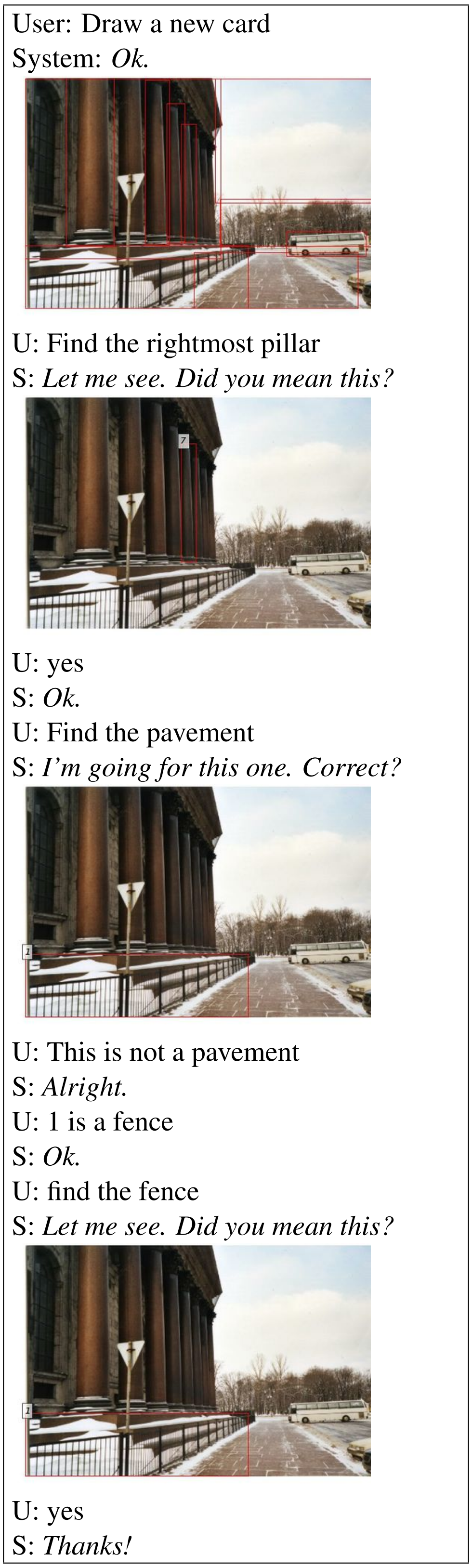
This grounding is not static, however, but rather is built up and maintained in interaction, through *Conversational Grounding*, which is the interactive process with which interlocutors build mutual understanding; *Justification*, the ability to explain and provide reasons for one's language use; *Adaptation*, the ability to accept corrections and adapt future language use accordingly.

Our aim in this paper is to outline a model of semantic competence that can address these desiderata: That it explains what kind of discriminatory power constrains meaningful language use, and that this power is acquired, defended and adapted in interaction. Its basis is a "two dimensional" model of lexical knowledge. In this model, one dimension captures '*know-how*' such as the knowledge required to pick out the intended referent in a visually presented scene, and the other captures more semantic knowledge ('*know-that*') that enables inferences, but can also, as we show, support visual reference resolution. (This distinction is inspired by that between *referential* and *inferential* lexical competence made by Marconi (1997), but further generalised. The visual-grounding model builds on (Kennington and Schlangen, 2015). See discussion below.) Both kinds of knowledge can

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 35-43.

be trained from interaction data. The lexical representations are used to compose meanings of larger phrases. This composition process is transparent (compared to composition in distributional semantics, as discussed below), and hence is accessible for inspection and correction.

To make our proposal concrete, and to investigate the utility of interaction that has the system's own semantic competence as its topic, we implemented an interactive system that tries to resolve references to objects in images and can accept corrections and provide explanations. Figure 1 shows an excerpt of an actual interaction with the system of a naive first-time user. (Image sources are credited in Section 7.1 below.)



User: Draw a new card
System: *Ok.*

U: Find the rightmost pillar
S: *Let me see. Did you mean this?*

U: yes
S: *Ok.*
U: Find the pavement
S: *I'm going for this one. Correct?*

U: This is not a pavement
S: *Alright.*
U: 1 is a fence
S: *Ok.*
U: find the fence
S: *Let me see. Did you mean this?*

U: yes
S: *Thanks!*

Figure 1: **Example Interaction** *Reference candidates outlined in red and without label; selected candidates with numeric label. (Best viewed in colour.)*

2 Related Work

The idea of connecting words to what they denote in the real world via perceptual features goes back at least to Harnad (1990), who coined "The Symbol Grounding Problem": "*[H]ow can the semantic interpretation of a formal symbol system be made* intrinsic *to the system, rather than just parasitic on the meanings in our heads?*" The proposed solution was to link 'categorial representations' with "learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections".

This suggestion has variously been taken up in computational work. An early example is Deb Roy's work from the early 2000s (Roy et al., 2002; Roy, 2002; Roy, 2005). In (Roy et al., 2002), computer vision techniques are used to detect object boundaries in a video feed, and to compute colour features (mean colour pixel value), positional features, and features encoding the relative spatial configuration of objects. These features are then associated in a learning process with certain words, resulting in an association of colour features with colour words, spatial features with prepositions, etc., and based on this, these words can be interpreted with reference to the scene currently presented to the video feed.

Of more recent work, that of Matuszek et al. (2012) is closely related to the approach we take. The task in this work is to compute (sets of) referents, given a (depth) image of a scene containing simple geometric shapes and a natural language expression. In keeping with the formal semantics tradition, a layer of logical form representation is assumed; it is not constructed via syntactic parsing rules, however, but by a learned mapping (*semantic parsing*). The non-logical constants of this representation then are interpreted by linking them to classifiers that work on perceptual features (representing shape and colour of objects). Interestingly, both mapping processes are trained jointly, and hence the links between classifiers and non-logical constants on the one hand, and non-logical constants and lexemes on the other are induced from data. In the work presented here, we take a simpler approach that directly links lexemes and perceptions, but does not yet learn the composition.

Most closely related on the formal side is recent work by Larsson (2015), which offers a very direct implementation of the 'words as classifiers' idea (couched in terms of type theory with records (TTR; (Cooper and Ginzburg, 2015)) and not model-theoretic semantics). In this approach, some lexical entries are enriched with classifiers that can judge, given a representation of an object, how applicable the term is to it. The paper also describes how these classifiers could be trained (or adapted) in interaction. The model is only specified theoretically, however, with hand-crafted classifiers for a small set of words, and not tested with real data. More generally, the claim that the ability to negotiate meaning is an important component of the competence of meaningful language use, which we also make here, has been forcefully argued for by Larsson and colleagues (Cooper and Larsson, 2009; Larsson, 2010; Fernández et al., 2011). (See also DeVault et al. (2006), who call this process *societal grounding* and outline a formal computational model of it.)

The second "dimension" in our semantic representations concerns language-to-language grounding. To explain within the framework of formal semantics how some statements can be necessarily true by virtue of meaning and not logical tautology (e.g., "bachelors are unmarried"), Carnap (1952) introduced *meaning postulates*, which are axioms that explicitly state connections between non-logical constants (e.g., $\forall x.bachelor(x) \rightarrow \neg married(x)$). The computational resource WORDNET (Fellbaum, 1998) can be seen as a large-scale realisation of this concept. It is a large database of word *senses*, different meanings that a word can have. Further semantic relations structure this lexicon (*antonymy, hyponomy, hypernymy, meronymy*). As described below, we use it as a starting point for encoding language-to-language grounding, together with the more directly perception-oriented feature norms of Silberer et al. (2013), which encode typical attributes ("is brown", "has feet") for about 500 concepts.

In the present work, our focus is on acquiring and using referential competence. On the ontological side, for now we simply use pre-compiled taxonomic/ontological resources. Methods exists for automating the construction of such resources (e.g., (Mitchell et al., 2015; Ganitkevitch et al., 2013)), some even using dialogue (Hixon et al., 2015). As another type of method, distributional semantics has recently become popular for the unsupervised acquisition of lexical relations (Turney and Pantel, 2010; Mikolov et al., 2013), particularly of the (typically rather vaguely specified) relation of 'similarity'. We will investigate the applicability of these methods in future work, but for now make use of the greater expressiveness and explicitness of more logic-inspired representations as used in WORDNET.

3 Overview of the Model

As stated in the introduction, a desideratum for the model is that it explains what kind of discriminatory power constrains meaningful language use, and how this power is acquired, defended and adapted in interaction. To make this more concrete (and to move from Putnam's tree example to a different biological kingdom), what we want to achieve is that our model can capture the knowledge required to deal satisfactorily both with (1-a) and (1-b).

(1) a. Find the Rottweiler in the picture.
b. Peter walked past a Rottweiler. The dog was barking ferociously.

But what is this knowledge? For (1-a), this must be information connected with the visual appearance of the object that is to be identified; for (1-b), the required knowledge is that a Rottweiler is a type of dog, and hence that the definite noun phrase in the second sentence can refer to the object introduced into the discourse with the indefinite in the first. These types of knowledge can interact: We'd still be satisfied if, when presented with an image containing one Rottweiler and, say, five cats, the addresse points at the Rottweiler, even if they don't actually know what distinguishes Rottweilers from other breeds of dog and all they knew was what visually distinguishes dogs from cats.

We take the basic idea from Marconi (1997) that there is a categorical difference between these types of knowledge. Marconi (1997) labels these aspects of lexical competence *referential* and *inferential*. While our focus in the work presented here is also on reference, we would argue that the distinction is more generally one between *know-how* and *know-that*, with the former covering the knowledge involved in executing actions ("cycling", "drawing an elephant") as well, and we will refer to the types with these labels. These "two dimensional" lexical semantic representations then must be composed into representations of phrases, where the composition process as well as what went into it must be open to justification and critique in interaction. We address these parts of the model in turn.

4 Two-Dimensional Lexical Semantics

(2) sketches the lexical entry for 'Rottweiler' with its two basic components, "know-how/referential" and "know-that/ontological", as it will be explained in the following.

$$(2)\quad \begin{bmatrix} \texttt{Rottweiler} \\ \textit{kh/ref} : \lambda \mathbf{x}.f_{rt}(\mathbf{x}) \\ \textit{kt/ont} : wn.hyponym, wn.hypernym, etc. \end{bmatrix}$$

4.1 Visual/Referential know-how

We follow Kennington and Schlangen (2015) and represent (and learn) visual-referential knowledge as classifiers on perceptual input. We briefly review their model here.

Let w be a word whose meaning is to be modelled, and let $\mathbf{x}$ be a representation of an object in terms of its visual features. The core ingredient then is a classifier that takes this representation and returns a score $f_w(\mathbf{x})$, indicating the "appropriateness" of the word for denoting the object. In (Kennington and Schlangen, 2015) and below, the classifier is a binary logistic regression and the score can be interpreted as a probability. Training of the classifier will be explained below.

Noting a (loose) correspondence to Montague's (1974) intensional semantics, where the intension of a word is a function from possible worlds to extensions (Gamut, 1991), the *intensional* meaning of w is then defined as the classifier itself, a function from a representation of an object to an "appropriateness score":[1]

$$[\![w]\!]_{obj} = \lambda \mathbf{x}.f_w(\mathbf{x}) \qquad (1)$$

(Where $[\![.]\!]$ is a function returning the meaning of its argument, and $\mathbf{x}$ is of the type of feature given by f_{obj}, the function computing a feature representation for a given object.)

The *extension* of a word in a given (here, visual) discourse universe W can then be modelled as a probability distribution ranging over all candidate objects in the given domain, resulting from the application of the word intension to each object ($\mathbf{x}_i$ is the feature vector for object i, $normalize()$ vectorized normalisation, and I a random variable ranging over the k candidates):

$$\begin{aligned} &[\![w]\!]^{W}_{obj} = \\ &\quad normalize(([\![w]\!]_{obj}(\mathbf{x}_1), \ldots, [\![w]\!]_{obj}(\mathbf{x}_k))) = \\ &\quad normalize((f_w(\mathbf{x}_1), \ldots, f_w(\mathbf{x}_k))) = P(I|w) \end{aligned} \qquad (2)$$

4.2 Taxonomic/Ontological know-that

As mentioned above, for now we use pre-existing resources as source of the initial ontological knowledge. There is some selection of available sources besides WORDNET (e.g., Freebase (Bollacker et al., 2008) and ConceptNet[2]), but we start with the former, as it is well-curated and stable. It provides us mostly with hypernomy (or "*is a*") relations. Notoriously, these can contain rather arcane categories; (3) shows this information for the lexical entry for "Rottweiler" with the less common categories (such as *placental* or *chordate*) left out.

$$(3)\quad \begin{bmatrix} \texttt{Rottweiler} \\ \textit{kt/ont/hyp} : \text{shepherd\_dog|working\_dog|dog|...} \end{bmatrix}$$

An additional, but with 509 entries compared to the over 200k entries of WORDNET much smaller information resource is the set of feature norms of McRae et al. (2005), a collection of attributes typically associated with a given object. (We use the version prepared by Silberer et al. (2013), which is filtered for being backed up with visual evidence.)

This resource does not contain an entry for *Rottweiler*, but one for *dog*, which is shown in (4).

---

[1] (Larsson, 2015) develops this intension/extension distinction in more detail for his formalisation.

[2] `http://conceptnet5.media.mit.edu`

$$
(4)\qquad \begin{bmatrix} \texttt{dog} \\ \textit{kt/ont/isa} : \text{animal}|\text{mammal} \\ \textit{kt/ont/properties} : \\ \begin{bmatrix} \textit{anatomy/has} : \text{mouth, head, whiskers,} \\ \text{claws, jaws, neck, snout, tail, 4\_legs, teeth,} \\ \text{eyes, nose, fur, ears, paws, feet, tongue} \\ \textit{behaviour} : \text{walks, runs, eats} \\ \textit{colour\_patterns} : \text{grey, black, brown, white} \\ \textit{diet} : \text{drinks\_water} \end{bmatrix} \end{bmatrix}
$$

We have explored two other kinds of automatically acquired lexical relations, but postpone their description until we have described the data sets that we used for our implementation.

5 Composition

5.1 Visual/Referential know-how

In the Kennington and Schlangen (2015) approach, composition of visual word meanings into phrase meanings is governed by rules that are tied to syntactic constructions. In the following, we only use simple multiplicative composition for nominal constructions:

$$
[\![[_{nom} w_1, \ldots, w_k]]\!]^W = [\![\text{NOM}]\!]^W [\![w_1, \ldots, w_k]\!]^W = \circ_{/N} \left([\![w_1]\!]^W, \ldots, [\![w_k]\!]^W\right) \qquad (3)
$$

where $\circ_{/N}$ is defined as

$$
\circ_{/N} \left([\![w_1]\!]^W, \ldots, [\![w_k]\!]^W\right) = P_\circ(I|w_1, \ldots, w_k)
$$

$$
\text{with } P_\circ(I = i|w_1, \ldots, w_k) = \frac{1}{Z}(P(I = i|w_1) * \cdots * P(I = i|w_k)) \text{ for } i \in I \qquad (4)
$$

(Z takes care that the result is normalized over all candidate objects.)

To arrive at the desired extension of a full referring expression—an individual object, in our case—, one additional element is needed, and this is contributed by the determiner. For uniquely referring expressions (“the red cross”), what is required is to pick the most likely candidate from the distribution:

$$
[\![the]\!] = \lambda x. \operatorname*{arg\,max}_{Dom(x)} x \qquad (5)
$$

$$
[\![[the]\ [_{nom} w_1, \ldots, w_k]]\!]^W = \operatorname*{arg\,max}_{i \in W} [\ [\![[_{nom} w_1, \ldots, w_k]]\!]^W\ ] \qquad (6)
$$

5.2 Taxonomic/Ontological know-that

Composition of the ontological information is less fully developed at the moment. We can describe the requirements, though. For a phrase like “*the black dog*”, we would want the general terminological knowledge encoded in (4) (“a dog is an animal, and (typically) is grey or brown or . . . ”) to be specialised to this particular instance (“this dog is an animal . . . ”) and the disjunctive attribute information to be restricted (“. . . and it is black”). This corresponds to the distinction between ‘terminological axioms’ in the so-called TBox and ‘assertional axioms’ in the ABox in Description Logic (Krötzsch et al., 2014), which should also have the necessary expressiveness to realise this composition process.

6 Interaction

The final component is the actual meta-linguistic interaction that takes as topic the adequacy of the predictions made by the other components. As, unlike in distributional semantics or in approaches to language/image matching using deep learning approaches (e.g., (Hu et al., 2016; Mao et al., 2016)), we specify the composition process explicitly, we have access to all its intermediate steps. We can hence provide justifications for object selection decisions that can adress the individual words as well as their composition. This will be described in more detail in the next section.

7 Implementation

7.1 Learning Visual Meanings

The visual classifiers are trained on large sets of images that are segmented into objects, for which referring expressions exist. This is described in more detail for a static recognition task in (Schlangen et al., 2016). We outline the process here, as the trained models form the basis for the interaction, which is the contribution of this paper.

One dataset is the SAIAPR/ReferIt set. It contains of 20k images with a tourism theme (Grubinger et al., 2006) for which object segmentations (Escalante et al., 2010) and, for these objects, referring expressions are available (120k altogether; Kazemzadeh et al. (2014)) . The second dataset is based on the “Microsoft Common Objects in Context” collection (Lin et al., 2014), which contains over 300k images with object segmentations (of objects from 80 pre-specified categories), object labels, and image captions. This has also been augmented with referring expressions by the same group as (Kazemzadeh et al., 2014), in as yet unpublished work (the authors provided us with this data on request). For evaluation, we downloaded 20k images from those synsets covered in ImageNet (Deng et al., 2009) for which object segmentations were available in ImageNet.

As representations of the images we used the final fully-connected layer before the classification layer of a deep convolutional neural network, "GoogLeNet" (Szegedy et al., 2015), that was trained on data from the Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) from the ImageNet corpus (Deng et al., 2009) to extract features, to give us a 1024 dimensional representation of the region. We augment this with 7 features that encode information about the region relative to the image: the (relative) coordinates of two corners, its (relative) area, distance to the center, and orientation of the image. The full representation hence is a vector of 1031 features.

In the training process then, each word occurring in the training part of the referring expression collection is paired with the region it describes (and with 5 randomly selected regions for which this word never ocurrs in the description) to train a logistic regression classifier. Here, we use the classifiers trained on the combination of the two datasets, which yields models for 933 words. As shown in (Schlangen et al., 2016), this model reaches accuracy on a static visual reference resolution task that is at the state of the art of end-to-end deep learning models. Our focus here, however, is on the use of these models in interaction, which we turn to next.

7.2 Interacting with Semantics

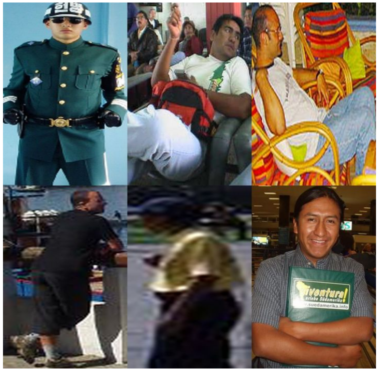
To get an impression of whether interaction ("meaning negotiation") can improve not only performance of the system but also the impression of considered language use, we implemented a simple dialogue system that uses these models. We interfaced it with a popular chat tool,[3] so that interactions can be done remotely.[4]

Together with a simple dialogue logic that handles misunderstandings and state updates (Larsson and Traum, 2000), we implemented the following commands (each accessible through a variety of formulations, only one of which we give here):

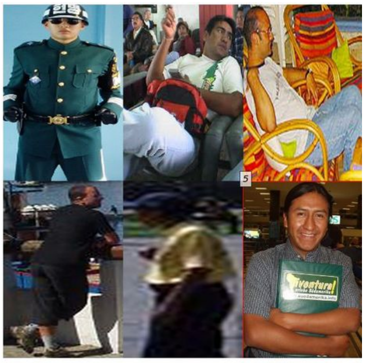
- *draw a new card*, which presents a new image with a natural scene to the user, with the candidate objects (for which manual segmentations are available) outlined.

- *draw a new collage*, which presents a collage consisting of 6 sub-images that each contain a single object; the understanding here is that

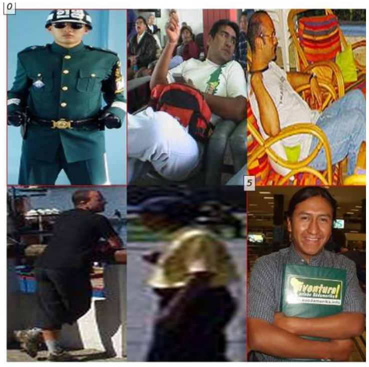
User: get a new collage men
System: *Ok.*



U: Find the man with the book
S: *Let me see. Did you mean this?*



U: yes
S: *Thanks!*
U: explain
S: *Okidoki.*



S: *0 is best for "man"*
S: *5 is best for "book"*
S: *Overal best: 5*
S: *Rank of region 5 for "man": 4*
S: *Rank of region 5 for "book": 1*

Figure 2: **Example Interaction II** *Reference candidates outlined in red and without label; selected candidates with numeric label.*

---

[3] Slack, `http://www.slack.com`

[4] By the time of the conference, we will release all code as open-source. We will also provide a publicly accessible installation of the system and will use it to collect more labels from users, and to learn online from them.

User: what do you associate with trees?
System: *palm, church, tree, house, fountain, beach, grass, statue, road*
U: what looks similar to trees?
S: *tree, plants, leaves, grassy, bushes*
U: what looks similar to bicycle?
S: *bike, rider, four, biker*

Figure 3: **Example Interaction III**

the system “sees” the same sub-images, but not necessarily in the same layout.

- *find X* (e.g., “find the person in the background on the left”), to which the system responds by outlining its selection and asking whether this selection is correct.
- *explain*, which asks for explanation of the selection decision, which the system provides by listing what fits best to each word individually, and then how the selection is composed.
- *what looks similar to X?*; this gives indirect information about the visual classification that the system has learned, as described below.
- *what do you associate with X?*, which similarly gives information about the learning situation, as also described below.
- *X is (not) Y* (e.g., “this is not a boat”, or “3 is not black”), which adds this information to the set of labels, which can then be used for adapting the classifiers.

Information about what looks similar is computed as follows: We randomly select 2000 image regions from a held out set and run all word classifiers on them. This results in 2000 responses (probabilities of fit) for each word, or in other words a 2000-dimensional vector that represents the reactions of this word-classifier to the sample objects. Similarity can then be computed in the usual way as a relation between vectors (we use the cosine); but the resulting type of similarity is a visual one. (More details and evaluations will be given elsewhere.)

The associative information is compiled by computing pointwise mutual information between words ocurring in descriptions of objects within the same scene. This brings objects that often occur together in the same image (such as houses and roads) together.

So far, we have run informal tests during development of the system. In one such test with a naive user, the user interacted for 30 minutes and added more than 40 facts in this time. In a post-experiment questionnaire, they ranked the system highly for the interest that the interactions generated, and they indicated that the interaction helped them form hypotheses about the word meanings learned by the system, better than looking at examples of successful and unsuccessful reference resolutions would have. More formal and comprehensive testing is of course still required.

8 Conclusions

We have outlined a model of grounded semantics that combines perceptual grounding with ontological grounding. This model serves as the basis of a dialogue system that can play a simple reference game, and can provide justifications for the decisions it makes, and accept corrections.

The visual-perceptual part of the model is fairly well-developed, and has been shown elsewhere to achieve good accuracy on an offline task (Schlangen et al., 2016), and has shown some promise as a bidirectional model that can also be used for generation (Zarrieß and Schlangen, 2016). Based on the preliminary tests reported here, embedding it in an interaction seems promising. Much still remains to be done, however. First, the way how what we call the lexical ‘know-how’ here and the ‘know-that’ is combined needs to be more fully formalised, and the reasoning this requires and enables must be described. Second, the taxonomic and ontological knowledge should also be acquired in interaction and be negotiable in interaction. The implementation should form a good basis for making these extensions.

Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology ’CITEC’ (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Thanks are due to the anonymous reviewers for their helpful comments.

The image processing code can be found at `https://github.com/dsg-bielefeld/image_wac`.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of ACM SIGMOD*.

Rudolf Carnap. 1952. Meaning postulates. *Philosophical Studies*, 3:65–73.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantic Theory 2nd edition*. Wiley-Blackwell.

Robin Cooper and Staffan Larsson. 2009. Compositional and ontological semantics in learning from corrective feedback and explicit definition. In *Proceedings of "Diaholmia" (semdial 2009)*, pages 10–14.

Jia Deng, W. Dong, Richard Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

David DeVault, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, USA, July.

Hugo Jair Escalante, Carlos a. Hernández, Jesus a. Gonzalez, a. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

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 2011)*, Barcelona, Spain, July.

L. T. F. Gamut. 1991. *Logic, Language and Meaning: Intensional Logic and Logical Grammar*, volume 2. Chicago University Press, Chicago.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDN: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 13–23, Genoa, Italy.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *NAACL 2015*.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of CVPR 2016*, Las Vegas, USA, June.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

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, July. Association for Computational Linguistics.

Markus Krötzsch, František Simančík, and Ian Horrocks. 2014. A description logic primer. In Jens Lehmann and Johanna Völker, editors, *Perspectives on Ontology Learning*, chapter 1. IOS Press.

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

Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 83–90, Poznan, Poland.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of logic and computation*, 25(2):335–369.

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.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR 2016*, Las Vegas, USA, June.

Diego Marconi. 1997. *Lexical Competence*. MIT Press, Cambride, Mass., 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*.

Ken McRae, George S. Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments & Computers*, 37(4):547—-559, feb.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, pages 1–9.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Hilary Putnam. 1973. Meaning and reference. *Journal of Philosophy*, 70:699–711.

Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. 2002. A trainable spoken language understanding system for visual object selection. In *Proceedings of the International Conference on Speech and Language Processing 2002 (ICSLP 2002)*, Colorado, USA.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3).

Deb Roy. 2005. Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciene*, 9(8):389–396.

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, August.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *ACL 2013*, pages 572—-582, Sofia, Bulgaria.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA, June.

Richmond H. Thomason, editor. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven and London.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of ACL 2016*, Berlin, Germany, August.

Comparing dialogue strategies for learning grounded language from human tutors

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

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

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

Abstract

We address the problem of interactively learning perceptually grounded word meanings in a multimodal dialogue system. Human tutors can correct, question, and confirm the statements of a dialogue agent which is trying to interactively learn the meanings of perceptual words, e.g. colours and shapes. We show that different learner and tutor dialogue strategies lead to different learning rates, accuracy of learned meanings, and effort/costs for human tutors. For example, we show that a learner which can handle corrections in dialogue, and its own uncertainty about what it sees, can learn meanings that are as accurate as a fully-supervised learner, but with less cost/effort to the human tutor.

1 Introduction

Identifying, classifying and talking about objects or events in the surrounding environment are key capabilities for intelligent, goal-driven systems that interact with other agents and the external world (e.g. smart phones, robots, and other automated systems), as well as for image search/retrieval systems. To this end, there has recently been a surge of interest and significant progress made on a variety of related tasks, including generation of Natural Language (NL) descriptions of images, or identifying images based on NL descriptions (Karpathy and Li, 2015; Bruni et al., 2014; Socher et al., 2014). Another strand of work has focused on learning to generate object descriptions and object classification based on low level concepts/features (such as colour, shape and material), enabling systems to identify and describe novel, unseen images (Farhadi et al., 2009; Silberer and Lapata, 2014; Sun et al., 2013).

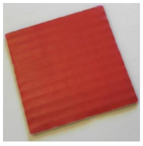
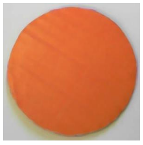
Dialogue	**Image**	**Final semantics**
S: Is this a green square? T: No it's red S: Thanks.		$\begin{bmatrix} x_{=o1} & : & e \\ p2 & : & red(x) \\ p3 & : & square(x) \end{bmatrix}$
T: What can you see? S: something orange. What is it? T: A circle. S: Thanks.		$\begin{bmatrix} x1_{=o2} & : & e \\ p & : & circle(x1) \\ p1 & : & orange(x1) \\ p2 & : & see(sys, x1) \end{bmatrix}$

Figure 1: Interactively agreed semantic contents

Our goal is to build *interactive* systems that can learn grounded word meanings relating to their perceptions of real-world objects – this is different from previous work such as e.g. (Roy, 2002), that learn groundings from descriptions without any interaction, and more recent work using Deep Learning methods (e.g. (Socher et al., 2014)).

Most machine learning approaches to this type of problem rely on training data of high quantity with no possibility of online error correction. Furthermore, they are unsuitable for robots and multimodal systems that need to continuously, and incrementally learn from the environment, and may encounter objects they haven't seen in training data. These limitations are likely to be alleviated if systems can learn concepts, as and when needed, from situated dialogue with humans. Interaction with a human tutor also enables systems to take initiative and seek the particular information they need or lack by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 1). For example, a robot could ask questions to learn the colour of a "square" or to request to be presented with more "red" things to improve its performance on the concept (see e.g. Fig. 1). Furthermore, such systems could allow for meaning negotiation in the form of clarification interactions with the tutor.

This setting means that the system must be *trainable from little data, compositional, adaptive,*

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 44-54.

and able to handle natural human dialogue with all its glorious context-sensitivity and messiness – for instance so that it can learn visual concepts suitable for specific tasks/domains, or even those specific to a particular user. Interactive systems that learn continuously, and over the long run from humans need to do so *incrementally*, *quickly*, and *with minimal effort/cost to human tutors*.

In this paper, we first outline an implemented dialogue system that integrates an incremental, semantic grammar framework, especially suited to dialogue processing – Dynamic Syntax and Type Theory with Records (DS-TTR[1] (Kempson et al., 2001; Eshghi et al., 2012)) with visual classifiers which are learned during the interaction, and which provide perceptual grounding for the basic semantic atoms in the semantic representations (Record Types in TTR) produced by the parser (see Fig. 1, Fig. 2 and section 3). In effect, the dialogue with the tutor continuously provides semantic information about objects in the scene which is then fed to online classifiers in the form of training instances. Conversely, the system can utilise the grammar and its existing knowledge about the world, encoded in the meanings it has already learned, to make reference to and formulate questions about the different attributes of an object identified in the visual scene.[2].

We then go on to use this system, in interaction with a simulated human tutor, to test hypotheses about how the accuracy of learned meanings, learning rates over time, and the overall cost/effort for the human tutor is affected by different dialogue policies and capabilities.

2 Related work

In this section, we will present an overview of vision and language processing systems, as well as multi-modal systems that learn to associate them. We compare them along two main dimensions: *Visual Classification methods: offline vs. online* and *the kinds of representation learned/used*.

Online vs. Offline Learning. A number of implemented systems have shown good performance on classification as well as NL-description of novel physical objects and their attributes, either using offline methods as in (Farhadi et al., 2009; Lampert et al., 2014; Socher et al., 2013; Kong et al., 2013), or through an incremental learning process, where the system's parameters are updated after each training example is presented to the system (Furao and Hasegawa, 2006; Zheng et al., 2013; Kristan and Leonardis, 2014). For the interactive learning task presented here, only the latter is appropriate, as the system is expected to learn from its interactions with a human tutor over a period of time. Shen & Hasegawa (2006) propose the SOINN-SVM model that re-trains linear SVM classifiers with data points that are clustered together with all the examples seen so far. The clustering is done incrementally, but the system needs to keep all the examples so far in memory. Kristian & Leonardis (2014), on the other hand, propose the oKDE model that continuously learns categorical knowledge about visual attributes as probability distributions over the categories (e.g. colours). However, when learning from scratch, it is unrealistic to predefine these concept groups (e.g. that red, blue, and green are colours). Systems need to learn for themselves that, e.g. colour is grounded in a specific sub-space of an object's features. For the visual classifiers, we therefore assume no such category groupings here, and instead learn individual binary classifiers for each visual attribute (see section 3.1 for details).

Distributional vs. Logical Representations. Learning to ground natural language in perception is one of the fundamental problems in Artificial Intelligence. There are two main strands of work that address this problem: (1) those that learn distributional representations using Deep Learning methods: this often works by projecting vector representations from different modalities (e.g. vision and language) into the same space in order to be able to retrieve one from the other (Socher et al., 2014; Karpathy and Li, 2015; Silberer and Lapata, 2014); (2) those that attempt to ground symbolic logical forms, obtained through semantic parsing (Tellex et al., 2014; Kollar et al., 2013; Matuszek et al., 2014) in classifiers of various entities types/events/relations in a segment of an image or a video. Perhaps one advantage of the latter over the former method, is that it is strictly compositional, i.e. the contribution of the meaning of an individual word, or semantic atom, to the whole representation is clear, whereas this is hard to say about the distributional models. As noted, our work also uses the latter methodology, though it is

---

[1] Download from `http://dylan.sourceforge.net`

[2] Here we assume that the words being grounded are in the lexicon, i.e. that their syntactic and semantic type are known: we leave the problem of grammar induction to one side here, though see (Eshghi et al., 2013)

dialogue, rather than sentence semantics that we care about. Most similar to our work is probably that of Kennington & Schlangen (2015) who learn a mapping between individual words - rather than logical atoms - and low-level visual features (e.g. colour-values) directly. The system is compositional, yet does not use a grammar (the compositions are defined by hand). Further, the groundings are learned from pairings of object references in NL and images rather than from dialogue.

What sets our approach apart from others is: a) that we use a domain-general, incremental semantic grammar with principled mechanisms for parsing and generation; b) Given the DS model of dialogue (Eshghi et al., 2015), representations are constructed jointly and interactively by the tutor and system over the course of several turns (see Fig. 1); c) perception and NL-semantics are modelled in a single logical formalism (TTR); d) we effectively induce an ontology of atomic types in TTR, which can be combined in arbitrarily complex ways for generation of complex descriptions of arbitrarily complex visual scenes (see e.g. (Dobnik et al., 2012) and compare this with (Kennington and Schlangen, 2015), who do not use a grammar and therefore do not have logical structure over grounded meanings).

3 System Architecture

We have developed a system to support an attribute-based object learning process through natural, incremental spoken dialogue interaction. The architecture of the system is shown in Fig. 2. The system has two main modules: a vision module for visual feature extraction and classification; and a dialogue system module using DS-TTR. Below we describe these components individually and then explain how they interact.

3.1 Attribute-based Classifiers used

Yu et. al (2015a; 2015b) point out that neither multi-label classification models nor ‘zero-shot’ learning models show acceptable performance on attribute-based learning tasks. Here, we instead use Logistic Regression SVM classifiers with Stochastic Gradient Descent (SGD) (Zhang, 2004) to incrementally learn attribute predictions.

All classifiers will output attribute-based label sets and corresponding probabilities for novel unseen images by predicting binary label vectors. We build visual feature representations to learn classifiers for particular attributes, as explained below.

3.1.1 Visual Feature Representation

In contrast with previous work (Yu et al., 2015a; Yu et al., 2015b), to reduce feature noise through the learning process, we simply extract a a 1280-dimensional feature vector consisting of only two base feature categories, i.e. the colour space for colour attributes, and a ‘bag of visual words’ for the object shapes/class (as shown in Fig. 2).

Colour descriptors, consisting of HSV colour space values, are extracted for each pixel and then are quantized to a $16\times4\times4$ HSV matrix. These descriptors inside the bounding box are binned into individual histograms. Meanwhile, a bag of visual words is built in PHOW descriptors using a visual dictionary (that is pre-defined with a handmade image set). These visual words are calculated using 2x2 blocks, a 4-pixel step size, and quantized into 1024 k-means centres.

3.2 Dynamic Syntax and Type Theory with Records

Dynamic Syntax (DS) a is a word-by-word incremental semantic parser/generator, based around the Dynamic Syntax (DS) grammar framework (Cann et al., 2005) especially suited to the fragmentary and highly contextual nature of dialogue. In DS, dialogue is modelled as the interactive and incremental construction of contextual and semantic representations (Eshghi et al., 2015). The contextual representations afforded by DS are of the fine-grained semantic content that is jointly negotiated/agreed upon by the interlocutors, as a result of processing questions and answers, clarification requests, corrections, acceptances, etc. We cannot go into any further detail here due to lack of space, but proceed to briefly describe Type Theory with Records, the formalism in which the DS contextual/semantic representations are couched.

Type Theory with Records (TTR) is an extension of standard type theory shown to be useful in semantics and dialogue modelling (Cooper, 2005; Ginzburg, 2012). TTR is particularly well-suited to our problem here as it allows information from various modalities, including vision and language, to be represented within a single semantic framework (see e.g. Larsson (2013); Dobnik et al. (2012) who use it to model the semantics of spatial language and perceptual classification).

In TTR, logical forms are specified as *record*

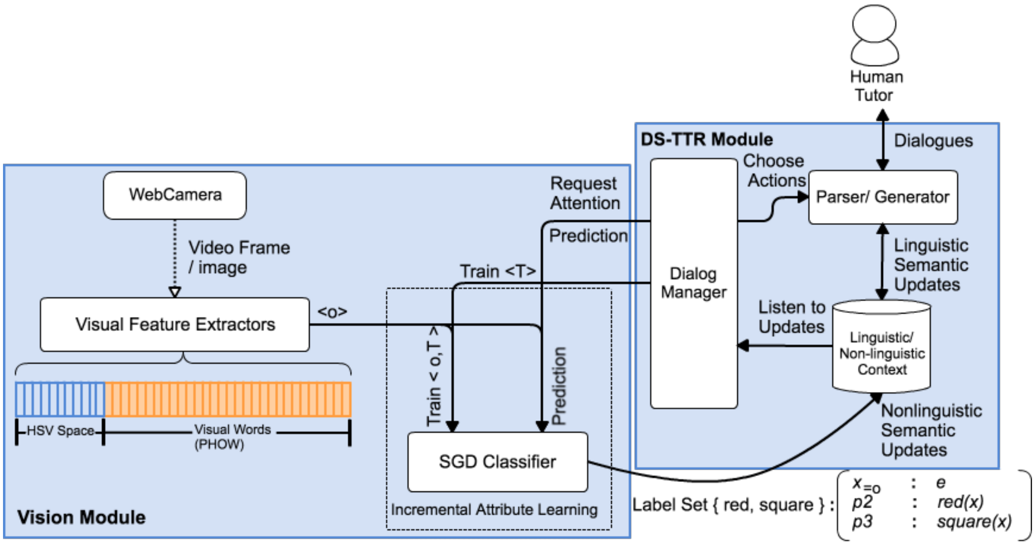


Figure 2: Architecture of the teachable system

types (RTs), which are sequences of *fields* of the form $[\, l : T \,]$ containing a label l and a type T. RTs can be witnessed (i.e. judged true) by *records* of that type, where a record is a sequence of label-object pairs $[\, l = v \,]$. We say that $[\, l = v \,]$ is of type $[\, l : T \,]$ just in case v is of type T. Importantly for us here, TTR has a subtyping relation, in terms of which inference is defined; but it also allows semantic information to be incrementally specified, i.e. record types can be indefinitely extended with more information/constraints. This is a key feature since it allows the system to encode *partial* knowledge about objects, and for this knowledge to be extended in a principled way, as and when it becomes available.

For further detail on TTR, see Cooper (2005) and Dobnik et al. (2012) among others.

3.3 Integration

Fig. 2 shows how the various parts of the system interact. At any point in time, the system has access to an ontology of (object) types and attributes encoded as a set of TTR Record Types, whose individual atomic symbols, such as ‘red’ or ‘square’ are grounded in the set of classifiers trained so far.

Given a set of individuated objects in a scene, encoded as a TTR Record, the system can utilise its existing ontology to output some maximal set of Record Types characterising these objects (see e.g. Fig. 1). Since these representations are shared by the DS-TTR module, they provide a direct interface between perceptual classification and semantic processing in dialogue: they can be used directly at any point to generate utterances, or ask questions about the objects.

On the other hand, the DS-TTR parser incrementally produces Record Types (RT), representing the meaning jointly established by the tutor and the system so far. In this domain, this is ultimately one or more type judgements, i.e. that some scene/image/object is judged to be of a particular type, e.g. in Fig. 1 that the individuated object, $o1$ is a red square. These jointly negotiated type judgements then go on to provide training instances for the classifiers. In general, the training instances are of the form, $\langle O, T \rangle$, where O is an image/scene segment (an object or TTR Record), and T, a record type. T is then decomposed into its constituent atomic types $T_1 \ldots T_n$, s.t. $\bigwedge T_i = T$. The judgements $O : T_i$ are then used directly to train the classifier that grounds the T_i.

4 Experiments and Results

In general, in real-world problems, there are a variety of dialogue behaviours that human tutors might adopt to teach the learner with novel knowledge, and these might lead to different reactions from the learner/system as well as different outcomes for the recognition performance of the learned concepts/meanings, effort from the tutor and trade-offs between these. Moreover, a learner with different capabilities (described below) can also affect these performances through dialogue. Our goal in this paper is therefore to explore the effects of these dialogue behaviours and capabilities on the overall performance of the learning agent by measuring the trade-off between recognition performance and tutoring cost.

4.1 Design

Before explaining the experiment configurations, there are several notions that need to be defined in terms of basic dialogue capabilities, tutor behaviours, and learner dialogue capabilities –

Basic Dialogue Capabilities: The following capabilities are explored for both the tutor and the learner (see examples in Fig. 3):

- **Listening**: this only refers to a *learner*, while the *tutor* is making a statement about a specific object/attribute;
- **Statement**: the ability for both *learners* and *tutors* to describe attributes of an object, e.g. "this is a red square" or "this is red";
- **Correction**: the ability to process corrections only from the *tutor*, e.g. "no, this is green" or "no, this is a circle';
- **Implicit/explicit confirmation**: the ability to process confirmations from the *tutor*, e.g. "Yes, it's a square";
- **Question-answering**: the ability to answer questions from both the *tutor* and the *learner*, e.g. "T: what is this? S: this is a red square.";
- **Question-asking**: the ability to ask WH or polar questions requesting correct information, e.g. "what colour is this?" or "is this a red square?".

Tutor Behaviours: Following previous work (Skočaj et al., 2009), we generally identify tutor behaviours based on how he/she treats the learner into two groups: **1)** ***Tutor-Driven (TD)***: The tutor always gives available information about a particular object, i.e. supervised learning (always providing labels), by directly making statements (e.g. "this is a square" or "this is a red square"). This means that the whole learning process is an unidirectional interaction only handled by the tutor. In this case, the learner only needs to listen and update its learning models (i.e. the visual classifiers) upon what information the tutor presented. **2)** ***Tutor-Corrected (TC)***: while the learner is describing or asking something about the object, the tutor only asks WH questions and corrects mistakes of the learner, and otherwise confirms correct statements (e.g. "T: what is this? L: this is a red square. T: yes/no, it is a green square" in Fig. 3). In contrast to the TD behaviour, the learner performs more actively to get involved with the learning process with its own predictions/knowledge. It will update its classifiers only when the tutor provides answers or confirms.

According to the previous work from Skočaj et. al. (2009), both tutor strategies are frequently adopted in a perceptual learning process, which may lead to different levels of learner involvement. They assumed that the tutor can always perform well through the entire learning process. However, this may be extremely idealised for real-world problems, in which human tutors may not always supply all their knowledge when informing about a visual object. In this paper, we therefore also take the following situations into account:

- ***"Good-Tutor" (GT)***: the tutor always gives all the labels for each image, always corrects all the mistakes of the learner, and always confirms correct statements by the learner.
- ***"Lazy-Tutor" (LT)***: this tutor only gives one of the correct labels at a time (e.g. "it's red" or "it's a square"), and only corrects one mistake at a time. It always confirms when asked to. This tutor is more similar to what we can expect from real human behaviour when teaching robots than the Good Tutor.

Learner Dialogue Capabilities: In real-world learning tasks, a learner might be required to consider several additional capabilities, which may enable it to respond to tutor behaviours in a more natural way, especially with a *Lazy-Tutor* situation. Moreover, these capabilities are also likely to help improve the overall performance of the learner at the end of learning process, i.e. achieving a better trade-off between the performance of object/attribute recognition and the cost for the tutor. In this paper, we use the dialogue system outlined above to carry out experiments with three dialogue capabilities with two levels: **1) Uncertainty (+UC/-UC)**: determines whether the learner takes into account, in its dialogue behaviour, its own subjective confidence about the attributes of the presented object. The confidence is the score assigned by any of its attribute classifiers of the object being a positive instance of an attribute (e.g. 'red') - see below for how a confidence threshold is used here. In condition +UC, the agent will not ask a question if it is confident about the answer, and it will hedge the answer to a tutor question if it is not confident, e.g. "T: What is this? L: is this a red square?". In -UC, the learner

	TD	TC (-UC)	TC (+UC)
Good(Ideal) Tutor	**T:** this is red. **L:** okay. **T:** this is a square. **L:** okay. ----------- or ------------- **T:** this is a red square. **L:** okay.	**T:** what is this? **L:** this is a red square. **T:** no, this is a green square. **L:** okay.	**T:** what is this? **L:** is this a red square? **T:** no, this is a green square. **L:** okay.
Lazy(natural) Tutor	Without Knowledge-Demanding (-KD)		
	T: this is red. **L:** okay.	**T:** what is this? **L:** this is a red square. **T:** yes, it is a square. **L:** okay.	**T:** what is this? **L:** is this a red square? **T:** yes , this is a square. **L:** okay.
	With Knowledge-Demanding (+KD)		
	T: this is red. **L:** okay. what shape is it? **T:** this is a square. **L:** okay.	**T:** what is this? **L:** this is a red square. **T:** no, this is a circle. **L:** okay. Is the colour correct? **T:** yes. **L:** okay.	**T:** what is this? **L:** is this a green circle? **T:** no, this is a square. **L:** okay. Is the colour correct? **T:** no, this is red. **L:** okay.

Figure 3: Example dialogues in different conditions (TD = tutor-driven, TC = tutor-corrected, -UC= no learner uncertainty, +UC= learner uncertainty)

is confident and always takes itself to know the attributes of the given object (as given by its currently trained classifiers), and behaves according to that assumption. **2) Knowledge-Demanding (+KD/-KD)**: this determines whether the learner can request further details/information about objects, which may be useful when interacting with a "Lazy" Tutor (described above). In condition +KD, the learner is able to request more information by asking extra questions (see Fig. 3 e.g. "what (colour/shape) is it? or "is the colour correct?". Otherwise, the learner with -KD will only update the classifiers based on the information provided.

Confidence Threshold: To determine when and how the agent properly copes with its attribute-based predictions, we use confidence-score thresholds. It consists of two values, a base threshold (e.g. 0.5) and a positive threshold (e.g. 0.9).

If the confidences of all classifiers are under the base threshold (i.e. the learner has no attribute label that it is confident about), the agent will ask for information directly from the tutor via questions (e.g. "L: what is this?").

On the other hand, if one or more classifiers score above the base threshold, then the positive threshold is used to judge to what extent the agent trusts its prediction or not. If the confidence score of a classifier is between the positive and base thresholds, the learner is not very confident about its knowledge, and will check with the tutor, e.g. "L: is this red?". However, if the confidence score of a classifier is above the positive threshold, the learner is confident enough in its knowledge not to bother verifying it with the tutor. This will lead to less effort needed from the tutor as the learner becomes more confident about its knowledge.

However, since a learner with high confidence will not ask for assistance from the tutor, a low positive threshold may reduce the opportunities that allow the tutor to correct the learner's mistakes. With an additional experiment (*note*: we will not explain it here due of lack of space), we determined a 0.5 base threshold and a 0.9 positive threshold as the most appropriate values for an interactive learning process - i.e. this preserved good classifier recognition while not requiring much effort from the tutor. In (Yu et al., 2016) we show how these thresholds can be optimised.

4.2 Experimental Setup

We carried out a set of experiments to investigate the effects of these dialogue policies on an interactive learning process with a tutor. We compare different behaviours and capabilities with two baseline policies without corrections (NC), in which the learner cannot process corrections but only confirmations from the tutor. This means that the learner can update its classifiers only when its own predictions are correct. There are several settings related to these experiments below:

Table 1: Recognition Score Table

	Yes	LowYes	LowNo	No
Yes	1	0.5	-0.5	-1
No	-1	-0.5	0.5	1

Table 2: Tutoring Cost Table

C_{inf}	C_{yes}	C_{crt}	C_{ign}	C_{turn}
1	0.25	1	0	0.15

Tutor Simulation and Policy: To run our experiment on a large-scale, we have hand-crafted an *Interactive Tutoring Simulator*, which simulates the behaviour of a human tutor[3]. The tutor policy is set up based on different tutor-based behaviours and situations as mentioned above.

Evaluation and Cross-validation: To evaluate the performance of the system in each condition, we performed a 100-fold cross validation with 500 images for training and 100 for testing within a handmade object set[4]. For each training instance, the learning system interacts with the simulated tutor. We define a **Learning Step** as comprised of 25 such dialogues. At the end of each learning step, the system is tested using the test set. The values used for the Tutoring Cost and the Recognition Score at each learning step correspond to averages across the 100 folds.

4.3 Evaluation Metrics

To test how the different dialogue capabilities and strategies affect the language learning process, we follow metrics proposed by Skočaj et al.(2009), that consist of two main evaluation measures, i.e. *Recognition Scores* and *Tutoring Costs*. We tweak the details below to reflect our own dialogue system configurations.

Recognition score: This is a metric measuring the overall accuracy of the learned word meanings / classifiers, which "rewards successful classifications (i.e. true positives and true negatives) and penalizes incorrect predictions (i.e. false positives and false negatives)" (Skočaj et al., 2009) [5]. As the proposed system considers both correctness of predicted labels and prediction confidence on learning tasks, the measure will also take the true labels with lower confidence into account, as shown in Table 1; "LowYes" means that the system made positive predictions but with lower confidence. In this case, the system can generate a polar question for requesting tutor feedback. "LowNo" is similar to "LowYes", but only works on negative predictions.

Cost: The cost measure reflects the effort needed by a human tutor in interacting with the system. Skocaj et. al. (2009) point out that a comprehensive teachable system should learn as autonomously as possible, rather than involving the human tutor too frequently. There are several possible costs that the tutor might incur, see Table 2: C_{inf} refers to the cost of the tutor providing information on a single attribute concept (e.g. "this is red" or "this is a square"), and we set this cost as 1; C_{yes} is the cost of a simple confirmation (like "yes", "right") and set it to be 0.25; C_{crt} is the cost of correction for a single concept (e.g. "no, it is blue" or "no, it is a circle") and is also set to be 1. Moreover, the number of dialogue turns from the tutor was also taken into account in measuring total cost: each single turn costs 0.15 in this experiment. These values are based on the intuition that it is just as much effort for the Tutor to provide a concept as to correct one, and that confirmation has a smaller cost, while each turn also requires a small effort from the Tutor.

Performance Score As mentioned above, an efficient Learner dialogue policy should consider both classification accuracy (Recognition score) and tutor effort (Cost). We thus defined an integrated measure – the *Performance Score* (S_{perf}) – that we use to compare the general performance across different dialogue policies and capabilities:

$$S_{perf} = \frac{S_{recog}}{C_{tutor}}$$

i.e. the ratio of Recognition Score achieved by the Learner to the effort/Cost required by the Tutor. We seek dialogue strategies that balance these metrics.

4.4 Results

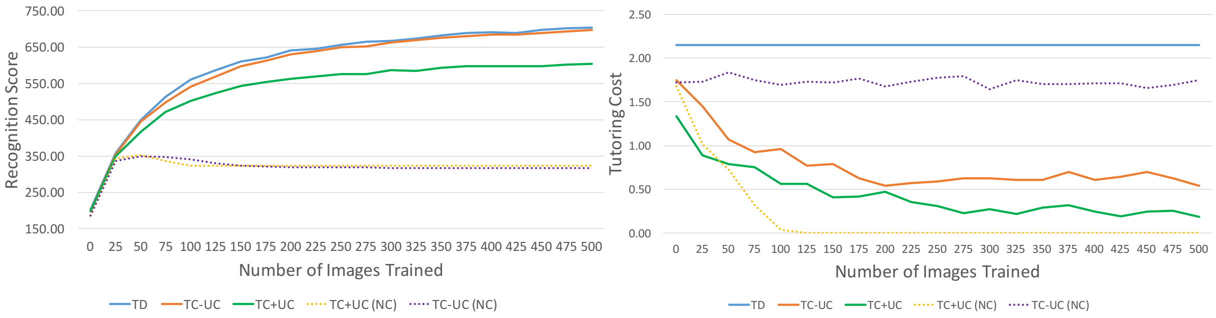
We first investigate the improvement of learning performance over time for different learner policies and capabilities with an ideal tutoring situation (*Good* Tutor) (see Fig. 4). We compared both tutor policies (TD and TC) with correspond-

---

[3] The experiment involves hundreds of dialogues, so running this experiment with real human tutors has proven too costly at this juncture, though we plan to do this for a full evaluation of our system in the future.

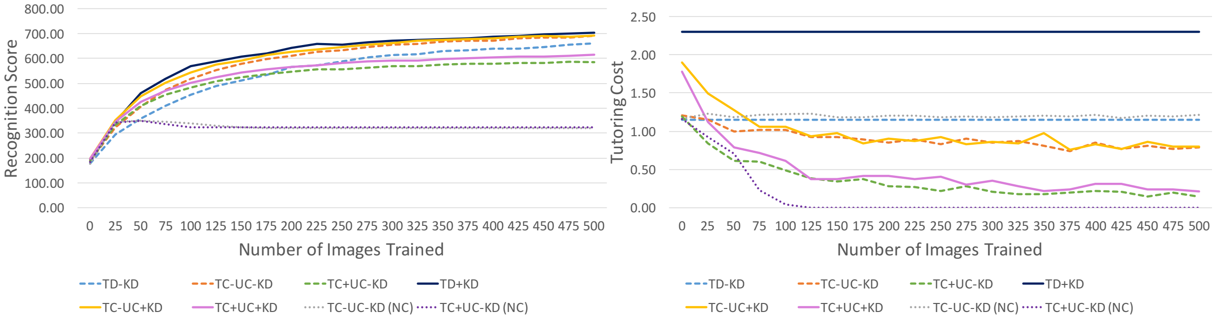
[4] All data from this paper will be made freely available.

[5] we use recognition score instead of accuracy because it better handles uncertainty predictions than accuracy, which could be more similar to a human-like learning task.



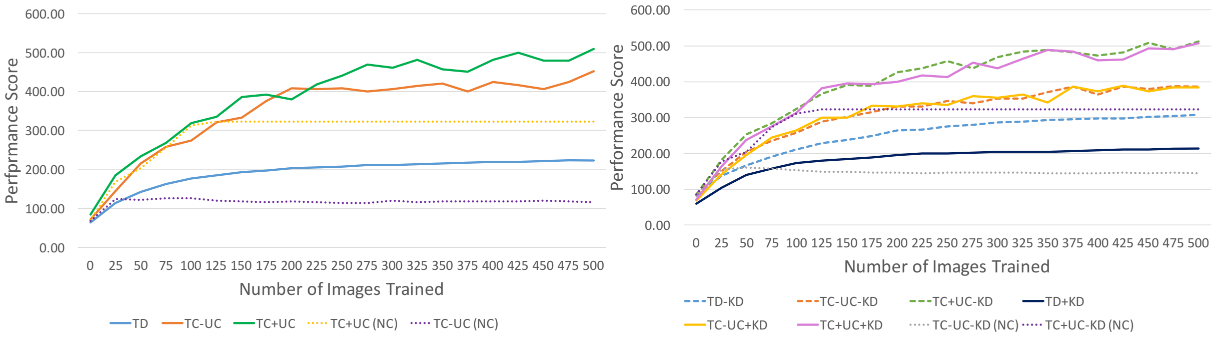
(a) Recognition Score (b) Tutoring Cost

Figure 4: Evolution of Learning Performance in the *Good Tutor* Condition (TD = tutor-driven, TC = tutor-corrected, -UC= no learner uncertainty, +UC= learner uncertainty, NC= no corrections)



(a) Recognition Score (b) Tutoring Cost

Figure 5: Evolution of Learning Performance in the *Lazy Tutor* Condition (TD = tutor-driven, TC = tutor-corrected, UC= learner uncertainty, NC= no corrections, KD= Knowledge-demanding)



(a) Recognition/Cost Ratio for “Good tutor” (b) Recognition/Cost Ratio for “Lazy Tutor”

Figure 6: Learner policy Performance with both Tutor types TD = tutor-driven, TC = tutor-corrected, UC= learner uncertainty, NC= no corrections, KD= Knowledge-demanding)

ing learner strategies and capabilities (+/-UC and NC) in terms of Recognition Score and Tutoring Cost. (Note that in the Good Tutor case, +/-KD has no effect).

Here we see that the Tutor-Driven (TD, blue line) and Tutor-Corrected without Uncertainty (TC-UC, red line) conditions gain the highest Recognition scores, while conditions without the Learner ability to process tutor corrections (NC) perform badly, as expected. In terms of Tutoring Cost though, we see that TD has a high cost while TC-UC has quite low cost. Interestingly, TC+UC (Tutor-corrected, with Uncertainty, green line), has a lower cost than both of these conditions, while still achieving a high Recognition score. This is because the Learner which is aware of its uncertainty about classifier outputs requires fewer corrections from the Tutor, while the classifiers still become more accurate over time.

Similar to Fig. 4, Figs. 5a, b show the Recognition Score and Tutoring Cost respectively for the same learner strategies, but with a more natural tutoring situation (*Lazy* Tutor), and where the learner can be Knowledge-Demanding (+/-KD). In addition, Fig. 6 shows the overall performance of different learner strategies (i.e. the trade-offs between the recognition score and the tutoring cost) in the Good and Lazy Tutor situations separately.

Here, in the Good-Tutor condition, the TC-UC policy (orange line) shows better overall performance than TD (blue line) because of its lower tutoring cost. In addition, though the Uncertain Learner (TC+UC, green line) policy performs slightly worse on recognition score (this might be due to insufficient error detection and recovery), it also reduces the tutoring cost through time. Hence, this policy achieves better performance than the others in the final results (see Fig. 6a).

In terms of the *Lazy-Tutor* condition, both the TD and TC-UC policies, without Knowledge-demand (-KD), show slightly worse recognition performance than they did under the Good Tutor policy, because the learner does not gain as much knowledge from the tutor in each learning step. Whilst both policies cost much less than before for the same reason, they show better performance in the final results (as compared between Figures 6a and 6b). By contrast, as a situation with two incorrect predictions rarely occurs with the TC+UC-KD policy (for only about 20 out of 500 images), the *Lazy-Tutor* policy will not affect Recognition Score or Tutoring Cost very much for the TC+UC policy (see Fig. 5a, b). Therefore, its final performance shows a similar tendency as under the Good Tutor condition.

Moreover, the results in Figure 5 also show that a Knowledge-Demanding (+KD) learner policy may always improve recognition performance (Fig. 5a). For the Lazy Tutor condition, the conditions TC+UC+KD (pink line) and TC+UC-KD (green dotted line) have the best overall performance (Fig. 6b).

Since our ultimate goal here is to create a full dialogue system that can learn accurate concepts (word meanings) with little effort from human tutors, these results would lead us to choose a dialogue system that can can handle corrections – i.e. some variant of the Tutor Corrected system. The results show that, depending on the relative weight between Recognition Score and Tutor Cost, an optimal Learner Dialogue Policy could, for example, use TC-UC(NC) for the first 50 or 60 images, and then switch to TC+UC. We investigate such dynamic policies and their optimisation in a later study using Reinforcement Learning methods (Yu et al., 2016).

5 Conclusion

We have developed a multimodal dialogue interface to explore the effectiveness of situated dialogue with a human tutor for learning perceptually-grounded word meanings. The system integrates semantic representations from an incremental semantic parser/generator, DS-TTR, with attribute classification models that ground the semantic representations.

We compared the system's performance (its Recognition Score and Tutor Cost) under several different dialogue policies for interactive language grounding, on a hand-made dataset of simple objects. Overall, we see that dialogue interaction is important for teachable agents as it reduces the effort required from the human tutor. The fully supervised cases (TD) have a high cost for the Tutor, and equivalent final recognition performance can be reached with less effort when using a Tutor-Corrected (TC) dialogue policy where the Learner can process corrections in dialogue. Final Recognition performance is slightly less good with learners which take their own uncertainty into account (TC+UC), but they require much less effort from Tutors, resulting in better overall performance.

Ongoing work explores full Learner dialogue policies (i.e. turn-based decisions about what to say next) and their optimisation using Reinforcement Learning methods (Rieser and Lemon, 2011; Yu et al., 2016).

Acknowledgements

This research is supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project[6]), and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688147 (MuMMER[7]).

---

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

[7] `http://mummer-project.eu/`

References

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(1–47).

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

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

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing (CSLPÄô12)*, pages 51–63.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.

Arash Eshghi, 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, August. Association for Computational Linguistics.

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.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*.

Shen Furao and Osamu Hasegawa. 2006. An incremental network for on-line unsupervised classification and topology learning. *Neural Networks*, 19(1):90–106.

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

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137.

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

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

Thomas Kollar, Jayant Krishnamurthy, and Grant Strimel. 2013. Toward interactive grounded language acqusition. In *Robotics: Science and Systems*.

Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou. 2013. Transductive multilabel learning via label set propagation. *IEEE Trans. Knowl. Data Eng.*, 25(3):704–719.

Matej Kristan and Ales Leonardis. 2014. Online discriminative kernel density estimator with gaussian kernels. *IEEE Trans. Cybernetics*, 44(3):355–365.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of logic and computation*.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 2556–2563.

Verena Rieser and Oliver Lemon. 2011. Learning and evaluation of dialogue strategies for new applications: Empirical methods for optimization from small data sets. *Computational Linguistics*, 37(1):153–196.

Deb Roy. 2002. A trainable visually-grounded spoken language generation system. In *Proceedings of the International Conference of Spoken Language Processing*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 721–732, Baltimore, Maryland, June. Association for Computational Linguistics.

Danijel Skocaj, Matej Kristan, Alen Vrecko, Marko Mahnic, Miroslav Janícek, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*, pages 3387–3394.

Danijel Skočaj, Matej Kristan, and Aleš Leonardis. 2009. Formalization of different learning strategies in a continuous learning framework. In *Proceedings of the Ninth International Conference on Epigenetic Robotics; Modeling Cognitive Development in Robotic Systems*, pages 153–160. Lund University Cognitive Studies.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems*, pages 935–943, Lake Tahoe, Nevada, USA.

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*, 2:207–218.

Yuyin Sun, Liefeng Bo, and Dieter Fox. 2013. Attribute based object identification. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2096–2103. IEEE.

Stefanie Tellex, Pratiksha Thaker, Joshua Mason Joseph, and Nicholas Roy. 2014. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 94(2):151–167.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2015a. Comparing attribute classifiers for interactive language grounding. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 60–69, Lisbon, Portugal, September. Association for Computational Linguistics.

Yanchao Yu, Oliver Lemon, and Arash Eshghi. 2015b. Interactive learning through dialogue for multimodal language grounding. In *SemDial 2015, Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue, Gothenburg, Sweden, August 24-26 2015*, pages 214–215.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *(under review)*.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM.

Jun Zheng, Furao Shen, Hongjun Fan, and Jinxi Zhao. 2013. An online incremental learning support vector machine for large-scale data. *Neural Computing and Applications*, 22(5):1023–1035.

Evaluating conversational success: Weighted Message Exchange Games

Nicholas Asher and **Soumya Paul**
IRIT, Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse, France
`nicholas.asher@irit.fr, soumya.paul@gmail.com`

Abstract

We analyze evaluations of conversational success and how such evaluations relate to notions of discourse content and structure. To do so, we extend the framework of Message Exchange (ME) games by adding weights or scores to the players' moves and then accumulating these weights using discounting to evaluate a conversationalist's performance. We illustrate our analysis on a fragment of a recent political debate.

1 Introduction

As is by now well accepted, a discourse is more than an unstructured set of utterances; these utterances should, for example, be related to one another in a coherent fashion. But in general, not just any coherent arrangement of utterances will do. If one's goal is merely to avoid an awkward silence, then maintaining conversational coherence might suffice to achieve one's ends, but conversational goals are frequently more ambitious than this. Sometimes interlocutors converse to get to the truth of a matter; other times, a speaker says what she does to convince her interlocutor or a third party, an observer, to do something or to adopt a certain belief; in the latter case, the truth of what she says might be less important than its persuasiveness. One might win a political debate, for instance, even if the majority of the claims one asserts in that debate are false, as the 2016 series of debates between Republican candidates for the U.S. Presidency illustrates.

With (Grice, 1975), we hold that conversations are rational activities, and that agents act so as to maximize their conversational success. But in order for that to be possible, conversational agents, and observers, must be able to evaluate conversations for such success, and this requires moving beyond evaluations of discourse content in terms of truth or satisfaction. In particular, we want to know how the linguistic and discourse structure and content of a speaker's contributions affect that evaluation. In this paper, we propose a model of context-sensitive *evaluations of conversational success* and investigate how such evaluations relate to notions of discourse content and structure. In our view, a better understanding of conversational success will shed light on how agents structure their contributions and how these contributions affect the overall shape and content of the conversation.

Conversational success need not be shared by all members of a conversation; speakers can have different and even opposed conversational goals. We thus develop our model of conversational success using the framework of Message Exchange (ME) Games (Asher et al., 2016), in which a conversation is understood as a sequential, extended game that does not require interlocutors to share interests or goals. To avoid troublesome backwards induction results that predict that no conversation takes place in cases of opposed interests (Crawford and Sobel, 1982; Asher et al., 2016), an ME game conceives of a conversation without a commonly known, set end and thus models conversations as infinitary games. Such games are evaluated by a *Jury*. Intuitively, a Jury is any entity or a group of entities that evaluates a conversation and decides the winner. For example, in a courtroom situation, the Jury is the courtroom Jury itself whereas in a political debate, the Jury is the audience of the debate which maybe the entire citizenry of a country. A Jury can even be one of the participants of the conversation itself. Thus, a Jury for a particular conversation setup depends entirely on the context. But given such a setup, it is always clear who or what constitutes the Jury. We formalize the Jury here as a *weight function* or *scoring function* over the sequence of

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 55-64.

conversational moves. To accumulate the individual weights to obtain a global score of a conversation for the players, we will use techniques of discounting (Shapley, 1953).

To motivate these decisions in our analysis ofconversational success, consider a recent example from the U.S. Republican primary debates (February 6, 2016) where things go dramatically wrong for a candidate Marco Rubio (R), the junior US senator from Florida. The crucial episode can be viewed at (Christie-Rubio-debate, 2016), and the transcript at (Christie-Rubio-transcript, 2016).

We describe the relevant part of that conversation below where the numbers correspond to blocks of sequential discourse moves making up a coherent unit. In terms of the linguistic theory SDRT, these blocks correspond to *complex discourse units* or CDUs (Asher and Lascarides, 2003). A CDU is a structure consisting of elementary discourse units (typically clauses) that are linked together by discourse relations and, crucially, that bear together some rhetorical relation to another discourse unit. For example, the block (3) below in an SDRT analysis would yield a CDU consisting of several EDUs; the first sentence (a) yields an EDU that is elaborated on by the EDU derived from (b), with the (c) and (d) elaborating on (b). The division of the conversation into CDUs and their numbering will help us in carrying out a detailed analysis in Section 4.

Fielding a question about his experience to be president given that he is a very junior US senator, R initially responds with (1) a summary of his record in the Senate, (2) a short argument that experience isn't sufficient for being President and then concludes (3) by drawing a comparison between himself and Obama, who, like R, had only one term of political experience at the national level before running for President:

(3) *"(a) And let's dispel once and for all with this fiction that Barack Obama doesn't know what he's doing. (b) He knows exactly what he's doing. (c) Barack Obama is undertaking a systematic effort to change this country, (d) to make America more like the rest of the world."*

(3) is a coherent move when R utters it. The question to which R is responding carries with it an implicit argument against him. The major premise of that argument is that no one who has had only one term of legislative experience could be a President who *"knows what he's doing."* R argues that Obama was very effective, thus challenging this premise.

The floor then goes to Governor Christie (C) of New Jersey, who takes issue with R's response and attacks his record in the Senate (4) and picks up the comparison to Obama (5). R responds by attacking C's record as governor (6), which is a natural move. But then something strange happens: in (7), R goes back and repeats (3) almost verbatim:

(7) *"But I would add this. Let's dispel with this fiction that Barack Obama doesn't know what he's doing. He knows exactly what he's doing. He is trying to change this country. He wants America to become more like the rest of the world."*

C then characterizes R's response in an extremely damaging way:

(8) *"That's what Washington, D.C. Does. The drive-by shot at the beginning with incorrect and incomplete information and then the memorized 25-second speech that is exactly what his advisers gave him."*

The debate continues with R again attacking C's record (9). Had R stuck to this strategy, he might have recovered from his faux pas repetition; but instead, he goes back and repeats in block (10) the material in (3) and (7) without any attempt to respond to C's characterization of the repetition in (7). In block (11) C once again points out the *"memorized text"* to R's detriment. The effect of this repetition and his failure to counter C's negative characterization of it was disastrous for R as pundits claimed and subsequent polls confirmed; C's characterization gave a label for R's *"robotic performance,"* and the video in (Christie-Rubio-debate, 2016) went viral.

While prior work on a conversationalist's success or 'power status' has focused on superficial features like the number of turns the speaker has, the length of time she has spoken, or word bigrams (Prabhakaran et al., 2012; Prabhakaran et al., 2013), examples like the Rubio gaffe show that a dialogue participant's success in meeting her conversational objectives depends upon the individual moves that she makes *in the particular dialogue context*. When pundits and the public evaluated the debate performance of the candidates, they justified their evaluations by making reference to particular moves in the debate, including R's 'robotic' repetitions. Had R simply given (3) in his response to the moderator's leading question, the response would have been fine. But same message (e.g. (7) and (10)) in a different context (e.g. following (4) and (5) and then (8)) gets a very different and bad score. Further, R's 'robotic' response affects the evaluation of the rest of the conversation, penalizing his subsequent performance.

To model evaluations of conversational success, we need to answer three questions: (a) how do we characterize the context upon which the evaluation is based? (b) in virtue of what does one give such an evaluation? (c) how does the evaluation proceed? Given our characterization of Rubio's performance, evaluators are sensitive to the exact words used, to the conversational string, but they also evaluate whether a particular discourse move or sequence of moves performs a coherent rhetorical role, like answering a question, amplifying on a response to a question, rebutting a prior attack move by another participant, and so on.

With respect to question b, evaluators exploit criteria like responsiveness and coherence, taking, e.g., an attack on an agent i to which i has no coherent rebuttal to contribute to a negative evaluation of a response given by i. Evaluation of conversational success also depends, however, on what is needed to persuade the evaluator that an agent has been successful. This may depend upon the agent's own global goals like defending a particular position, but it may also depend upon the evaluator's preconception of what a successful conversation for i would be.

Finally, to answer question c, a global evaluation of Player i's contributions depends on the contributions she makes on each of her turns and how they are related to the discourse context. The evaluation of i's performance in the conversation should be a function of the evaluation she receives on each turn. We examine a normalized, additive function that assigns to each turn for every debater i a score in $\{0, 1, \ldots, d\}$ where d is a positive integer. However, a bad evaluation on one turn like that of Rubio's (or 1988 Vice-Presidential candidate Quayle's famous gaffe (Asher and Paul, 2013)) colors the evaluation of further turns, and several bad evaluations can doom the entire conversation by heavily 'discounting' the value of future moves.

The rest of the paper is organized as follows. Section 2.1 introduces *weighted ME games*—that is, ME games with *weights* or *scores* for each move of a play. The weights are accumulated over the entire play by the method of discounting. Section 2.2 extensively discusses a discounting factor to account for the penalties that the speakers incur from making disastrous discourse moves. As we show in Section 3, the discounting factor entails the existence of ϵ-Nash equilibria for weighted ME games, meaning that a notion of optimal rational play exists for our games. Section 4 applies our notion of weighted ME games to the Rubio/Christie exchange, while Section 5 considers related work. Section 6 concludes the paper.

2 The model

In this section we introduce Weighted Message Exchange games and formulate a discounting mechanism to accumulate the weights of the moves along a play.

2.1 ME and WME games

Definition 1 (ME game (Asher et al., 2016)) *A Message Exchange game (ME game) is a tuple* $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win_0, Win_1)$ *with* $Win_0, Win_1 \subseteq (V_0 \cup V_1)^\omega$.

V_0 and V_1 are called the *vocabularies* of players 0 and 1 respectively. The intuitive idea behind an ME game is that a conversation proceeds in turns where in each turn one of the players 'speaks' or plays a string of letters from her own vocabulary. However, the player does not speak any garbled sequence of strings but sentences or sets of sentences that 'make sense'. We capture by setting V_0 and V_1 to be SDRSs (Asher and Lascarides, 2003). See (Asher et al., 2016) for a detailed discussion on this topic and the motivation behind the formal setting of ME games.

Formally the ME game $\mathcal{G}$ is played as follows. Player 0 starts the game by playing a non-empty sequence in V_0^+. The turn then moves to Player 1 who plays a non-empty sequence from V_1^+. The turn then goes back to Player 0 and so on. The game generates a play p_n after n (≥ 0) turns, where by convention, $p_0 = \epsilon$ (the empty move). A play can potentially go on forever generating an infinite play p_ω, or more simply p. Plays are segmented into *rounds*—a move by Player 0 followed by a move by Player 1. A finite play of an ME game is (also) called a history, and is de-noted by h. Let Z be the set of all such histories, $Z \subseteq (V_0 \cup V_1)^*$, where $\epsilon \in Z$ is the empty history and where a history of the form $(V_0 \cup V_1)^+ V_0^+$ is a 0-history and one of the form $(V_0 \cup V_1)^+ V_1^+$ is a 1-history. We denote the set of 0-histories (1-histories) by Z_0 (Z_1). Thus $Z = Z_0 \cup Z_1$. For $h \in Z$, $\mathsf{turns}(h)$ denotes the total number of turns (by either player) in h.

We are interested in an extension of ME games where a *Jury* assigns a non-negative integer *weight* or *score* to every move by each player. The Jury then accumulates these weights in a way it deems suitable to compute the global score of the play for each player. In what follows, unless otherwise mentioned, i will range over the set of players, here $\{0, 1\}$. Thus, Player $(1 - i)$ denotes Player i's opponent.

Let $\mathbb{Z}$ be the set of all integers and $\mathbb{Z}_+$ be the set of non-negative integers. For any $n \in \mathbb{Z}_+$ let $[n] = [0, n-1] \cap \mathbb{Z}_+ = \{0, 1, \ldots, n-1\}$. A weight function is a function $w : (Z_0 \times V_1^+ \cup Z_1 \times V_0^+) \to \mathbb{Z} \times \mathbb{Z}$. Intuitively, given a history $h \in Z$, w assigns a tuple of integers $(a_0, a_1) = w(h, x)$ to the next legal move x of the play h. Note that the weight function, w depends on the current history of the game in that, given two different histories $h_1, h_2 \in Z$, it might be the case that $w(h_1, x) \neq w(h_2, x)$ for the same continuing move x. For notational simplicity, in what follows, given a play $p = x_0 x_1 \ldots$ of $\mathcal{G}$, we shall denote by $w_i^n(p)$, the weight assigned by w to Player i in the nth turn of p ($n \geq 1$). That is, if $w(p_{n-1}, x_n) = (a_0, a_1)$, then $w_0^n(p) = a_0$ and $w_1^n(p) = a_1$

Definition 2 (WME game) *A weighted ME game (WME game) is a tuple* $\mathcal{G} = ((V_0 \cup V_1)^\omega, w)$ *where* w *is a weight function.*

In Section 3, We will formally define a *Jury* who assigns weights to the moves of the game in a play p and accumulates them in a way it deems suitable to have a global evaluation of p for both the players. One of the standard methods for performing such an accumulation is 'discounting' (Shapley, 1953). In discounting, along a play p, the immediate moves are assigned high values and the moves further and further into the future are assigned lower and lower values. This is achieved by multiplying the weight of every subsequent move by a factor λ, which is usually fixed to be a constant between 0 and 1. However, in our case, to capture the context dependence of evaluations, we shall set λ to be a function of the history h, $\lambda : Z \to (0, 1)$.

Before fixing λ, we define first the discounted weight of a play and a discounted WME game.

Definition 3 (Discounted-payoff) *Let* p *be a play of* $\mathcal{G}$ *and let* λ *be a discounting function. Then the discounted-payoff of* p *for Player* i *is given by*

$$w_i^D(p) = \sum_{n \geq 1} \lambda(p_{n-1})^{n-1} w_i^n(p)$$

Definition 4 (Discounted WME game) *Let* w *be a weight function and* λ *be a discounting function. A discounted WME game with discount* λ *is a tuple* $\mathcal{G}_D[\lambda] = ((V_0 \cup V_1)^\omega, w)$ *such that for every play* p*, Player* i *receives a payoff of* $w_i^D(p)$.

When λ is clear from the context, we shall simply write $\mathcal{G}_D$ instead of $\mathcal{G}_D[\lambda]$. A (pure) strategy σ_i for Player i is defined in the standard way, $\sigma_i : Z_{1-i} \to V_i^+$. A play $p = x_0 x_1 x_2 \ldots$ conforms to a strategy σ_i of Player i if she always plays according to σ_i in p, that is, for every $j > 0$, $j - 1 = i \pmod 2$ implies $x_j = \sigma_i(p_{j-1})$. We denote by $p_{(\sigma_0, \sigma_1)}$ the unique play conforming to the tuple of strategies (σ_0, σ_1).

Definition 5 (Best-response / Nash-equilibrium) *A strategy* σ_i *of Player* i *is a best-response to a strategy* σ_{1-i} *of Player* $(1 - i)$ *if for every other strategy* σ_i' *of Player i, we have*

$$w_i^D(p_{(\sigma_i, \sigma_{1-i})}) \geq w_i^D(p_{(\sigma_i', \sigma_{1-i})})$$

Given $\epsilon > 0$, σ_i *is an* ϵ*-best-response to* σ_{1-i} *if for every other strategy* σ_i' *of Player i, we have*

$$w_i^D(p_{(\sigma_i, \sigma_{1-i})}) \geq w_i^D(p_{(\sigma_i', \sigma_{1-i})}) - \epsilon$$

A tuple of strategies (σ_0, σ_1) *is a Nash equilibrium (resp.* ϵ*-Nash equilibrium) if* σ_0 *and* σ_1 *are mutual best-responses (resp.* ϵ*-best-responses).*

We can also define natural notions of a *win*, *winning-strategy* etc. as follows, for both zero sum and non-zero sum games.

Definition 6 (Winning and winning strategy) *Let* $\mathcal{G}_D[\lambda] = ((V_0 \cup V_1)^\omega, w)$ *be a discounted WME game. Then (i)* **Zero-sum:** *Player* i *wins a play* p *of* $\mathcal{G}_D[\lambda]$ *if* $w_i^D(p) \geq w_{1-i}^D(p)$. *Player* $(1 - i)$ *wins* p *otherwise. (ii)* **Non-zero sum:** *Fix constants* $\nu_i \in \mathbb{R}$ *called 'thresholds'. Then Player* i *wins a play* p *if* $w_i^D(p) \geq \nu_i$. *(iii) A strategy* σ_i *is winning for Player* i *if she wins all plays* p *conforming to* σ_i.

2.2 The discounting factor

We now fix the exact form of the discounting factor λ to suit evaluations of conversational success. We assume that w is both integral and bounded, that is, the range of w is $[d]$ for some constant $d \in \mathbb{Z}_+$. A move with a weight of '0' is a 'failure' or a 'disastrous move' and heavily penalizes a player's future play. Also a move that gets weight 'd' is a 'brilliant move'; if such a move follows a disastrous move then it is a 'recovery move'.

For any history h, the function λ consists of two terms

$$\lambda(h) = \lambda_1 \lambda_2^{\frac{\mathsf{rec}_i(h)}{\mathsf{turns}(h)-1}}$$

The first is the global discounting which weighs initial moves more than later ones. This reflects the intuition: "get your best licks in first" - the player who does better initially often has an upper hand throughout the course of the debate. The second term is the 'punishing factor' that heavily discounts disastrous moves of a player. It 'kicks in' after the first disastrous move made by the player and gets worse if she keeps making such moves. A player may also recover from a disastrous move by making a number of brilliant moves, after which the punishing factor disappears, but might kick in again in the future. $\mathsf{rec}_i(h)$ is thus the 'recovery index' of Player i at history h and is computed using Algorithm 1 [note that the denominator of $(\mathsf{turns}(h) - 1)$ occurs in the index of λ_2 so that the number of turns does not affect it like it does for the global discounting λ_1].

Algorithm 1: REC$_i(h)$

```
data:$h$; result:$\mathsf{rec}_i(h)$
let rec_i = 0; good = 0
for j=1 to $\mathsf{turns}(h)$ do
    if $w_i^j(h) = 0$ then rec_i++
    if rec_i=0 then good=0
    if rec_i > 0 then
        if $w_i^j(h) = d$ then good++
    if good=c then rec_i--; good=0
return rec_i
```

Intuitively, Algorithm 1 starts accumulating the number of disastrous moves occurred. If Player i plays 'c' recovery moves after having played one or more disastrous move, the accumulated count of the disastrous moves decreases by 1. If i has fully recovered, it stops keeping track of the brilliant moves. The process repeats when i plays a disastrous move again.

3 Finite satisfiability and the Jury

We can now formalize the notion of the Jury. The Jury fixes the weights of the moves of the Players and also the parameters of the discounting function λ. That is, it fixes λ_1, λ_2 and c. Thus

Definition 7 (Jury) *The Jury for a discounted WME game $\mathcal{G}_D$ is a tuple $\mathcal{J} = (w, \lambda_1, \lambda_2, c)$ where w is a weight function.*

Although the game $\mathcal{G}_D$ can potentially go on forever, the Jury has to decide the winner after a finite number of turns. We can compute a bound on the number of turns after which the Jury can confidently decide the winner of the game. This is facilitated by the discounting of the weights and also the fact that w is integral and bounded. We have

Proposition 1 *Fix a discounted WME game $\mathcal{G}_D$ with a Jury $\mathcal{J} = (w, \lambda_1, \lambda_2)$ such that the range of w is $[d]$. Then given $\epsilon > 0$ we have for Player i and any play p of $\mathcal{G}_D$*

$$w_i^D(p) \leq \sum_{j=1}^{n_\epsilon} \lambda(p_{j-1})^{j-1} w_i^j(p) + \epsilon$$

where $n_\epsilon \leq \frac{\ln[\frac{\epsilon}{d}(1-\lambda_1)]}{\ln \lambda_1} - 1$.

Proof Suppose Player i does not play any disastrous move after n_ϵ turns. The maximum payoff she can gain after n_ϵ turns is $\lambda_1^{n_\epsilon+1} \frac{1}{1-\lambda_1} d$. Setting

$$\lambda_1^{n_\epsilon+1} \frac{1}{1-\lambda_1} d \leq \epsilon$$

we have $n_\epsilon \leq \frac{\ln[\frac{\epsilon}{d}(1-\lambda_1)]}{\ln \lambda_1} - 1$. ∎

Thus, if the Jury stops the game after n_ϵ turns, they can be sure no player would have gained more than ϵ, had the game been allowed to continue forever. Note that this result is fully general, but that values for n_ϵ will very much depend on the values set for λ_1 and λ_2.

Remark Note that it is crucial to assume that the players are unaware of the parameters of the Jury, w, λ_1, λ_2 and c. Otherwise, they can compute n_ϵ on their own. The game then becomes equivalent to a finite extensive form game with a set end, which is against the view on modeling strategic conversations defended in (Asher et al., 2016) that we have adopted. Thus, although the Jury takes a decision on the outcome of the game after a finite number of turns, the players do not know when that decision takes place. Thus, the game still appears to the players as potentially unbounded.

From Proposition 1, it also follows that ϵ-Nash equilibria always exist in our discounted WME games in pure strategies. However, since our space of strategies is uncountably infinite, the existence of Nash equilibria is a delicate matter (see for e.g. (Levy, 2013)) and we intend to explore it further in future work.

Corollary 1 *Given $\epsilon > 0$, a discounted WME game always has an ϵ-Nash equilibrium.*

Proof Consider the 'finite' discounted WME game for n_ϵ turns where n_ϵ is given by Proposition 1. Define the relation $\sim$ on plays of n_ϵ turns as: for two plays p and p', $p \sim p'$ iff for all $j : 1 \leq j \leq n_\epsilon$, $w_i^j(p) = w_i^j(p')$ and $w_{1-i}^j(p) = w_{1-i}^j(p')$. Clearly, $\sim$ is an equivalence relation. Also, since w is integral and bounded, there are only a finitely many possibilities for the weights of each Player i along any play p, and thus $\sim$ has finitely many equivalence classes. Thus there is a finite number of discounted payoffs possible (one for each equivalence class of $\sim$) after n_ϵ turns. A backward induction procedure on the equivalence classes of $\sim$ gives an ϵ-Nash equilibrium tuple of strategies $([\sigma_0], [\sigma_1])$ on these classes. Indeed, since by Proposition 1, no player can gain more than ϵ by deviating from it. Lifting $[\sigma_0]$ and $[\sigma_1]$ to corresponding representative elements of functions over actual histories gives us a required ϵ-Nash equilibrium (σ_0, σ_1). ∎

4 Applications

In Section 2, we developed weighting functions with two discounting parameters, λ_1 and λ_2 and a recovery constant c. λ_1 discounts future moves in the standard way agreeing with our intuition that good moves carry more value if played earlier than later. λ_2 is particular to WME games, that derives from agents' bad moves a penalty that adversely affects their score. c represents the number of brilliant moves required by a player to recover from a single disastrous move. These parameters are decided by the Jury. In this section we examine an WME game evaluation of our example dialogue, framed by the question as to whether Rubio has the experience to be president to be a dialogue on its own. The exchange is rather lengthy from the perspective of giving a complete discourse structure in which each clause is linked to other clauses via one or more rhetorical relations; this particular part of the political debate has over 200 clauses or elementary discourse unit (EDU). However, SDRT groups EDUs into more complex units or CDUs, small discourse graphs on their own that also have rhetorical links to other discourse units (Asher and Lascarides, 2003). As coherence is assured amongst the EDUs within the blocks, we will look only at the organization of CDUs and their relation to the whole dialogue, for it is there where the Jury has an important effect.

Our example is a fragment of a zero sum WME game. Let us denote the actual debate that unfolded between Rubio (R) and Christie (C), which is a play of the above game, as p_{RC}. Rubio's goal is to provide a convincing answer to the moderator's (M) question: to convince the public that he has the experience to be President. The goal of the antagonist, here C, is to destroy that answer, and C is very effective in doing that. Let us see how.

To do so, we will examine the role of the CDU blocks of the debate, which we've numbered in the introduction as (1)-(11), in the context of the Jury which is here the audience in the debate. For the sake of concreteness, we will take a particular integer scale and discount values for the weighting scheme; we feel that the scheme is defensible, though we acknowledge that there are many weighting schemes to choose from and we are unsure at this point exactly how to determine optimal weighting schemes or even whether such exits. We will also leave the tie between the details of the discourse structure and the weighting scheme relatively programmatic for now, as we have not fully figured out at present all the parameters of variation in this relation. Based on the Jury's evaluations and its applause reactions, we fix the range of w to be $[5] \times [5]$. We also fix $\lambda_1 = 0.9$ and $\lambda_2 = 0.5$. Thus the global discounting λ_1 is more or less gradual whereas the penalty discounting λ_2 for disastrous moves by either player is pretty severe. Let us also assume that the recovery constant $c = 5$. As we will show, these values fit the facts of the conversational sequences we have analyzed.

After the CDU introducing the question of political experience to R, R's response has 3 CDUS: 1) he talks about his record; 2) he argues that years of experience is not sufficient; years of experience aren't necessary either; 3) Obama with little experience knows exactly what he's doing (not necessary). We'll call (3) the *Obama CDU*. This seems to be a perfectly adequate response; it is responsive to the question and internally coherent. The audience applauds politely, and we could fix $w(\epsilon, \langle 1\rangle\langle 2\rangle\langle 3\rangle) = (3, 1)$. That is R (Player 0) gets a score of 3 for his points 1,2 and 3 which satisfactory but not overwhelming and C (Player 1) reaps only a minimal reward of 1 at this stage.

The moderator then invites C to comment on R's prior response. 4) C mounts a direct attack on R's record. 5) C also picks up on R's reference Obama but uses Obama as an example of disastrous government on the part of an inexperience one time senator, which indirectly attacks R as well. There are two points at which the audience applauds so we might set $w(\langle 1\rangle\langle 2\rangle\langle 3\rangle, \langle 4\rangle\langle 5\rangle) = (1, 4)$. C has a forceful reply and R gains only minimally from C's response.

Now R in (6) briefly responds with an attack on C's record as a problem solver but then in (7) returns to the Obama CDU. The problem is that the Obama CDU does not cohere with (6). R flubs the connection between the attack by implicating contrast (*"but let me add this"*), when he should have made an explicit reference back to C's use of Obama's record. While the point could have been effective, it wasn't rhetorically crafted in the right way, and the Obama CDU seems just to hang there, in addition to (7)'s being an almost verbatim repetition of (3). We could even imagine that C actually gains from R's dubious move. So here we let $w(\langle 1\rangle\langle 2\rangle\langle 3\rangle\langle 4\rangle\langle 5\rangle, \langle 6\rangle\langle 7\rangle)) = (1, 2)$. This inept response nevertheless does not kick in the penalty discount λ_2 for R yet, as λ_2 only makes a difference if there are moves evaluated with 0.

R's inept rhetorical connection and reuse of the Obama CDU gives C a crucial opening; C characterizes R's attack and the incoherently linked Obama CDU in a devastating way in (8). That is, (8) has the rhetorical function of commenting on the Obama CDU, not its content but its representation. With (8), C provides an evaluation of R's turn that capitalizes on its inept rhetorical structure. The audience sees the aptness of the characterization and roars its approval. Their evaluation coincides with C's, which means: $w(\langle 1\rangle\langle 2\rangle\langle 3\rangle\langle 4\rangle\langle 5\rangle\langle 6\rangle\langle 7\rangle, \langle 8\rangle) = (0, 5)$.

λ_2 now kicks in and since it is relatively low (0.5), R would have to do very well for the rest of the debate while C has to do very badly in order for R to win. We do allow that a long sequence of very good moves resets λ_2, but this seems to happen rarely. Actually, things get worse for R. In (10) R starts to deliver the Obama CDU again. Given (8), we can set $w(\langle 1\rangle\langle 2\rangle\langle 3\rangle\langle 4\rangle\langle 5\rangle\langle 6\rangle\langle 7\rangle\langle 8\rangle, \langle 9\rangle\langle 10\rangle)) = (0, 5)$, that is, it is a disastrous move for R while C's reputation is not hampered in any way. Moreover, C in (11) reuses his characterization again on R's contribution in (10), making $w(\langle 1\rangle\langle 2\rangle\langle 3\rangle\langle 4\rangle\langle 5\rangle\langle 6\rangle\langle 7\rangle\langle 8\rangle\langle 9\rangle\langle 10\rangle, \langle 11\rangle)) = (0, 5)$. At this point the contribution of the penalty discount, λ_2, is cubed ($= 0.125$), which is terrible for R. This makes C's characterization of his performance stick and affects the audience's (Jury) evaluations for the rest of R's turns.

We can now compute the discounted payoff to R and C respectively after these 3 rounds of p_{RC} as:

$$R : 3 + (0.95) \cdot 1 + (0.9)^2 \cdot 1 + (0.9)^3 (0.5) \cdot 0 + (0.9)^4 (0.5)^2 \cdot 0 + (0.9)^5 (0.5)^3 \cdot 0 = 4.76$$

$$C : 1 + (0.9) \cdot 4 + (0.9)^2 \cdot 2 + (0.9)^3 \cdot 5 + (0.9)^4 \cdot 5 + (0.9)^5 \cdot 5 = 16.10$$

Thus we see that after just 6 turns C has a overwhelming advantage over R in terms of his discounted payoff. Now suppose R tries to recover by playing brilliant moves (so as to neutralize the penalty discounting λ_2). That is, suppose he scores 5 for each of the subsequent 15 turns. Since $c = 5$, after each set of 5 turns the index of λ_2 will reduce by 1. A simple calculation shows us that the payoff to R after these 15 turns (that is a total of 6+15=21 turns) would be 9.63. After that, the penalty discounting λ_2 would disappear. But from then on the global discount λ_1 itself would start contributing heavily to the weights of the moves and we can show that even if R keeps playing brilliant moves forever, the maximum payoff he can receive from then on is just 5.47. Thus his total payoff in the infinite game after the initial slump is 9.63+5.47=15.10 which is still less than what C has amassed in the first 6 rounds (16.10). This justifies Proposition 1 and shows that the Jury can already offer the win to C (which it implicitly does).

What is crucial here is that C's attack on R's delivery rings true, and the fact that R could have attempted to rebut C's commentary but did not, confirms C's characterization of it. This affects the rest of the debate's evaluation; R's subsequent moves never mattered. In other words, the fate of R's evaluation was sealed after this initial exchange of 3 rounds. Thus, not responding to an attack on either the style or the substance of ones contributions forces the evaluation to go negative as in (Asher et al., 2016)'s general constraint.

(Asher and Paul, 2013) gives another example of a disastrous debate move. Though (Asher and Paul, 2013) does not use a weighing function and discounted payoffs, we can still apply our formalism to that example. The example concerns Senator Dan Quayle's (Q) reply to a similar question about his experience to be President in the 1988 Vice-Presidential debate, in which he drew a parallel between his own experience and that of President John Kennedy (K). His opponent, senator Lloyd Bentsen (B), took a weak implicature from Q's response, that Q had the potential to be a similar president to K, and attacked it forcefully, drawing a roar of appreciation from the audience, giving Q a score of 0 for that move. Q's subsequent rejoinder *"that was unfair Senator, unfair,"* was a comment that did not take issue with B's drawing of the implicature concerning Q and K. This amounted to a tacit acceptance of the implicature. Given that B had refuted that implicature, Q was saddled with having conveyed an implicit content that he was unable to defend but accepted, which netted him a second zero, which was enough to sink his performance for the rest of the debate. B's attack move, though different from C's in (8) in that it attacked content not presentation, also colored Q's performance for the rest of the debate. Q's evaluation went to the bottom of the scale for the rest of the debate and stayed there, making B the clear winner.

We have modeled the consequences of disastrous moves on evaluations of a conversational play. But what about brilliant moves that are not attacks, how do they function? One memorable line used over by Ronald Reagan during the 1980 US Presidential campaign was *"Are you better off than you were four years ago?"* In one question, Reagan was able to remind Americans that they were worse-off under the incumbent Carter; inflation and unemployment had dramatically risen under Carter and purchasing power has waned. Carter himself described the American mood as a *"malaise"* during his Presidency. This one move set the tone for the discussion and put Reagan in a winning position, as Carter could not convincingly counter the obvious *"no"* answer to Reagan's question.

We can model the above in our setting of WME games with w assigning a 5 to this move by Reagan and a 0 to Carter. Carter's inability to respond convincingly saddles him with another 0 and this colors the evaluation by w of the ensuing debate, heavily favoring Reagan. Reagan continues to get high scores for all his moves while Carter fares badly, which accords with history: Reagan was pronounced a clear winner of the exchange.

5 Related Work

As alluded to in the introduction, game theory has been used before in the literature for the analysis of strategic message-exchange. The focus for the purpose has mostly been on the use of signaling games (Spence, 1973). However, signaling games lack the necessary tools to model situations where the interests of the players are opposed, as is the case in the current setting. Noteworthy also is the work on persuasion games (Glazer and Rubinstein, 2001; Glazer and Rubinstein, 2004) which has the setup similar to that of signaling games where a 'speaker' is trying to persuade an uninformed 'listener' about the current state of the world. Despite being hugely successful in modeling many different economic and strategic situations, signaling games have certain drawbacks which restricts their applicability to dynamic strategic conversations, as in the current setting. This issue has been extensively discussed in (Asher et al., 2016).

Our notion of evaluation makes use of discourse structural moves and depends on work on discourse structure and rhetorical relations like that of (Asher and Lascarides, 2003); to our knowledge, we are the first to model evaluations of conversational success by exploiting ideas of discourse coherence and discourse structure, along with techniques of discounting from game theory. Our account also makes at least informal use of the notion of an attack, and is thus related to work on argumentation (Dung, 1995; Besnard and Hunter, 2008). (Besnard and Hunter, 2008) also considers a definition for evaluating an argument by an audience. They structure arguments as trees, which roughly parallels the notion of a discourse graph in SDRT (Stede et al., 2016). They also use a discounting function, so that more deeply embedded arguments (responding to prior attacks) are weighted less than the main arguments and counterarguments at the top. This discounting function is similar to our λ_1. However, there is nothing in the argumentation literature of the form of our penalty discount λ_2 for convincing attacks and very bad moves. And to our knowledge, no one in the argumentation literature, or anywhere else, has tried to formalize an evaluation of attacks and refutations over the course of a dialogue. The analysis of argumentation in game theoretic terms, which is a consequence of our approach, is also the first of its kind to our knowledge.

Evaluations of conversational success are also related to linguistic work on predicates of taste (Lasersohn, 2005; Glanzberg, 2007; Crespo and Fernández, 2011), in that our evaluations are relative to the standards of a person or group. It may be that two people may disagree over a evaluation of i's contributions, because they have incompatible views of what constitutes conversational success for i, just as people may disagree about whether say blood sausage is tasty or not. The received wisdom about predicates of taste, however, is there is 'no fact of the matter' as to whether blood sausage is tasty or not. We do not believe this carries over to evaluations of conversational success. Given that players in a political debate have the goal of convincing the public, it is really the public's evaluation that counts and gives an 'objective' evaluation of the player's success in terms of their own interests. Work on automatic debate evaluation in terms of an audience's reactions has attracted interest in NLP (Prabhakaran et al., 2012; Prabhakaran and Rambow, 2013; Prabhakaran et al., 2013), for which weighted ME games provide a formal framework.

6 Conclusions

We have presented a model of the evaluation of conversational success, WME games. Extending the framework of infinite ME games for modeling conversations introduced in (Asher et al., 2016), we have shown how a Jury can concretely evaluate a player's conversational success. We have illustrated how such evaluations depend upon the structure and content of a person's contributions as well as on discounting functions, and we have analyzed at length one sample conversation to show an evaluation process at work. Our discounting functions entail: (i) it is best to get one's very good moves in early, (ii) a sequence of moves that are bad by Player i affects the evaluation of future moves, and in particular, (iii) a failure by i to respond effectively to a convincing attack on i's earlier moves is disastrous, because λ_2 becomes very significant.

There are many ways in which we wish to extend this work. First, we want to explore further the space of weighting and discounting functions; different functions will yield new and potentially interesting evaluation schemes. Secondly, we wish to enrich our model with an epistemic framework by introducing imperfect information (Harsanyi, 1968). In the present abstract, as remarked, we assume that the players are unaware of the parameters of the Jury. Elaborating on this, we might assume that a Jury can be of different 'types'. For instance, it may be 'biased' towards a particular player or may be 'fair' to everybody. It may be 'patient' (with high λ_1) or 'impatient' (with low λ_1); 'strict' (with low λ_2) or 'lenient' (with high λ_2). In addition, the players might themselves be of different types: risk-takers, risk-aversers, rational, irrational etc. Players are aware of their own types but are uncertain about the types of the other players and that of the Jury; they hold certain 'beliefs' about these unknown types. A player's strategy now depends not only on the history but her own type and her beliefs about the types of the other players and that of the Jury. Such an approach is standard in epistemic game-theory and we believe that augmenting the current framework of WME games with it will lead to a much more complete analysis of the behavior of conversationalists and evaluations of conversations.

Finally, we wish to explore the existence of Nash equilibria and other solution concepts in our WME games and explore rationality criteria.

References

Nicholas Asher and Alexandra Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Nicholas Asher and Soumya Paul. 2013. Conversations and incomplete knowledge. In *SEMDIAL 2013: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, pages 173–176. Citeseer.

Nicholas Asher, Soumya Paul, and Antoine Venant. 2016. Message exchange games in strategic conversations. *Journal of Philosophical Logic*. In press.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Christie-Rubio-debate. 2016. `http://www.realclearpolitics.com/video/`

`2016/02/06/christie_vs_rubio_the_memorized_30-second_speech_where_you_talk_about_how_great_america_is_doesnt_solve_anything.html`.

Christie-Rubio-transcript. 2016. `https://www.washingtonpost.com/news/the-fix/wp/2016/02/06/transcript-of-the-feb-6-gop-debate-annotated/`.

Vincent Crawford and Joel Sobel. 1982. Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Inés Crespo and Raquel Fernández. 2011. Expressing taste in dialogue. In *SEMDIAL 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 84–93. Citeseer.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Michael Glanzberg. 2007. Context, content, and relativism. *Philosophical Studies*, 136(1):1–29.

Jacob Glazer and Ariel Rubinstein. 2001. Debates and decisions, on a rationale of argumentation rules. *Games and Economic Behavior*, 36:158–173.

Jacob Glazer and Ariel Rubinstein. 2004. On optimal rules of persuasion. *Econometrica*, 72(6):119–123.

Herbert Paul Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press.

John Harsanyi. 1968. Games with incomplete information played by ‘bayesian’ players, parts i, ii & iii. *Management Science*, 14(7):486–502.

Peter Lasersohn. 2005. Context dependence, disagreement, and predicates of personal taste*. *Linguistics and philosophy*, 28(6):643–686.

Yehuda Levy. 2013. Discounted stochastic games with no stationary nash equilibrium: Two examples. *Econometrica*, 81(5):1973–2007, September.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *IJCNLP*, pages 216–224.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ajita John, and Dorée D Seligmann. 2013. Who had the upper hand? ranking participants of interactions based on their relative power. In *IJCNLP*, pages 365–373.

Lloyd S. Shapley. 1953. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095–1100.

Michael A. Spence. 1973. Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jeremy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of LREC*.

Disagreement with evidentials: A call for subjectivity

Natalia Korotkova
UCLA & University of Tübingen
`n.korotkova@ucla.edu`

Abstract

Across languages, grammatical evidentials (linguistic expressions of information source) exhibit the property of non-challengeability: they resist direct denial in dialogues. The literature attributes this property to the not-at-issue status of the information contributed by evidentials. I argue against this view and show that with respect to disagreement, evidentials pattern with subjective expressions such as first-person belief and pain reports. Like other subjective expressions and unlike e.g. appositives, evidentials ban *all* kinds of disagreement about content and not just explicit denial. This novel observation has no account in the literature. It falls out naturally once a theory of evidentiality incorporates subjectivity. It is thus unnecessary to appeal to a special discourse status of evidentials to explain their behavior in conversations.

1 Introduction

Evidentiality is a linguistic category that marks the information source for the proposition expressed by a sentence (a.o. Chafe and Nichols, 1986; Willett, 1988; Aikhenvald, 2004; de Haan, 2013b,a). English can express information source lexically, e.g. by adverbials as in (1) below:

(1) Threatened by climate change, Florida **reportedly** bans term ‘climate change’.[1]

The sentence in (1) conveys, via *reportedly*, an evidential adverbial, the idea that the speaker does not have firsthand knowledge about the terminology ban but rather learned about it via hearsay.

This paper focuses on grammaticalized evidentials. Many of the world’s languages—237 out of 414 surveyed by de Haan (2013b)—have designated morphological means to talk about information source, e.g. visual vs. non-visual perception, inference from reasoning, or hearsay; see (Willett, 1988) for an overview of source types. Three types of information source commonly signalled by evidential markers are exemplified below by the Cuzco Quechua evidential paradigm:

(2) Cuzco Quechua (Quechuan)

a. Para-sha-n=**mi**. [PERCEPTION]
rain-PROG-3=**DIR**
‘It is raining, *I see*.’

b. Para-sha-n=**si**. [HEARSAY]
rain-PROG-3=**REP**
‘It is raining, *I hear*.’

c. Para-sha-n=**chá**. [CONJECTURE]
rain-PROG-3=**CONJ**
‘It must be raining, *I gather*.’

(based on Faller, 2002: 3, ex.2a-c)

Each sentence in (2) talks about the proposition ‘It is raining’, which will be referred to as the *scope proposition* throughout (after Murray, 2014). Evidentials =*mi*, =*si* and =*chá* specify the way the speaker learned the scope proposition: firsthand as in (2a), via hearsay as in (2b), or via conjecture as in (2c). The contribution of evidentials will be referred to as the *Evidential Requirement* (ER).

A hallmark of morphological evidentials is their *non-challengeability*: a direct denial can only target the scope proposition, but not the ER ((Izvorski, 1997) and later work).[2] This property is illustrated with a Cuzco Quechua dialogue:[3]

(3) Cuzco Quechua

A. Ines-qa qaynunchay ñaña-n-ta=**n**
Inés-TOP yesterday sister-3-ACC=**DIR**
watuku-rqa-n
visit-PST-3
‘Inés visited her sister yesterday, *I saw*’.

B. Mana=n chiqaq-chu.
not=DIR true-NEG
‘That’s not true.’

(i) $= \neg$ [Inés visited her sister]
(ii) $\neq \neg$ [You saw that Inés visited her sister]

(based on Faller, 2002: 156, ex. 116-117b)

---

[1] From *Washington Post*; `https://goo.gl/XVx6JN`.

[2] Premises for making a conclusion may be challenged, e.g. faulty logic or an untrustworthy source (Faller, 2007).

[3] In (3), =*n* is a morphophonemic variant of =*mi*.

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*,
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 65-75.

The reaction in (3) can only indicate disagreement with the scope proposition. For instance, it can be followed up with 'Inés only visited her mother' (Faller, 2002: 158: ex.119). However, (3) cannot be understood as a disagreement with the ER, and the follow-up 'You didn't see it' results in infelicity (Faller, 2002: 158: ex.118). The same holds for other Cuzco Quechua evidentials. To sum up, it is impossible to challenge the fact that the speaker acquired the scope proposition in a way lexically specified by the evidential.[4]

The pattern illustrated in (3) is observed in many other geographically unrelated languages, e.g. in Bulgarian (South Slavic; Izvorski, 1997), Cheyenne (Algonquian; Murray, 2014), Georgian (South Caucasian; Korotkova, 2012), German (Germanic; Faller, 2007) and St'át'imcets (Salish; Matthewson et al., 2007). Based on the data from available studies of evidentiality within formal semantics, the non-challengeability of the ER is a *universal* property of morphological evidentials.

The central puzzle addressed in this paper is as follows: what bans disagreement with the ER?

The near-consensus in the literature (a.o. Izvorski, 1997; Matthewson et al., 2007; Murray, 2014) is that statements with evidentials make two contributions. The scope proposition constitutes the main point of an utterance and thus enjoys the *At-Issue* (AI) status. The ER, one the other hand, is analyzed as a kind of peripheral, *Not-At-Issue* (NAI), information (in the sense of Tonhauser et al., 2013). Relying on the view on discourse wherein conversational disagreement is derived solely from the AI vs. NAI divide (e.g. Potts 2005, Roberts 1998/2012), the non-challengeability of the ER is explained via its discourse status: by definition, NAI cannot be targeted by direct responses.

I argue that the view above is not justified empirically. The argument proceeds in two steps.

First, I show that non-challengeability does not carve out just NAI (pace Tonhauser et al., 2013). It is also an inherent trait of a host of expressions that I will call *subjective*, such as first-person belief reports or statements about pain. The source of non-challengeability is different in each case. For NAI, non-challengeability results from the special discourse status of the information conveyed by a given construction. For subjective expressions, it is their lexical semantics that bans disagreement. Such expressions describe experiences to which individuals have exclusive access (e.g. mental states) and which others have no grounds to contest. This non-linguistic fact has linguistic consequences: denial is an infelicitous reaction to statements with subjective expressions. Summing up, non-challengeable content comes in at least two varieties: (i) NAI content and (ii) *Subjective Content* (SC). This means that direct denials alone are not indicative of the NAI nature of the ER, contrary to the accepted wisdom.

Second, I show that NAI on the one hand, and SC on the other, part company when examined against a variety of disagreement strategies. While there are ways to disagree with presuppositions or appositives (typical representatives of the NAI class), subjective expressions resist *all* kinds of disagreement about content. Based on novel data from Bulgarian and Turkish,[5] I argue that evidentials exhibit the same kind of strong non-challengeability as subjective expressions do. I further demonstrate that the only kind of disagreement allowed for evidentials and e.g. first-person belief reports is what I refer to as "performance disagreement" (the term from (Anand, 2009) on similar facts about taste ascriptions): a situation when the speaker is considered incompetent (e.g. drunk) or insincere (e.g. lying) by their addressee. The overall pattern of disagreement with evidentials is not easily amenable to an NAI analysis. Such an analysis incorrectly predicts that disagreement with evidentials should be possible modulo the constraints on propositional anaphora. I thus conclude that the data from various kinds of denials (A) do not support the NAI view of evidentiality and (B) call for a new, subjective, approach.

The paper is structured as follows. Section 2 presents two analytical options that explain non-challengeability away, the NAI status and subjectivity, and explores their applications to evidentiality. Section 3 demonstrates that non-challengeable expressions do not form a uniform class with respect to various kinds of disagreement and that evidentials do not pattern with NAI. Section 4 is on performance disagreement. Section 5 concludes.

---

[4] The current paper only discusses root declarative clauses, so it is correct to say that evidentials are always anchored to the speaker. Elsewhere, they may flip: to the addressee in questions (Lim, 2010; Murray, 2010), and to the attitude holder in attitude reports (Korotkova, 2015).

[5] The data come from my work with consultants.

2 Direct denial

This section discusses two routes to banning direct denials: (1) via NAI status, and (2) via subjectivity. The former route reflects the now-standard view that disagreement is reducible to the AI vs. NAI distinction, and it is widely taken in the literature on evidentiality. The latter route is never addressed head-on with respect to evidentials. I will show that this neglected route is a viable alternative to the ER-as-NAI mantra.

2.1 Route 1: NAI content

Issues in discourse Recent research on conversational dynamics identifies different types of content (Potts, 2005; Simons et al., 2010; Tonhauser, 2012; Tonhauser et al., 2013; Gutzmann, 2015):

- ○ AT-ISSUE (AI): information central to the issues discussed
- ○ NOT-AT-ISSUE (NAI): peripheral information

NAI does include presuppositions (what is taken for granted), but also new information that constitutes a comment rather than the main point of an utterance, e.g. conventional implicatures (Potts, 2005) (though see (Schlenker, 2013) for a presuppositional analysis of Potts' cases).

Is there a relation between the structure of discourse and grammar? Natural language is sensitive to the AI vs. NAI divide and has designated means to mark it, e.g. focus:

(4) Where did Kit spend his vacation?
  a. ✓Kit flew to CALIFORNIA.
  b. #KIT flew to California.

As examples like (4) show, English prosodic focus highlights what the issue under discussion is. Only (4a) is a felicitous reply while (4b) is out, as it suggests that the question asked is about people who flew to California.

Non-challengeability of NAI The divide is obviously important in determining the range of replies to questions and reactions to assertions. Often it is argued that the divide is *solely* responsible for patterns of conversational disagreement (cf. (Amaral et al., 2007; Anderbois et al., 2015) and diagnostics 1a,b,c in (Tonhauser, 2012)):

- ○ A direct response has to target AI.
- ○ NAI cannot be targeted by a direct response.

These patterns are familiar from presuppositions, which one cannot explicitly deny (5ii):

(5) PRESUPPOSITIONS
  A. The queen of the US visited Jupiter.
  B. That's not true.
 (i) $=$ She visited Mars.
 (ii) $\neq \neg$ [The US has a queen].

More recently, a number of constructions have been analyzed as a vehicle for the not-at-issue content based in particular on their non-challengeability: appositives and non-restrictive relative clauses (Potts, 2005), expressives such as *darn* (McCready, 2008, 2010), and various parentheticals (Potts, 2002; Simons, 2007):

(6) EXPRESSIVES
  A. That **damn** Ortcutt lost his passport.
  B. That's not true.
 (i) $= \neg$ [Ortcutt lost his passport]
 (ii) $\neq \neg$ [There is something wrong with Ortcutt]

(7) APPOSITIVES
  A. Ortcutt, **a spy**, lost his passport.
  B. That's not true.
 (i) $= \neg$ [Ortcutt lost his passport]
 (ii) $\neq \neg$ [Ortcutt is a spy]

Direct responses such as *That's not true* cannot target the semantic contribution of *damn* (6ii), or the content of an appositive (7ii). Similar results hold for other types of response, such as *That's right*: one can only agree with what is at-issue.

ER as NAI Recall from (3) that a direct denial can only target the scope proposition and never the ER. This is the same pattern as the one exhibited by expressions under the NAI umbrella. Not surprisingly, formally different approaches to evidentiality meet at one point: the ER is treated as a kind of NAI content (first proposed by Izvorski (1997)).

The ER-as-NAI view is widely accepted. The approaches range from presuppositional (Izvorski, 1997; McCready and Asher, 2006; Matthewson et al., 2007; Lee, 2013) to ones where the ER is a part of sincerity conditions associated with a speech act (Faller, 2002) to ones where the ER is paralleled to Pottsian supplements (Murray, 2010, 2014; Koev, 2016). Modulo the technical and conceptual differences, the key intuition of these theories is that the ER is an *automatic* restriction on the common ground and as such is never up for negotiation by the interlocutors. The ban on explicit denial is thus correctly predicted.

A common trait of the above proposals is that, out of several empirical means to diagnose discourse status (see e.g. Tonhauser, 2012), the only one used is the non-challengeability test. As I will argue throughout the paper, the denial pattern lends itself to an alternative explanation and thus is not indicative of the NAI status of the ER.

Additional arguments for the ER-as-NAI view come from projection (=escaping the scope of entailment-cancelling operators). However, recent research challenges Simons et al. (2010)'s idea that discourse status and projection go hand in hand (see (Jasinskaja, 2016) for discussion)—it is possible to project and exhibit properties of AI (sentence-final appositives, see section 3). Furthermore, the overall cross-linguistic profile of evidentials with respect to projection is largely understudied. For instance, across languages the ER is not affected by the clause-mate negation (de Haan, 1997: 146-170), which is almost always taken as an instance of projection:

(8) Georgian
sup'-i ar **gauk'etebia**
soup-NOM NEG make.**IND**.PST
$p =$ 'S/he made a soup'
(i) $\neq$ [I hear/infer p] $\wedge$ [$\neg p$] **projects**
(ii) $\neq \neg$ [I hear/infer p] **narrow scope**, EV$< \neg$
(iii) $=$ [I hear/infer $\neg p$] **wide scope**, EV$> \neg$

As Murray (2010) (but not Murray, 2014) and Tonhauser (forth.) correctly point out, the only available interpretation is an instance of the evidential outscoping clause-mate negation, which in turn creates an illusion of projection.[6] Moreover, non-challengeability does not correlate with projection: while all evidentials are non-challengeable, some of them may have narrow scope in conditionals, e.g. Tagalog (Kierstead, 2015), or in attitudes, e.g. Turkish and Korean (Korotkova, 2015). In light of this, the data on disagreement are essential for modeling evidentiality.

In all incarnations of the view above, the speaker's having acquired p in a particular way is treated as an objective fact. That this information has to be channeled as NAI seems to be an arbitrary property of grammar, and things could have been otherwise. I present an alternative view wherein the non-challengeability of some elements is a direct effect of what they mean.

2.2 Route 2: Subjectivity

Subjectivity Individuals have privileged and exclusive access to certain information about themselves, through senses and introspection: (A) mental states, e.g. having a desire, (B) feelings, e.g. being angry or sad, (C) some bodily sensations, e.g. pain or hunger. Self-knowledge obtained via these channels is *incorrigible*: the experiencer has a special epistemic status and others have no grounds to deny such knowledge.[7] If I am, say, tired, I am the only authority over this state of mine.[8]

I will call linguistic expressions that describe such experiences as above *subjective*. The category of Subjective Content (SC) includes, e.g., first-person (A) attitude reports (*I hope*), (B) taste ascriptions (*It tastes good to me*), (C) psych verbs (*I am excited*), and (D) statements about pain (*It hurts*).[9] I demonstrate, using conversational disagreement as an example, that some features of the linguistic behavior of SC stem from intrinsic properties of the experiences it talks about.

Non-challengeability of SC Incorrigibility of knowledge obtained via subjective experiences restricts the range of reactions to SC in the following way. Only the experiencer has access to said experiences, so genuine disagreement is impossible:

(9) FIRST-PERSON PAIN REPORT
A. **I** have a splitting headache.
B. #No, that's not true.

By virtue of self-knowledge about pain being incorrigible (a non-linguistic fact), B cannot felicitously disagree (a linguistic fact) with A about A's pain (9).[10] In third-person pain reports (10), the speaker and the addressee both have low epistemic status, and non-challengeability evaporates:

(10) THIRD-PERSON PAIN REPORT
A. **Mo** has a splitting headache.
B. ✓No, that's not true.

Other subjective expressions exhibit the same pattern with respect to non-challengeability of first-person statements (11, 13) and lack thereof for their third-person counterparts (12, 14).

---

[6] Sharvit (2015) makes a similar observation about the pseudo-projective behavior of *only*.

[7] I am not taking sides in the debate on the infallibility—complete immunity to error—of such self-knowledge (see e.g. (Aydede, 2013) on pain). Of importance here is that only the experiencer has access to certain experiences, regardless of whether it is logically possible for them to be mistaken.

[8] Bodily awareness isn't always incorrigible (de Vignemont, 2015). Even though proprioception offers a unique experience of one's body, mistakes about e.g. spatial orientation are possible and may be corrected by others.

[9] The notion is broader than the usually recognized first-person content such as attitudes 'de se' (Moltmann, 2012).

[10] B may disagree with (9) if B thinks that A (a) is being insincere or (b) is not correctly assessing their own experience. I ignore such pragmatically odd situations until section 4.

(11) FIRST-PERSON PSYCH PREDICATE
  A. Sauerkraut disgusts **me**.
  B. #That's not true.

(12) THIRD-PERSON PSYCH PREDICATE
  A. Sauerkraut disgusts **all vegans**.
  B. ✓That's not true.

(13) FIRST-PERSON BELIEF REPORT
  A. **I** think that there is life on Mars.
  B. That's not true.
 (i) $= \neg$ [There is life on Mars]
 (ii) $\neq \neg$ [You think that there is life on Mars]

(14) THIRD-PERSON BELIEF REPORT
  A. **Mo** thinks that there is life on Mars.
  B. That's not true.
 (i) $= \neg$ [There is life on Mars]
 (ii) $\neq$ [She thinks that there is life on Mars]

SC resists third-party assessment in general, which is responsible for its non-challengeability in dialogues. Therefore, non-challengeability does not uniquely diagnose NAI (Anand (2007: 203) makes a similar point).[11] Still, subjectivity is not brought up in the literature on conversational dynamics (Simons et al., 2010; Tonhauser, 2012).

On the other hand, that subjectivity bans some kinds of disagreement is explicitly acknowledged for epistemics such as *likely* and Predicates of Personal Taste (PPT) such as *delicious*. While it is possible to disagree with e.g. a PPT-statement, the disagreement is of a different nature:

(15) A. Sauerkraut is disgusting.
  B. ✓No, it is delicious.

Dialogues as in (15) are referred to as *faultless disagreement* (Kölbel, 2003), a situation when the two parties disagree without one of them being strictly wrong. B's statement is only felicitous so long as B is making a claim about oneself or a generic statement ($\approx$ 'People in general like sauerkraut'). Such a *No* never contests the speaker's epistemic state or perception, a move deemed infelicitous precisely on the grounds I discuss—these are private experiences (Stephenson, 2007; Anand, 2009; von Fintel and Gillies, 2011). Subjectivity also predicts the infelicity of a *No* in reply to a PPT-statement with an overt experiencer, such as *It is disgusting to me*.

The possibility of faultless disagreement, as well as the the possibility of retractions—disagreement with one's previous statements, is at the core of the contextualism-relativism-expressivism debate on the proper analysis of PPTs and epistemics; see (Weatherson and Egan, 2011) and (MacFarlane, 2014: 1-25) for an overview. Using Weatherson and Egan's helpful analogy, epistemics and PPTs resemble *we* in that they have a 'communal' component, formalized e.g. as assessment-sensitivity (Stephenson, 2007; MacFarlane, 2014), genericity (Anand, 2009; Pearson, 2013) or group-relativity (von Fintel and Gillies, 2008). This component is part of their conventional meaning and is not inherent to subjective expressions, as evidenced e.g. by the contrast between taste ascriptions via PPTs (15) and via psych verbs (11); cf. also discussion in (Anand, 2009).

ER as SC Recall that it is illicit to deny the ER:

(16) German (Germanic)

A.

Es	**soll**	regnen	am	Wochenende.
It	**REP**	to.rain	at.DAT.DEF	weeekend

'It will rain on the weekend, *I hear*'.
$\approx$'It is supposed to rain on the weekend.'

B.

Nein,	das	stimmt	nicht.
No	that	be.true.PRES	NEG

'No, that's not true'

 (i) $= \neg$ [It will rain]
 (ii) $\neq \neg$ [You heard it it will rain]
 (iii) $\neq \neg$ [I/we all heard it will rain]

The hearsay use of German *sollen* ($\approx$'must') also exhibits non-challengeability (Faller, 2007). I propose that this universal pattern is amenable to a subjective analysis.

Acquisition of some proposition is always associated with a mental state formed thereafter. Some *conjectural* and *inferential* evidentials, e.g. Cuzco Quechua =*chá* (2c), refer to mental states directly by indicating that the scope proposition was acquired via reasoning from general knowledge. Other evidentials describe mental states mediated by perception. (A) *Direct* evidentials such as Cuzco Quechua =*mi* (2a, 3) involve immediate perception. (B) *Hearsay* evidentials such as Cuzco Quechua =*si* (2b) and German *sollen* (16) denote having heard (or read) a report. (C) *Indirect* evidentials—ones denoting either hearsay or inference from results, such as Bulgarian *-l* and Turkish *mIş* (discussed in section 3 below)—refer to, respectively, perceiving results or reports. Whichever the channel, denying that the speaker acquired the scope proposition in a given

---

[11] (13) is susceptible to an explanation along the lines of (Simons, 2007): the proferred content, but not the matrix verb, constitutes the main point. Such an analysis fails to predict (A) the contrast between first- and third-person attitudes with respect to disagreement, and (B) the pattern exhibited by SC across the board. In e.g. (9) and (12) one's headache and preferences are clearly at-issue as the sentences can answer questions about, respectively, one's well-being and likes.

way amounts to questioning their introspection and perception—and this, in turn, is infelicitous. The formal analysis is proposed in (Korotkova in prep.), where I also argue that faultless disagreement (16iii) is banned due to indexicality.

Similar effects in fact hold for English. Even though the language lacks grammatical evidentials, information source can be signalled by other means, as in *I saw that it hailed.* A reply *No, you didn't* is infelicitous: regardless of what the addressee thinks the speaker has observed, only the actual speaker has access to their perception.[12]

If explicit performatives (*I promise*) are true by say-so, linguistic subjectivity can be described as true by *feel-so*: *It hurts* is true if the speaker is sincere. The ER, under the view sketched above, behaves the same way and thus is non-challengeable. Faller (2002), who likens the ER to mental acts of evaluation, observes this parallel between performatives and evidentials. However, Faller does not discuss linguistic and non-linguistic subjectivity, and derives the non-challengeability of the ER from the level of meaning evidentials operate at.

Garrett (2001: Chapter 4, 102-206), too, discusses the truth by say-so effects of evidentials and appeals to the privileged status of some information to describe constraints on what he calls *ego evidentiality* in Tibetan, a category that describes internal knowledge about a situation. The proposal I put forth is different. I argue that *all* evidentials denote experiences to which individuals have exclusive access, regardless of the source. Besides, the status of ego evidentiality as evidentiality proper is debated, and it may better fit under the egophorocity umbrella (Floyd et al. forth).

2.3 Interim summary

The landscape of disagreement patterns requires rethinking. I show that (not-)at-issue status is not the only source of impossibility of direct denials and that subjectivity is another plausible solution. To this end, I delineate an approach to evidentiality such that the speaker is the one and only authority over the way they acquired the scope proposition. This view derives direct denials equally well compared to the ER-as-NAI approaches. The next section discusses where the two options diverge.

3 Other types of denial

Direct denials of the form *No, that's not true* do not distinguish between NAI content and SC: both are non-challengeable. Thus, as far as evidentials are concerned, each line of analysis will get the direct denial data right. I show that the two different sources of non-challengeability yield different patterns with respect to other denial strategies and argue that evidentials pattern with SC.

NAI content is backgrounded, which limits the range of discourse operations applicable to it. In this case, form matters. Direct denials become possible if the same content is conveyed via regular clausal coordination. Direct denials are more likely (Syrett and Koev, 2015: Experiment 2) for sentence-final non-restrictive relatives (17a) as opposed to non sentence-final ones (17b):

(17) a. The photographer took a picture of Catherine, **who is an experienced climber**.
b. Catherine, **who is an experienced climber**, made it to the summit.
(Syrett and Koev, 2015: App.A, ex.5)

Jasinskaja (2016) argues that positional effects follow from a more general constraint on salience associated with propositional anaphora such as *that*.

Additionally, special discourse moves are allowed to target NAI. *Hey, wait a minute* (proposed by von Fintel (2004) for identifying presuppositions) may target appositives, and in fact prefers to (Syrett and Koev, 2015: Experiment 1).

SC, on the other hand, cannot be challenged across the board: the addressee has no epistemic authority for disagreement, and e.g. *Hey, wait, you are not* in reply to *I'm in pain* is bizarre at best.

The asymmetry in licensing disagreement can be used as a benchmark for evidentials. If some kinds of disagreement are allowed, it is an argument for the dominant ER-as-NAI view (section 2.1). If denials are banned altogether, it is an argument for a subjective approach (section 2.2):

	NAI	SC	ER
That's not true	☹	☹	☹
Other types of denial	✓	☹	??

Table 1: Licensing disagreement

Below I discuss novel data from Bulgarian (South Slavic) and Turkish (Turkic) on the availability of two kinds of denials, *No, that's not true* and *You are mistaken*, for (A) NAI: presuppositions and appositives, (B) SC: pain and attitude reports, and (C) evidentials. None of these

---

[12] A reviewer notes that it is possible to disagree with such English statements. I argue that such cases can be subsumed under performance disagreement, discussed in section 4.

expression types allow *No, that's not true*. *You are mistaken*, being more flexible than propositional anaphora, may target NAI but, given the lack of epistemic authority on part of the addressee, cannot target SC. Evidentials ban both reactions.[13]

NAI In both Bulgarian and Turkish, presuppositions introduced by *too* (Appendix A) and the content of appositives (18, 19) can be disagreed with using *You are mistaken* (with a follow-up specifying what the mistake is about) but cannot be targeted by direct denial (even with a follow-up).[14]

APPOSITIVES

(18) Bulgarian

A. Kalifornija, **naj-golemijat** **štat**, legalizira
California **the.largest** **state** legalize.PST
marixuana-ta
marijuana-DEF
'C., the largest state, legalized marijuana.'.

B. Ne, ne e vjarno.
No NEG be.PRES true
'No, that's not true'.

(i) = ¬ [California legalized]
(ii) ≠ ¬ [California is the largest state] (even with a continuation such as *Alaska is the largest state*)

B'. Bărkaš.
be.mistaken.2SG
'No, you're mistaken'.

(i) = ¬ [California legalized]
(ii) = ¬ [California is the largest state] (if there is a continuation such as *Alaska is the largest state*)

(19) Turkish

A. Kaliforniya, **Amerika'nin** **en** **büyük**
California **America's** **most** **big**
eyaleti, otu yasallaştır-dı
state weed legalize-PST
'C., A.'s largest state, legalized marijuana.'

B. Hayır. Bu doğru değil.
no this true NEG
'No, that's not true'.

(i) = ¬ [California legalized]
(ii) ≠ ¬ [California is the largest state] (even with a continuation such as *Alaska is the largest state*)

B'. Yanıl-ıyor-sun.
be.mistaken.PROG-2SG
'You're mistaken'.

(i) = ¬ [California legalized]
(ii) = ¬ [California is the largest state] (if there is a continuation such as *Alaska is the largest state*)

In (18) and (19), the appositives are sentence-medial to compensate for potential positional effects. Given that this position does not facilitate denials, unlike the sentence-final position in English (which would be especially interesting to test in languages with other word-order patterns, such as Turkish), the contrast between the two strategies is even more marked.

SC In both Bulgarian and Turkish, first-person pain (20, 21) and attitude (Appendix A) reports ban all kinds of disagreement, while third-person statements can be disagreed with using both strategies in question (see section 2.2 on English).

PAIN REPORTS

First person

(20) Bulgarian

A. Glava-ta **me** boli strašno
head-DEF **I**.DAT ache.PRES awfully
'I have an awful headache'.

B. #No, that's not true.

B'. #You are mistaken.

(21) Turkish

A. Can-**ım** yan-ıyor
life-**1SG**.POSS burn-PROG
'I am in pain; lit. My life is burning'.

B. #No, that's not true.

B'. #You are mistaken.

Third person

(22) Bulgarian

A. **Lora** **ja** boli glava-ta strašno
Laura **she**.DAT ache head-DEF awfully
'Laura has an awful headache'.

B. ✓No, that's not true.

B'. ✓You are mistaken.

(23) Turkish

A. **Canın** can-ı yan-ıyor
John's life-**3SG**.POSS burn-PROG
'John is in pain; lit. John's life is burning'.

B. ✓No, that's not true.

B'. ✓You are mistaken.

ER In Bulgarian and Turkish, evidentiality is morphologically part of the tense system. Indirect evidential morphemes *-l* (Bulgarian; Izvorski 1997) and *-mIş* (Turkish; Şener 2011) denote, depending on the context, either inference from results or hearsay. The ER contributed by each morpheme cannot be challenged using either of the strategies in question (24, 25):

Context 1, hearsay: I read a note in LA Times.
Context 2, inference: I come to Venice Beach. Lots of people are smoking weed.

(24) Bulgarian

A. Kalifornija legalizira-**l**-a
California legalize-**IND**.PST-F
marixuana-ta
marijuana-DEF
'C. legalized marijuana, *I hear/infer*'.

B. That's not true.

(i) = ¬ [California legalized]

---

[13] I avoid using 'Hey, wait a minute' for methodological reasons: the cross-linguistic validity of this test as a NAI diagnostic is yet to be established. In particular, the test may not have a direct equivalent in other languages and the results are controversial even for e.g. presuppositions.

[14] As translations show, English behaves the same way.

(ii) ≠ ¬ [You hear/infer it]
B'. You are mistaken.
(i) = ¬ [California legalized]
(ii) ≠ ¬ [You hear/infer it]

(25) Turkish
A. Kaliforniya otu yasallaştır-**mış**
California weed legalize-**IND**.PST
'C. legalized marijuana, *I hear/infer*.'
B. That's not true.
(i) = ¬ [California legalized]
(ii) ≠ ¬ [You hear/infer it]
B'. You are mistaken.
(i) = ¬ [California legalized]
(ii) ≠ ¬ [You hear/infer it]

Below is a detailed summary of applicability of the two disagreement strategies, *No, that's not true* and *You are mistaken* across different kinds of expressions in Bulgarian and Turkish:

	Not true	*Mistaken*
too (30, 31)	☹	✓(w/ follow-up)
appositive (18, 19)	☹	✓(w/ follow-up)
1-person pain (20, 21)	☹	☹
3-person pain (22, 23)	✓	✓
1-person hope (32, 33)	☹	☹
3-person hope (34, 35)	✓	✓
ER (24, 25)	☹	☹

Table 2: Licensing disagreement, itemized

That *No, that's not true* and *You are mistaken* can target the scope proposition in (24) and (25) is predicted both by the NAI and the subjective view. What is surprising under the ER-as-NAI view is that *You are mistaken* can be directed at appositives and presuppositions but not at the ER, even with an explicit follow-up such as 'You didn't hear it', 'Nobody told you so', 'You don't infer it', 'You don't have evidence for it' and so on. The NAI view on evidentiality—especially approaches that model appositives and the ER in the same fashion (Murray, 2014)—fails to predict and explain the pattern. If all types of NAI content were created equal, the difference would be a mystery. Using Jasinskaja (2016)'s insight, one may argue that (a) *You are mistaken* requires a particular level of salience, and that (b) only the content of appositives, but not the ER, satisfies it. While analytically an option, this argument currently has no empirical basis. Furthermore, the NAI approaches predict that all kinds of propositional anaphora would be banned with evidentials, but the prediction is not borne out:

(26) Bulgarian
A. Ana se ozheni-**l**-a.
Ana REFL marry-**IND**.PST-F
'Ana got married, *I hear/infer*'.
B. (Tova e) stranno. Tja mi kaza
(that be.3SG) weird she me say.PST
da go pazja v tajna.
to it keep in secret
'That's surprising. She told me to keep it as a secret.'

The bottom line The lesson learned from the data presented in this section is as follows. Denials make it possible to draw a line between NAI on the one hand and evidentials on the other. If the ER were a type of NAI content, at least *some* kinds of disagreement about content would be possible. This expectation is not borne out. The ER behaves in the same way as subjective expressions such as *I hope* in that disagreement is generally infelicitous, which makes a subjective analysis not just possible but empirically advantageous:

	NAI	SC	ER
That's not true	☹	☹	☹
You're mistaken	✓	☹	☹

Table 3: Licensing disagreement, revisited

4 Performance disagreement

The paper argues that disagreement with SC is infelicitous because self-knowledge described by subjective expressions is not available to the addressee (a non-linguistic fact), so they have no reasonable basis to contest it (a linguistic fact). Such disagreement would signal that the addressee assumes being in a better position to evaluate the speaker's mental state than the actual speaker is. Under normal circumstances, such behavior is outright weird and possibly violates social norms. However, even though the weirdness is rooted in the lexical semantics of the items in question, which in turn is rooted in the qualities of experiences described, the ban is of a pragmatic nature. If so, under less-normal circumstances *some* kind of disagreement should be possible. The prediction is borne out.

It is possible to disagree with SC if the addressee thinks that the speaker is insincere or is impaired in judgment. Consider (27) below:

(27) A. It hurts so much!
B. No, it doesn't.

(27) is common in caretaker-child interactions. B may think that A is faking. Or B may deem A's reaction inappropriate as nothing really serious has

happened. Either way, B is in disagreement with A. But the disagreement is not about the content of A's utterance: after all, B has no access as to what A truly experiences.[15] B is challenging the premises for said utterance.

I will call cases such as (27) *performance disagreement*: the situation when the addressee challenges not the content, but the speaker's performance and thus the grounds for an assertion (the term from (Anand, 2009) on taste ascriptions).

As section 3 shows, genuine disagreement is impossible with first-person statements about pain, first-person attitude ascriptions and evidentials. But performance disagreement is allowed.

In the case of pain, both Bulgarian and Turkish allow dialogues like (27) in scenarios with children and caretakers. This use is highly restricted though, likely due to societal norms. It is infelicitous to challenge an adult's statement about their pain even if you think they are under anesthesia and should not feel anything.

Performance disagreement with attitudes (28) and evidentials (29) is exemplified below.

FIRST-PERSON HOPE

(28) Bulgarian

Context: A is a devout Democrat.

A. Nadjava-**m** se [če Tramp šte spečeli].
hope-**1SG** REFL [that Trump will win]
'I hope that Trump will win.'

B. Ne, kazvaš go samo za provokacija
no say.2SG it only for provocation
'No, you say this only for provocation.'

In (28), B is challenging A's sincerity (or sanity).

EVIDENTIALS

(29) Bulgarian

A. Teksas legalizira-**l** marixuana-ta.
Texas legalize-**IND**.PST marijuana-DEF
'T. legalized marijuana, *I hear/infer*'.

B. Njamaš nikakvo osnovanie za tova.
have.NEG.2SG no ground for that
Prosto si pijan.
just be.2SG drunk
'You have no grounds for saying that. You're just drunk.'

In (29), B is challenging A's competence, suspecting they are drunk. Dialogues similar to (29) and (28) are also possible in cases of assumed hallucinations and other types of impaired performance, or if the addressee thinks that the speaker is lying.

Summing up, evidentials pattern with subjective expressions even with respect to substandard disagreement. This new data point is not immediately handled in current approaches to evidentiality.

5 General discussion

The non-challengeability of the ER has been one of the keystones of NAI approaches to evidentials. Based on the behavior of different types of content with respect to different types of denial, I argue that the ER patterns with subjective expressions and not with NAI.

The main empirical contributions are twofold. (A) Subjective content resists denial. Direct denial thus cannot be used as a two-way diagnostics that separates AI (=denial possible) from NAI content (=denial impossible). (B) In the case of SC, all kinds of denial render the infelicity of response, except for performance disagreement. Evidentials exhibit this very pattern—at least the morphological ones, in contrast with lexical means such as English *allegedly* and *reportedly*. In the case of NAI, denial is contingent on the strategy used: *You are mistaken* is allowed and *That's not true* is banned. I leave investigating the source of flexibility of *You are mistaken*, as well as the behavior of other disagreement techniques, for future research.

The main theoretical claim is that the strong non-challengeability of the ER necessitates a subjective analysis of evidentiality. Certain experiences, such as mental states, are inherently first-person and thus incorrigible, i.e. immune to third-party assessment. In dialogues, these properties give rise to non-challengeability. Evidentials make reference to mental processes such as perception and reasoning, therefore it is only natural to treat them as subjective. And once subjectivity is in place, the NAI analysis is no longer needed.

Acknowledgments

I thank Pranav Anand, Adam Bjorndahl, B.R. George, Adam Maruszak, Kathleen Chase O'Flynn, Yael Sharvit, James Shaw, Mandy Simons, Dominique Sportiche, Tom Werner, and Igor Yanovich, three anonymous SemDial reviewers, and also my language consultants, especially Roumi Pancheva for help with Bulgarian and Deniz Özyildiz for help with Turkish. All errors are mine.

---

[15] I am not concerned here with brain-in-a-vat kind of scenarios where a third party might gain access to one's experiences. I focus not on the logical (im)possibility to assess someone's exclusive states, which is a question for philosophy of mind, but on particularities of dialogues that feature subjective expressions in worlds similar to ours.

References

Alexandra Aikhenvald. 2004. *Evidentiality*. OUP.

Patricia Amaral, Craige Roberts, and E. Allyn Smith. 2007. Review of *The Logic of Conventional Implicatures* by Chris Potts. *Linguistics and Philosophy* 30(6):707–749.

Pranav Anand. 2007. Re-expressing judgment. *Theoretical Linguistics* 33(2):199–208.

Pranav Anand. 2009. Kinds of taste. Ms. UCSC.

Scott Anderbois, Adrian Brasoveanu, and Robert Henderson. 2015. At-issue proposals and appositive impositions in discourse. *Journal of Semantics* 32:93–138.

Murat Aydede. 2013. Pain. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2013 edition.

Wallace Chafe and Johanna Nichols, editors. 1986. *Evidentiality: the linguistic coding of epistemology*. Ablex Publishing Corporation, Norwood.

Martina Faller. 2002. *Semantics and pragmatics of evidentials in Cuzco Quechua*. Ph.D. thesis, Stanford.

Martina Faller. 2007. Evidentiality above and below speech acts. Ms. University of Manchester.

Kai von Fintel. 2004. Would you believe it? The King of France is back! presuppositions and truth-value intuitions. In M. Reimer and A. Bezuidenhout, editors, *Descriptions and Beyond*, OUP, Oxford.

Kai von Fintel and Anthony S. Gillies. 2008. CIA leaks. *Philosophical review* 117(1):77–98.

Kai von Fintel and Anthony S. Gillies. 2011. 'Might' made right. In Andy Egan and Brian Weatherson, editors, *Epistemic modality*, OUP, pages 108–130.

Simeon Floyd, Elisabeth Norcliffe, and Lila San Roque, editors. Forthcoming. *Egophoricity*. John Benjamins, Amsterdam.

Edward John Garrett. 2001. *Evidentiality and Assertion in Tibetan*. Ph.D. thesis, UCLA.

Daniel Gutzmann. 2015. *Use-conditional meaning: studies in multi-dimensional semantics*. OUP, Oxford.

Ferdinand de Haan. 1997. *The Interaction of Modality and Negation: A Typological Study*. Outstanding dissertations in linguistics. Garland.

Ferdinand de Haan. 2013a. Coding of evidentiality. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ferdinand de Haan. 2013b. Semantic distinctions of evidentiality. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Roumyana Izvorski. 1997. The present perfect as an epistemic modal. In *Proceedings of SALT XII*. pages 222–239.

Katja Jasinskaja. 2016. Not at issue any more. Ms. University of Cologne.

Gregory Kierstead. 2015. *Projectivity and the Tagalog Reportative Evidential*. Master's thesis, OSU.

Todor Koev. 2016. Evidentiality, learning events and spatiotemporal distance: The view from Bulgarian. *Journal of Semantics* .

Max Kölbel. 2003. Faultless diasgreement. *Proceedings of the Aristotelian society* 104:53–73.

Natalia Korotkova. 2012. Evidentiality in the Georgian tense and aspect system. Ms. UCLA.

Natalia Korotkova. 2015. Evidentials in attitudes: do's and dont's. In Eva Csipak and Hedde Zeijlstra, editors, *Proceedings of Sinn und Bedeutung (SuB) 19*. pages 340–357.

Natalia Korotkova. In prep. *Heterogeneity and universality in the evidential domain*. Ph.D. thesis, UCLA.

Jungmee Lee. 2013. Temporal constraints on the meaning of evidentiality. *Natural Language Semantics* 21:1–41.

Dong Sik Lim. 2010. *Evidentials as interrogatives: a case study from Korean*. Ph.D. thesis, USC.

John MacFarlane. 2014. *Assessment sensitivity: relative truth and its applications*. Oxford University Press.

Lisa Matthewson, Henry Davis, and H. Rullman. 2007. Evidentials as epistemic modals: Evidence from St'át'imcets. In Jeron van Craenenbroeck, editor, *Linguistic Variation Yearbook*, John Benjamins, volume 7, pages 201–254.

Eric McCready. 2008. What man does. *Linguistics and Philosophy* 31(6):671–724.

Eric McCready. 2010. Varieties of conventional implicature. *Semantics and Pragmatics* 3:1–57.

Eric McCready and Nicholas Asher. 2006. Modal subordination in Japanese: Dynamics and evidentiality. *Penn Working Papers in Linguistics* 12(1):237–249.

Friederike Moltmann. 2012. Two kinds of first-person-oriented content. *Synthese* 184(2):157–177.

Sarah Murray. 2010. *Evidentiality and the Structure of Speech Acts*. Ph.D. thesis, Rutgers.

Sarah Murray. 2014. Varieties of update. *Semantics and Pragmatics* 7:1–53.

Hazel Pearson. 2013. A judge-free semantics for predicates of personal taste. *Journal of Semantics* 30(1):103–154.

Christopher Potts. 2002. The syntax and semantics of *As*-parentheticals. *Natural Language and Linguistic Theory* 20(3):623–689.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. OUP, Oxford.

Craige Roberts. 1998/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6):1–69.

Philippe Schlenker. 2013. Supplements without bidimensionalism. Ms. NYU / Institut Jean Nicod.

Nilufer Şener. 2011. *Semantics and Pragmatics of Evidentials in Turkish*. Ph.D. thesis, UConn, Storrs.

Yael Sharvit. 2015. The onliest NP: Non-definite definites. In *Proceedings of* WCCFL 32. Cascadilla Proceedings Project, Somerville, MA, pages 171–190.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117(6):1034–1056.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Proceedings of* SALT XX. pages 309–327.

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

Kristen Syrett and Todor Koev. 2015. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics* 32(3):525–577.

Judith Tonhauser. 2012. Diagnosing (not-)at-issue content. In *Proceedings of* SULA 6. GLSA, Umass, Amherst, pages 239–254.

Judith Tonhauser. Forth. Reportative evidentiality in Paraguayan Guaraní. In *Proceedings of* SULA 7.

Judith Tonhauser, David Beaver, Craige Roberts, and Mandy Simons. 2013. Towards a taxonomy of projective content. *Language* 89(1):66–109.

Frédérique de Vignemont. 2015. Bodily awareness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2015 edition.

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

Thomas Willett. 1988. A cross-linguistic survey of the grammaticization of evidentiality. *Studies in Language* 12(1):51–97.

Appendix A: Presuppositions and attitudes

PRESUPPOSITIONS: TOO

(30) Bulgarian

A. Kalifornija **sǎshto** legalizira marixuana-ta
California **too** legalize.PST marijuana-DEF
'California, too, legalized marijuana'.

B. No, that's not true.
(i) $= \neg$ [California legalized]
(ii) $\neq \neg$ [Some other state legalized]

B'. You are mistaken.
(i) $= \neg$ [California legalized]
(ii) $= \neg$ [Some other state legalized]

(31) Turkish

A. Kaliforniya **da** otu yasallaştır-dı
California **too** weed legalize-PST
'California, too, legalized marijuana.'

B. No, that's not true.
(i) $= \neg$ [California legalized]
(ii) $\neq \neg$ [Some other state legalized]

B'. You are mistaken.
(i) $= \neg$ [California legalized]
(ii) $= \neg$ [Some other state legalized]

ATTITUDE ASCRIPTIONS: HOPE

Since it is pragmatically odd to evaluate the truth of one's aspirations, only reactions that target the entire sentence are included.

First person

(32) Bulgarian

A. Nadjava-**m** se [če Tramp šte spečeli].
hope-**1SG** REFL [that Trump will win]
'I hope that Trump will win.'

B. #No, that's not true.

B'. #You are mistaken.

(33) Turkish

A. [Trump'ın kazancağını] um-uyor-**um**
[Trump's winning] hope-PROG-**1SG**
'I hope that Trump will win'.

B. #No, that's not true.

B'. #You are mistaken.

Third person

(34) Bulgarian

A. **Republikanci-te** se nadjava-**t** [če Tramp šte spečeli].
Republican.PL-DEF REFL hope-**3PL** [that Trump will win]
'The Republicans hope that Trump will win.'

B. ✓No, that's not true.

B'. ✓You are mistaken.

(35) Turkish

A. **Can** [Trump'ın kazancağını] um-uyor
John [Trump's winning] hope-PROG
'John hopes that Trump will win'.

B. ✓No, that's not true.

B'. ✓You are mistaken.

Appendix B: Abbreviations

1,2,3 person; ACC accusative; CONJ conjectural; DAT dative; DEF definite; DIR direct; F feminine; IND indirect; NEG negation; PL plural; PROG progressive; POSS possessive; PST past; PRES present; REFL reflexive; REP reportative; SG singular; TOP topic

Prominence Shifts in English and Spanish Parallel Constructions

Jeffrey Klassen
McGill University
1085 Doctor Penfield Avenue
Montreal, QC

Michael Wagner
McGill University

Heather Goad
McGill University

Annie Tremblay
University of Kansas
1541 Lilac Lane
Lawrence, KS

Contact: `jeffrey.klassen@mail.mcgill.ca`

Abstract

Certain information theoretical distinctions that are encoded by prosody in English are encoded by word order in Spanish (Bolinger, 1954), a fact often related to the freer word order in Spanish (Lambrecht, 1994; Büring, 2010). This study reports on a production experiment that compares whether and how the two languages mark focus in cases of parallelism, where a change in word order is not an option in Spanish. Prior studies have claimed that Spanish marks focus prosodically only if the focus involved is 'contrastive' or 'corrective' (Zubizarreta, 1998), whereas English marks all types of focus prosodically. Our production results are compatible with this claim, but we offer another interpretation of the results: That the focus operator involved in prosodic focus marking in Spanish necessarily has to take scope over the entire root clause (speech act), while in English it can take scope over a broader range of constituents.

1 Focus-Driven Shifts in Sentential Prominence

The pattern of prosodic prominence of utterances in which all encoded information is new and no constituent is construed as being contrastive is often viewed as the 'default' prosodic pattern. In English, the typical default prosodic pattern for most sentences involves a pitch accent on the last constituent (Chomsky and Halle, 1968; Cinque, 1993), and in Spanish this is even more likely to be the case (Ladd, 1990; Zubizarreta, 1998; Ladd, 2008; Büring, 2010). Examples of the default stress pattern in each language are illustrated in examples (1) and (2).

(1) A: What kept you up last night?
B: [A woman was SINGING]$_F$

(2) A: ¿Que te mantenía despierto anoche?
B: [Alguna mujer CANTÓ]$_F$

The final stressed syllable of the final sentential constituent is likely to be perceived as the most prominent syllable of the sentence, often referred to as its 'nuclear stress'. Sometimes, the main prominence is placed on a constituent other the one which would be expected to carry it by default. Such 'prominence shifts' encode what information is contextually given and what information is 'focused' or 'contrastive'. An example is given in (3):

(3) A: Who was singing last night in the street?
B: [A WOMAN]$_F$ was singing.

Contextually motivated shifts in sentential prominence are argued by Rooth (1992) to reflect the alternatives to an utterance that are relevant in the current context. In Rooth's theory, every constituent comes with a set of alternatives, its 'focus semantic value,' in addition to its regular denotation. When there is an antecedent for focus marking, prominence falls on those constituents that are substituted in the antecedent, and is shifted away from constituents that are the same. In the present case, the question serves as the antecedent, and the relevant alternatives to B's utterance are all of the form *x was singing*. Hence prominence is shifted away from from the predicate and placed on the subject by leaving the VP unaccented and/or by boosting the prominence on the subject (Breen et al., 2010).

It has long been noted that the marking of focus differs between English and

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 76-85.

Spanish (Bolinger, 1954), and more generally between Germanic and Romance languages (Vallduví, 1993; Lambrecht, 1994). One difference between focus-marking in English and Spanish is that Spanish makes use of certain word orders to mark focus that are not allowed in English. As has been described by many researchers (Bolinger, 1954; Lambrecht, 1994; Zubizarreta, 1998; Lozano, 2006; Büring, 2010; Hualde et al., 2012; Domínguez, 2013, i.a.), given or topical constituents are often placed earlier in an utterance, and new or focused constituents are often placed in the more prominent, sentence-final position. Under this view, a leftward shift in prominence, as was seen in English in (3), is not permitted, or at least not preferred when an alternative syntactic strategy is available:

(4) *Spanish*
A: ¿Quién cantó anoche en la calle?
"Who sang last night in the street?"
B: #[Alguna MUJER]$_F$ cantó
a woman sang
B′: Cantó [alguna MUJER]$_F$
sang a woman
"A woman sang."

Sometimes, however, prominence does shift even in Spanish. One instance are corrective utterances (Zubizarreta, 1998):

(5) A: Algun hombre cantó anoche en la calle.
"A man sang in the street last night."
B: No, [alguna MUJER]$_F$ cantó.
no a woman sang
B′: #No, cantó [alguna MUJER]$_F$.
no sang a woman
"No, a WOMAN sang."

One conclusion often drawn is that Spanish marks focus prosodically only in corrective utterances, while in English, focus is marked prosodically in a greater range of circumstances (Zubizarreta, 1998; Ladd, 2008). In fact, Zubizarreta (1998), López (2009), Büring (2010), and others have claimed that syntactic ways to mark focus trade off with prosodic means of focus marking, a claim that is used to explain the prosodic differences between English and Spanish.

The evidence in the literature for this interaction between focus type and prosodic marking has mostly been based on impressionistic observations. The only experimental study that we know of that directly compared English and Spanish with respect to their prosodic marking of different types of focus is Cruttenden (2006). Cruttenden looked at 10 different dialogues in a range of typologically different languages. Although Cruttenden does not identify the types of focus in each of the dialogues, two of the dialogues arguably do involve a corrective response (dialogue 5 and dialogue 7), of which one example is the following:

(6) a. A: I did all the work.
B: You mean your SISTER did all the work.
b. A: Yo hizo todo el trabajo.
I did all the work
B: Lo que quieres decir es que tu
it that want say is that your
hermana hizo todo el TRABAJO.
sister did all the work

Cruttenden found that in contrast to English, where 7 out of 7 speakers shifted prominence to *sister* in (6), 0 out of the 4 Spanish speakers shifted prominence to *hermana*, all instead re-accenting *trabajo*.[1] For the second corrective dialogue (not shown), again 0 out of 4 of the Spanish speakers showed a prominence shift.

Cruttenden's results therefore seem to contradict commonly held assumptions that Spanish shifts prominence for corrective focus. However, Cruttenden's experiment is based on a few isolated sentences (for example, only 2 dialogues that one could plausibly call 'corrective' per speaker), with few speakers (only four speakers in Spanish). A more detailed comparison of focus marking in English and Spanish is clearly needed, with a larger sampling of participants and more carefully controlled stimuli.

2 Prominence Shifts Under Parallelism

In constructions involving series of parallel linguistic constructions, contrastive intonation is necessary in English (Chomsky, 1971):

(7) John is neither EASY to please, nor EA-

---

[1] There are inconsistencies in Cruttenden's reporting of the results. The table of results (p.319) reports 1 out of 4 Spanish speakers shifting prominence in (6), whereas within the text (p.324) it is reported that "all four Spanish speakers re-accented *trabajo*." Additionally, the table of results reports that 7 out of 7 speakers re-accented *work* in the English dialogue, while it is clear within the text that this must be a mistake, and instead it must be that 7 out of 7 English speakers accented *sister* and de-accented *work*.

GER to please, nor CERTAIN to please, nor INCLINED to please, nor HAPPY to please, ...

Rooth (1992) analyzes this type of prosodic marking of contrast as an anaphoric phenomenon, similar to the use of pronouns. Prominence shifts like those observed in (7) require an appropriate antecedent. He introduces the presuppositional focus operator $\sim$. The operator $\sim$ introduces the presupposition that there is an antecedent similar to the constituent that $\sim$ attaches to (the 'scope' of $\sim$). The antecedent has to be identical, except that any F-marked constituent contained in the scope of $\sim$ (its 'focus', or 'foci' if there are multiple) has to be non-identical at least in one alternative. Under this theory, the focus structure for (7) is analyzed as follows:

(8) John is neither $\sim$[EASY$_F$ to please], nor $\sim$[EAGER$_F$ to please], nor $\sim$[CERTAIN$_F$ to please], nor $\sim$[INCLINED$_F$ to please], nor $\sim$[HAPPY$_F$ to please], ...

Usually, the antecedent for prosodic focus marking precedes the anaphor, as in (3), where the question in the context serves as the antecedent for the answer. But in cases of parallelism as in (8), a prominence shift is possible even in the first occurrence of the parallel structure, in which case prosodic focus marking is *cataphoric* rather than *anaphoric*. Put simply, the first instance of focus marking in (8) leaves the listener hanging: It sets up a contrast that requires an *post*cedent that has not been realized yet. A listener might use this information and expect a parallel structure to be imminent. Another way of thinking about cataphora is that the listener has to accommodate a prior (unmentioned) antecedent, to which all instances of the parallel structure are anaphoric (Williams, 1997). The exact conditions governing cataphoric prominence shifts are not yet known: Authors such as Rooth (1992) have provided the intuition that a shift in prominence within the first parallel constituent (e.g. *easy* in (8)) is optional English, which would make sense because it requires a level of foresight when planning the utterance that may not always be possible. Cataphoric focus marking requires a greater amount of look-ahead, and therefore its optionality might be due to limits on production planning.

The intuition sometimes reported for Romance languages is that there is no shift in prominence in cases of parallelism (Ladd, 1990; Ladd, 2008; Bocci, 2013). Vander Klok et al. (2014) provide experimental evidence that this is correct at least for French—note that French, unlike Spanish, does not employ the focus-driven changes in word order seen in (4).

The case of parallelism is particularly interesting because it provides an opportunity to constrain the scope of $\sim$ in order to observe how its scope affects prosodic focus marking. We use this method in the present study to test whether Spanish marks corrective focus differently from other types of focus. Consider the three cases of parallelism in English below:

(9) a. Move $\sim$[angel$_F$ number two] to $\sim$[donkey$_F$ number two].
  b. $\sim$[Click angel$_F$ number two]. Then $\sim$[click donkey$_F$ number two].
  c. $\sim$[Don't click angel$_F$ number two.] $\sim$[Click donkey$_F$ number two].

One characterization of the relevant differences between the three cases is that only the last example in (9-c) is 'corrective,' while the other two cases are merely 'contrastive'. Uses of focus can more generally be distinguished by their pragmatic function, and focus is often classified into different 'types of focus' along those lines. However, there is another way to characterize the distinctions in (9): They differ in the scope of $\sim$; that is, they differ in the the attachment height of $\sim$ (Vander Klok et al., 2014). Since each of the two parallel constituents has to serve as an antecedent for the focus marking of the other, the scope of $\sim$ is constrained and cannot be so wide that both parallel constituents fall within the constituent that $\sim$ attaches to. This means that in (9-a), the scope of $\sim$ cannot be wider than the NPs, and each focus operator has to attach to a separate individual-denoting NP within a single clause. In (9-b) and (9-c), $\sim$ cannot span both sentences, but $\sim$ can attach to nodes bigger than the individual NPs, that denote entire propositions. In the last example in (9-c), $\sim$ can attach to constituents that correspond to separate imperative speech acts.

If Spanish marks only *corrective* focus prosodically, we would expect focus marking only to be possible in the Spanish equivalent of (9-c). In this case, we could also characterize Spanish as restricting the scope of $\sim$ to constituents that denote entire speech acts (in an alternative terminology, we could refer to such constituents as "root clauses," where the root corresponds to a syntactic node used in a particular speech act). We refer to this as the "Corrective" or "Speech act scope" hypothesis:

(10) Spanish: Corrective Hypothesis
(or: Speech act scope) – see (9) for translation
  a. Ponga el ángel número dos en el burro número dos.
  b. Haga clic en el ángel número dos. Después haga clic en el burro número dos.
  c. No $\sim$[haga clic en el ángel$_F$ número dos.] $\sim$[Haga clic en el burro$_F$ número dos.]

Another possibility, however, is that Spanish also allows for a prominence shift in cases like (10-b), where $\sim$ can scope over constituents that denote propositions. Under this view, Spanish restricts $\sim$ to clausal scope. This is compatible with the claim that certain Romance languages do not allow multiple focus operators within a single clause (Calabrese, 1987; Stoyanova, 2008; Bocci, 2013). Then, English and Spanish should pattern similarly with respect to (10-b) and (10-c), shifting prominence where the two focus operators involved are in separate clauses. In (10-a), where two focus operators within a single clause are necessary in order to mark parallelism, prosodic focus marking should still be impossible in Spanish. This so-called "Propositional Scope" hypothesis therefore generates different predictions from the "Speech Act Scope" or "Corrective" hypothesis, the predictions of which which are illustrated in (10).

By comparing the realization of the three sentence types in (9) and (10), we can gain novel insights into the grammatical underpinnings of the differences and provide the first systematic experimental evidence of these claimed differences. At the same time, this comparison will serve as a test for the more basic claim, assumed by many previous authors, that Spanish reliably marks corrective focus as in (10-c) but not does not mark parallelism in cases like (10-b).

3 Materials

We designed our materials to test focus at three different levels with respect to the syntactic scope of the focus operator involved: Two parallel DPs within a single clause ("sub-clausal," as in (9-a) and (10-a)), two parallel clauses ("clausal," as in (9-b) and (10-b)), and two parallel speech acts, where the second corrects the first ("super-clausal," as in (9-c) and (10-c)). Within each type of parallelism, we controlled whether the first, the second, or both constituents were contrasted. For example, when the first constituent was contrasted, the head nouns would be in focus (e.g. Move [angel]$_F$ number two to [donkey]$_F$ number two). When the second was contrasted, the number modifiers would be in focus (e.g. Move angel [number two]$_F$ to angel [number three]$_F$). Finally, when both contrasted, the phrase contained two foci, on both the head noun and the modifier (e.g Move [angel]$_F$ [number two]$_F$ to [donkey]$_F$ [number three]$_F$). This resulted in a total of 9 conditions. An example set of 9 stimuli is summarized in Table 1 for English and in Table 2 for Spanish. We created 72 object-number combinations for each language (each with 2 objects $*$ 2 numbers), each with 9 variants (according to the 9 conditions; i.e. 72 $*$ 9 items per language).

The experiment was run in a Latin square design, where each participant only saw one condition from each object-number combination, but an equal number of 8 trials from each condition across the experiment. The items were presented in random order. The objects were chosen to be relatively high frequency nouns referring to concrete, easily illustratable objects like animals, articles of clothing and food items. They consisted of only disyllabic trochees in both Spanish and English. The numbers two, three and six (*dos*, *tres* and *seis*) were used because they are monosyllabic in both languages. Number modifiers ("number two", "number three") were used because they are postnominal in both English and Spanish.[2]

---

[2] Many previous studies of prosodic focus make use of noun-adjective combinations (Swerts et al., 2002; Hamlaoui et al., 2012), which somewhat limits crosslinguistic comparisons between Germanic and Romance languages: Germanic adjectives usually precede the noun while in Romance, the opposite is true. Using a modifier that is postnominal in both languages allowed us to control for potential effects of the syntax on prominence. Vander Klok et al. (2014), however, report experimental evidence suggesting that whether an adjectival modifier is prenominal or postnominal does not alter whether a prominence shift with an NP is possible or not.

	Sub-clausal Scope
Head Noun	Move angel number two to donkey number two.
Modifier	Move angel number two to angel number three.
Both	Move angel number two to donkey number three.
	Clausal Scope
Head Noun	Click angel number two. Then click donkey number two.
Modifier	Click angel number two. Then click angel number three.
Both	Click angel number two. Then click donkey number three.
	Super-clausal Scope
Head Noun	Don't click angel number two. Click donkey number two.
Modifier	Don't click angel number two. Click angel number three.
Both	Don't click angel number two. Click donkey number three.

Table 1: English production task conditions

	Sub-clausal Scope
Head Noun	Ponga el ángel número dos en el burro número dos.
Modifier	Ponga el ángel número dos en el ángel número tres.
Both	Ponga el ángel número dos en el burro número tres.
	Clausal Scope
Head Noun	Haga clic en el ángel número dos. Después haga clic en el burro número dos.
Modifier	Haga clic en el ángel número dos. Después haga clic en el ángel número tres.
Both	Haga clic en el ángel número dos. Después haga clic en el burro número tres.
	Super-clausal Scope
Head Noun	No haga clic en el ángel número dos. Haga clic en el burro número dos.
Modifier	No haga clic en el ángel número dos. Haga clic en el ángel número tres.
Both	No haga clic en el ángel número dos. Haga clic en el burro número tres.

Table 2: Spanish production task conditions

4 Research Questions and Predictions

Our first research question was whether the two languages mark focus prosodically in the final constituent (which we will call "NP2"), and if this would occur in all types of contexts. The Speech Act Scope hypothesis (and Corrective hypothesis) predict that prominence shifts in Spanish should occur only within the Super-clausal condition, while a Propositional Scope approach would predict prominence in Spanish to be shifted in both the Clausal and Super-clausal conditions. A second research question relates to the marking of cataphoric (anticipatory) focus in English and Spanish (within "NP1"), in order to see whether speakers of English or Spanish would ever shift prominence in anticipation of the upcoming constituent. We predicted prominence shifts to be less frequent in NP1 than in NP2 because of its optional nature (possibly due to limits in look-ahead when planning an utterance).

5 Procedures

Participants were recorded with the use of a digital head-mounted microphone. The participant sat in front of one computer screen and the experimenter sat at a second screen that was turned away from the participant at a perpendicular angle. The participant was required to instruct the experimenter to move or click images on the screen based on different symbols that appeared with the images. The experimenter performed the instructions on their own screen. Before the experimental trials, the participants practiced each type of instruction by running through a block of 9 practice trials. The practice block was repeated as needed to ensure that the participant gave the correct instruction corresponding to the symbol on the screen. The three types of instructions were "Move," which was indicated with an arrow (Figure 1), "Click...then, click," which was indicated with two green squares and an arrow (Figure 2), and "Don't click...click," which was indicated with a red square and a green square (Figure 3).

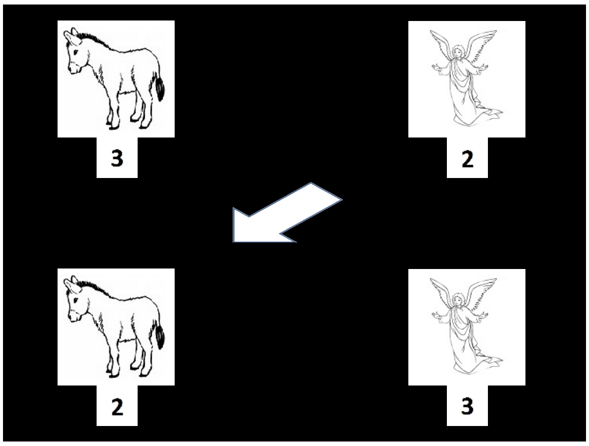


Figure 1: Visual Array – "Move angel number two to donkey number two."

The participant was told that the experimenter could not see the instructive symbols so that they would think that the verbal instructions were the only information available to the experimenter. In reality, both screens were completely synchronized, and the experimenter performed the moves

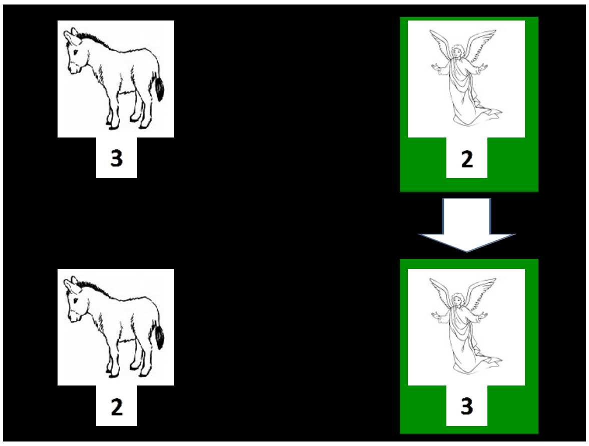


Figure 2: Visual Array – "Click angel number two. Then, click angel number three."

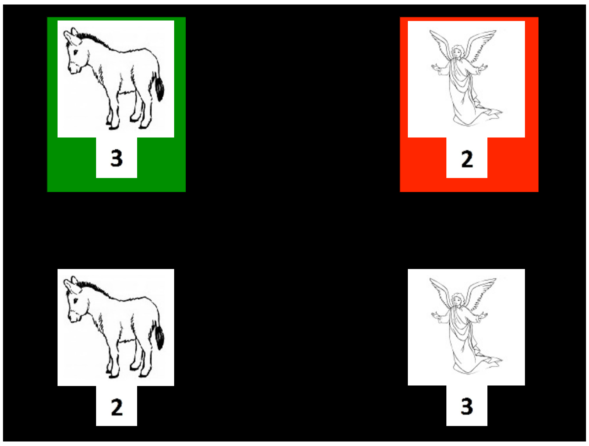


Figure 3: Visual Array – "Don't click angel number two. Click donkey number three."

as instructed, simply pretending to not know each move. During the experiment, each visual array was presented for 4 seconds before the appearance of the symbols in order for the participant to have enough time to activate the name of the objects on the screen. Once the symbols appeared, the recording began: The participant formed their utterance based on the symbols, and the experimenter gave verbal confirmation to continue once they had carried out the move.

6 Participants

Two groups of participants were recruited: a group of 16 North American English native speakers (10 female, born in USA and Canada[3]) and a group of 17 native speakers of Spanish (14 female) from Latin American countries.[4] Of the 17 Spanish speakers, 9 were born in Colombia, 3 were born in Mexico, 2 were born in Venezuela, and 1 each was born in Chile, Cuba, and the Dominican Republic.

Because they were concurrently participating in a second language study, all participants were Spanish-English bilingual to a limited extent: Both native speaker groups scored at an intermediate level of proficiency in their second language. In addition, participants were not excluded if they had knowledge of a third language (most commonly, French), but such participants were included in the study only if their third language was reported to be less dominant and less proficient than both Spanish and English.[5]

7 Data Analysis

Data were coded for prominence impressionistically by two trained annotators. A research assistant whose native language was English coded the English data and the first author, who is also an English native speaker, coded the Spanish data. For each recorded item, the annotator listened and noted whether the main stress of the phrase had been shifted leftward to the head noun (prominence shift) or if it remained in the default rightmost position (no prominence shift). Acoustic measures were extracted and the prominence annotations were validated by measuring the correlation between the annotations and the acoustic measures by means of a logistic regression: Items marked as "prominence shift" consistently showed a larger difference in prominence between the head noun and the modifier (relative prominence). In NP1, pitch and duration predicted prominence, and in NP2, intensity and duration were the significant predictors. Both languages were pooled together in the regression, and language did not lead to a significant interaction when included in the model (and therefore was excluded). The results

---

[3] One native English speaker was born in the United Kingdom but moved to the US at a young age and spoke with a North American accent.

[4] Argentinean Spanish speakers were excluded since the dialect is known to differ greatly from other dialects of the Americas, particularly with respect to information structural components (Gabriel, 2010).

[5] Language status was determined by means of a language background questionnaire that asked participants to report their proficiency in each language they knew, and asked about the amount of exposure they received from each language throughout all stages of infancy to late adolescence. All participants lived in an English-speaking country at the time of the study. All Spanish native speakers had arrived in the English-speaking country during adulthood (mean age of arrival: 26.18 years).

of these models are shown in Table 3.

It must be acknowledged that using impressionistic annotations from a single rater in each language is less than ideal, as is using non-native speakers to annotate their non-native language. It is therefore clear that additional analyses (a second annotator in English in order to establish inter-rater reliability, use of two Spanish native speakers for Spanish) are indeed necessary. In addition to this, the annotators could not be completely blind to the initial conditions because the sentence types were apparent upon listening. In future analyses, we will employ annotators who are blind to the experimental hypotheses.

	NP1	**NP2**
	Coeff (SE)	Coeff (SE)
(Intercept)	$-3.65\ (0.24)^{***}$	$-1.46\ (0.18)^{***}$
Intensity	$0.14\ (0.08)$	$0.15\ (0.05)^{**}$
Pitch	$0.34\ (0.09)^{***}$	$-0.04\ (0.05)$
Duration	$0.55\ (0.11)^{***}$	$0.39\ (0.06)^{***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Acoustic Predictors of Annotations

Despite the issues arising from employing impressionistic prominence annotations, we do not directly report on the raw acoustic measures because it is often the case that with acoustic prominence, several acoustic variables work together in a dependent fashion. For example, Breen et al. (2010) showed that relative intensity, duration and pitch worked additively to determine whether prominence had been perceived to have been shifted: One, two or all three cues may be present, but it is difficult to predict the exact mixtures required. The percept of relative prominence has been argued to be more robust than acoustic measures, and in general leads to high inter-annotator reliability (Klassen and Wagner, 2016).

Finally, given the results of the logistic regression in Table 3, it could very well be the case that cataphoric focus is marked using different combinations of cues in relation to those used in focus marking for the second constituent, as argued for example in (Rooth, 2015); this would need to be investigated in further research.

8 Results

8.1 Anaphoric Focus

We first look at the prosodic realization of the second constituent, that is, the case of anaphoric (as opposed to cataphoric) focus marking. As seen in Figure 4, English speakers shifted stress to the head noun of NP2 in the Head Noun condition in 92-93% of the trials, whereas they rarely shifted prominence in cases where the modifier was not given (Both and Modifier conditions). Scope was not a significant factor in determining prominence shift in English.

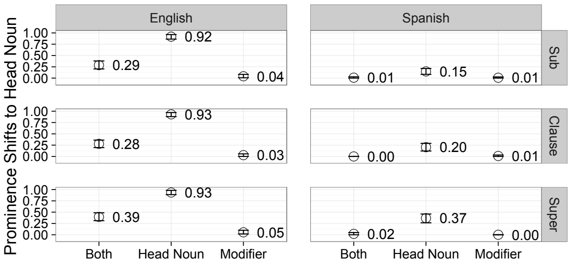


Figure 4: Prominence annotations for NP2 (Anaphoric focus)

	Coeff (SE)
(Intercept)	$-1.46\ (0.11)^{***}$
scope1 (Super vs. other)	$0.52\ (0.15)^{***}$
scope2 (Clause vs. Sub)	$-0.23\ (0.21)$
language	$2.26\ (0.22)^{***}$
scope1:language	$-0.61\ (0.29)^{*}$
scope2:language	$0.34\ (0.42)$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Generalized linear mixed model - Anaphoric Focus. Formula: Annotation $\sim$ scope $*$ language $+$ (1$|$participant) $+$ (0$+$scope$|$participant) + (0$+$scope$|$item)

Overall, Spanish speakers shifted prominence less frequently than what was seen for English speakers. As seen in Figure 4, Spanish speakers shifted prominence to the head noun within the Head Noun condition in only 20-37% of the trials, depending on the level of scope of the focus operator.

We tested for the significance of the observed differences using a logistic mixed-effects regression, outlined in Table 4. Our model included Scope and Language and their interaction as fixed effects, and random effects for by-item and by-participant differences. The random effects included slopes for the two fixed effects and their interaction. The three-level factor Scope was coded using Helmert Coding: The first contrast compared Super-scope vs. the two other scopes and the second contrast compared clausal scope vs.

sub-clausal scope. Helmert Coding was the type of coding best suited to our theoretical question: There was no true control condition with respect to Scope; the first comparison (Super vs. other) directly tested the Corrective/Speech Act Scope hypothesis while the second (Clausal versus Sub) tested the Propositional Scope hypothesis.

The results show a main effect of Language: A prominence shift was generally more likely in English than in Spanish. It also showed a main effect of Scope: Super-clausal scope (i.e., in our stimuli, corrective focus) was more likely to cause a prominence shift than other types of scope. Crucially, there was also a significant interaction between Language and Scope, showing that the difference between super-clausal and other types of scope observed in Spanish differed significantly from that difference in English. To our knowledge, this is the first experimental demonstration that indeed, the difference between corrective focus and other types of focus is important in determining prominence shifts in Spanish, but not in English.

8.2 Cataphoric Focus

When looking at prosodic focus marking in the first constituent, we see that the rate of prominence shifts in NP1 is much lower compared to that in NP2. However, as seen in Figure 5, English native speakers do, to some degree, shift prominence to the head noun in cataphoric focus contexts, while Spanish native speakers almost never do.

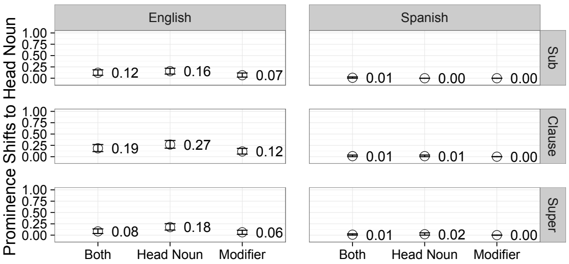


Figure 5: Prominence annotations for NP1 (Cataphoric focus)

The difference in the rate of cataphoric prominence shifts between the English and Spanish experiments is characterized by a significant main effect of test language in the model shown in Table 5. Scope was not a significant predictor.

9 Discussion and Conclusion

With respect to anaphoric focus, Spanish speakers shift prominence leftwards with greater frequency in the Super-clausal scope condition, where the focus operator took wide scope over the speech act, than in the other two conditions with narrower scope. This is compatible with the idea that in Spanish, only corrective focus is marked prosodically. It is also compatible with the idea that in Spanish, $\sim$ necessarily takes scope over constituents that correspond to entire speech acts (the 'root' level). This pattern of results cannot be explained by the Propositional Scope hypothesis—if this was the correct explanation, then the case in which both operators attach to constituents denoting propositions should allow for prominence shifts in Spanish.

	Coeff (SE)
(Intercept)	$-3.89\ (0.31)^{***}$
scope1 (Super vs. other)	$-0.19\ (0.56)$
scope2 (Clause vs. Sub)	$-0.42\ (0.67)$
language	$3.40\ (0.57)^{***}$
scope1:language	$-1.07\ (0.90)$
scope2:language	$-0.01\ (1.29)$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: Generalized linear mixed model - Cataphoric Focus. Formula: Annotation $\sim$ scope $*$ language $+$ (1|participant) $+$ (0$+$scope|participant) + (0$+$scope|item)

We found the rate of cataphoric focus marking to be fairly low in English, and essentially at floor in Spanish. A future study would need to find the cause of the optionality in English. It could be that cataphoric focus marking requires greater look-ahead, although it would be difficult to see how one could increase look-ahead in an experimental setting: Our participants already had plenty of time to plan their utterances. One idea could be to use written instructions instead of symbols, because in such a case the majority of the sentence (besides the prosodic realization) would already be planned out for the speaker. Another method that might encourage a higher degree of cataphoric focus marking in an experimental setting would be to create a situation in which the task is time-sensitive and the speaker is required to increase the response time of the listener: Perhaps in such a case, cataphoric focus would be employed more frequently, in order to help the listener anticipate the final game instruction and respond with greater speed.

An unexpected result is that the rate of prominence shift in Spanish for corrective focus was only 37% — fairly low in comparison to English. We have suggested that the source of the difference between English and Spanish could be explained by the scope of $\sim$. However, this could function in two different ways: Either Spanish allows only very wide (root-level) scope for $\sim$, or only wide-scope $\sim$ shows prosodic effects (see (Vander Klok et al., 2014) for a related proposal for French).

Under the view that $\sim$ can only attach to root-level nodes, one possible way to explain the low rate of Spanish prominence shifts is that our stimuli were in fact ambiguous between two different kinds of structures: The first structure involves a single speech act with two sub-commands, which we separate here with a comma. In this case, the utterance includes what is sometimes called 'contrastive negation' (McCawley, 1991) or 'replacive negation' (Jacobs, 1991). The second possible structure involves two independent speech acts:

(11) a. Don't click angel number two, click donkey number two.
b. Don't click angel number two. Click donkey number two.

If Spanish speakers only shifted prominence in one of these two structures, it would shed further light on the precise conditions governing prosodic focus marking in Spanish. We have not yet tried to test whether there is any evidence in our data for such a distinction (for example, the two structures might differ with respect to the boundary tones separating the two commands).

Another possibility is that $\sim$ can attach at all levels in Spanish, but only has prosodic effects attached at the root level. Our parallelism manipulation sets an upper bound for the scope of $\sim$ (it cannot attach to a node that includes both legs of the parallelism, since then there is no appropriate antecedent anymore). It also sets a lower bound (it cannot attach lower than the node that contains the F-marked constituent). But, as was pointed out to us by a reviewer, it is still compatible with several attachment sites (provided that lower scope is possible at all in Spanish). Our corrective condition is compatible with adjoining the $\sim$ operators at the NP level, since their presupposition is fulfilled for this contrast between smaller constituents as well:

(12) a. Narrow Scope: Don't click $\sim$[angel$_F$ number two], click $\sim$[donkey$_F$ number two].
b. Wide Scope: Don't $\sim$[click angel$_F$ number two], $\sim$[click donkey$_F$ number two].

Note that in English, either wide or narrow attachment of $\sim$ would lead to a prominence shift, but based on this hypothesis, only giving widest scope to $\sim$ would lead to a prominence shift in the second leg of the parallelism in Spanish. Variation in the scope of $\sim$ could therefore explain the lower accentuation rate in Spanish compared to English.

Understanding the relatively low rate of corrective focus marking in Spanish might prove crucial in order to further differentiate the different interpretations of the observed patterns. What our results clearly show is that corrective focus is indeed different from other types of focus in Spanish in terms of its prosodic realization. Furthermore, we maintain that corrective focus can be described in syntactic terms: corrective focus involves root-level scope of $\sim$, as was argued in Vander Klok et al. (2014).

Acknowledgments

Thanks to the members of prosody.lab for help with the alignment and annotation of the data, as well as Kalyna Franko and Kristina van Anne, who were involved in collecting data. Spanish intuitions were provided by José Alemán Bañon, Iris Hauser, and Maite Martínez García. Thanks as well to Lydia White, for comments on earlier drafts. This research was supported by funding from a SSHRC Canada Research Chair, as well as a SSHRC Doctoral Scholarship, and SSHRC grant 435-2014-1504.

References

[Bocci2013] Giuliano Bocci. 2013. *The Syntax-Prosody Interface*. John Benjamins, Amsterdam, NL.

[Bolinger1954] Dwight L. Bolinger. 1954. English prosodic stress and Spanish sentence order. *Hispania*, 37(2):152–156.

[Breen et al.2010] Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes*, 25(7):1044–1098.

[Büring2010] Daniel Büring. 2010. Towards a typology of focus realization. In Malte Zimmermann and

Caroline Féry, editors, *Information Structure: Theoretical, Typological and Experimental Perspectives*, pages 177–205. Oxford University Press, Oxford, UK.

[Calabrese1987] Andrea Calabrese. 1987. Focus structure in Berber: A comparative analysis with Italian. In Mohamed Guerssel and Kenneth L. Hale, editors, *Studies in Berber Syntax*, pages 103–120. MIT Press, Cambridge, MA, USA.

[Chomsky and Halle1968] Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York, NY, USA.

[Chomsky1971] Noam Chomsky. 1971. Deep structure, surface structure, and semantic interpretation. In D.D. Steinberg and L.A. Jakobovits, editors, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*. Cambridge University Press, Cambridge.

[Cinque1993] Gugliemo Cinque. 1993. A null theory of phrase and compound stress. *Linguistic Inquiry*, 24(2):239–297.

[Cruttenden2006] Alan Cruttenden. 2006. The deaccenting of given information: A cognitive universal? In Giuliano Bernini and Maria L. Schwartz, editors, *Pragmatic Organization of Discourse in the Languages of Europe*. Mouton de Gruyter, New York, NY, USA.

[Domínguez2013] Laura Domínguez. 2013. *Understanding Interfaces: Second Language Acquisition and Native Language Attrition of Spanish Subject Realization and Word Order Variation*, volume 55 of *Language Acquisition and Language Disorders*. John Benjamins, Amsterdam, NL.

[Gabriel2010] Cristoph Gabriel. 2010. On focus, prosody, and word order in Argentinean Spanish: A minimalist OT account. *Revista Virtual de Estudos da Linguagem*, 8(4).

[Hamlaoui et al.2012] Fatima Hamlaoui, Sascha Coridun, and Caroline Féry. 2012. Expression prosodique du focus et du donné au sein des groupes nominaux [N A] du français. In Franck Neveu, Valelia Muni Toke, Peter Blumenthal, Thomas Klinger, Pierliugi Ligas, Sophie Prévost, and Sandra Teston-Bonnard, editors, *Actes du 3e Congrès Mondial de Linguistique Française*, volume 1, pages 1505–1518. SHS Web of Conferences.

[Hualde et al.2012] José Ignacio Hualde, Antxon Olarrea, and Erin O'Rourke, editors. 2012. *The Handbook of Hispanic Linguistics*. Blackwell Handbooks in Linguistics. Wiley-Blackwell, Malden, MA, USA.

[Jacobs1991] Joachim Jacobs. 1991. Negation. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantik. Ein internationales Handbuch der zeitgenössischen Forschung. (HSK 6)*, pages 560–596. Mouton de Gruyter, Berlin, Germany.

[Klassen and Wagner2016] Jeffrey Klassen and Michael Wagner. 2016. Prosodic prominence shifts are anaphoric. Accepted for publication.

[Ladd1990] D. Robert Ladd. 1990. Intonation: Emotion vs. grammar. review of: Intonation and its uses, by Dwight Bolinger. *Language*, 66(4):806–816.

[Ladd2008] D. Robert Ladd. 2008. *Intonational Phonology*, volume 79 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge, UK.

[Lambrecht1994] Knud Lambrecht. 1994. *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, UK.

[López2009] Luis López. 2009. *A Derivational Syntax for Information Structure*. Oxford University Press, Oxford, UK.

[Lozano2006] Cristobal Lozano. 2006. Focus and split-intransitivity: the acquisition of word order alternations in non-native Spanish. *Second Language Research*, 22(2):145–187.

[McCawley1991] James D. McCawley. 1991. Contrastive negation and metalingistic negation. In *Proceedings of the 27th Annual Meeting of the Chicago Linguistic Society. Part II: The Parasession on Negation*, pages 189–206.

[Rooth1992] Mats Edward Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

[Rooth2015] Mats Rooth. 2015. Representing focus scoping over new. In Thuy Bai and Deniz Özyildiz, editors, *Proceedings of the Forty-Fifth Annual Meeing of the North East Linguisic Society*.

[Stoyanova2008] Marina Stoyanova. 2008. *Unique Focus. Languages without multiple wh-questions*. John Benjamins.

[Swerts et al.2002] Marc Swerts, Emiel Krahmer, and Cinzia Avesani. 2002. Prosodic marking of information status in Dutch and Italian: a comparative analysis. *Journal of Phonetics*, 30(4):629–654.

[Vallduví1993] Enric Vallduví. 1993. *The Informational Component*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

[Vander Klok et al.2014] Jozina Vander Klok, Heather Goad, and Michael Wagner. 2014. Prosodic focus in English vs. French: A scope account. LingBuzz: ling.auf.net/lingbuzz/002274.

[Williams1997] Edwin Williams. 1997. Blocking and anaphora. *Linguistic Inquiry*, 28(4):577–628.

[Zubizarreta1998] Maria Luisa Zubizarreta. 1998. *Prosody, Focus and Word Order*. MIT Press, Cambridge, MA, USA.

Impotent Speech Acts & Awareness

Phil Crone
Stanford University
Department of Linguistics
`pcrone@stanford.edu`

Abstract

Different types speech acts are associated with different types of effects on the discourse, e.g. assertions are associated with the addition of information to the common ground. I show that across three main clause types – declaratives, interrogatives, and imperatives – we find speech acts that do not have the expected effects on the discourse. At first glance, these speech acts appear to serve no conversational purpose at all. I argue that these "impotent" speech acts can be understood in terms of their ability to raise awareness or draw attention to issues. Although the ability to raise awareness is not a property unique to impotent speech acts, I argue that these speech acts are particularly well-suited for this purpose due to addressees' pragmatic reasoning regarding the speaker's beliefs about the addressee.

1 Introduction

Natural language clause types are typically associated with particular speech acts, which are, in turn, associated with particular effects upon the discourse context. To illustrate, consider (1).

(1) Trump is the Republican nominee.

The sentence in (1) is a declarative clause, and a speaker who utters (1) with falling intonation performs the speech act of an assertion. If (1) is asserted in a particular discourse and is not challenged by any discourse participants, its effect is that of adding the proposition that Trump is the Republican nominee to the common ground.

Although particular clause types are typically associated with particular speech acts, there is no one-to-one mapping between the two. This is perhaps most obvious in the case of imperative clauses, which can be used to perform commands, warnings, requests, well-wishes, invitations, etc. (Kaufmann, 2012). Despite the failure of clause types to associate with unique speech acts, I focus here on the more pedestrian cases in which clauses are used to perform the traditional speech acts associated with them. That is, I focus on declarative clauses that are used to assert, interrogative clauses that are used to ask questions, and imperatives that are used directively. I show that in each case, we find uses that are "impotent" in the sense that the expected effect of the speech act on the context must already be entailed by the context for the speech act to be felicitous. Note that although there is no one-to-one mapping between clause types and speech act types, I do assume that each type of speech act has a unique effect on the discourse.

The question arises why any speaker would ever use an impotent speech act. I argue that the utility of such expressions derives from their ability to draw attention to or raise awareness of particular facts about the context that interlocutors may be ignoring. In turn, these awareness-related effects may play a role in structuring the discourse or in guiding agents' resolution to decision problems. Although "potent" speech acts may also play the role of raising awareness, impotent speech acts are particularly well-suited for this function due to pragmatic reasoning on the part of the addressee(s) about the speaker's beliefs about the addressee(s) and the speaker's intentions.

2 Impotent Speech Acts across Clause Types

2.1 Declaratives

Declarative clauses are traditionally taken to be used to assert. Following Stalnaker (1978), the essential effect of assertions on the discourse can be seen as informing. Informativity is formalized

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 86-96.

in terms of the *common ground* (CG), the set of propositions that are commonly believed by all discourse participants or that all discourse participants act as if they commonly believe. A proposition p is a common belief iff each participant believes p, and each participant believes that every other participants believes p, etc. This contrasts with the weaker notion of mutual belief: p is mutually believed by a set of discourse participants iff each discourse participants believes p. The set of worlds consistent with the CG (i.e. $\cap CG$) is known as the *context set C* (Stalnaker, 1978; Gunlogson, 2001; Stalnaker, 2002). If an assertion with content p is accepted in a discourse, p enters the CG. Let C_1 and C_2 denote the context set before and after, respectively, an assertion a with content p. Then, a is informative iff $C_2 \subset C_1$.

If we view the essential function of assertion as informing, then an impotent assertion is one that is uninformative. In English, uninformative assertions include those of the form *As you already know, p* or *Needless to say, p*:

(2) As you already know/Needless to say, Trump terrifies me.

Example (2) entails that the addressee already knows that Trump terrifies the speaker. Of course, this is consistent with the speaker's attitude towards Trump being a mutual belief, but not a common belief. If this is the case, an assertion of (2) would add new information to the CG and would therefore be informative.

Yet it is possible to assert *As you already know/Needless to say, p* when p is already in the CG. For example, suppose the speaker of (2) had informed the addressee about their feelings towards Trump in a previous conversation. It would still be felicitous for the speaker to utter (2) at the start of a new discourse, perhaps one in which the speaker intends to elaborate on these feelings. Importantly, at the start of this discourse, that Trump terrifies the speaker would already be in the CG.

Barker and Taranto (2003) and Barker (2009) provide another example of necessarily uninformative assertions. First, Barker and Taranto observe that in a context in which all discourse participants have access to evidence supporting p, it is felicitous to assert that p is clear. For example, if we see a photograph of a woman dressed as a doctor, (3) may be felicitously asserted.

(3) It is clear that she is a doctor.

But if the photograph truly makes it clear that the woman is a doctor, then all discourse participants should be able to conclude that she is a doctor. Moreover, all discourse participants can conclude that the other discourse participants have concluded that she is a doctor, etc. Thus, in such contexts in which the evidence supporting a clarity assertion is part of the CG, the clarity assertion itself is necessarily uninformative.

Assertions containing the unfocused variants of the German discourse particles *ja* and *doch* provide another potential case of uninformative assertions. A full discussion of these data go well beyond the scope of this paper. However, the conventional wisdom on these particles as summarized by Kaufmann and Kaufmann (2012) suggests that they behave similarly to the phenomena discussed above: "It is widely agreed that both $ja(p)$ and $doch(p)$ commit the speaker to the belief that p is in some sense given, obvious, or uncontroversial" (210). The English expressions *after all* and *of course* may function similarly:

(4) Trump probably can't win in November. After all, many Republicans don't even like him.

According to my judgments, an assertion of (4) is only felicitous if all discourse participants already believe that most Republicans don't like Trump.[1]

Before moving on, I address a potential objection to the claims that the assertions discussed above are truly impotent. Speakers often assert *As you already know, p*, *Needless to say, p*, etc. when, in fact, the addressee does not already know p. In such cases, these assertions may be genuinely informative. These uses seem to be grounded in a desire to treat the addressee as more knowledgeable than they actually are for politeness reasons, an idea that is explored in §4 below. I maintain that when speakers use these expressions in this way, they assert something that is false. My focus is on truthful and uninformative uses of these expressions, notwithstanding other potential uses they may have.

---

[1] There is another use of *after all* that expresses an outcome contrary to expectations:

(1) Trump won the general election after all.

Interestingly, whereas the use of *after all* in (4) seems corresponds to German unfocused *doch*, the use in (1) seems to correspond to the focused variant of *doch* (Rojas-Esponda, 2015).

2.2 Interrogatives

Interrogative clauses are typically used to obtain information from the addressee or other discourse participants, although (5) shows that interrogatives can be used to achieve other purposes.

(5) a. What are we having for dinner?
b. Can you pass the salt?

Only (5a) would typically be used to elicit information from the addressee; (5b) is best understood as a request for the addressee to perform a non-linguistic action. I focus on the former type of speech act, which I refer to simply as a "question."

In order to understand the contextual effect of questions in discourse, I draw on ideas from Question Under Discussion (QUD) frameworks. In particular, I have in mind the approach of Roberts (2012), although I do not present the full formalism here. We begin with the idea that the goal of the conversation is to add information to the CG. Added to this is the notion that the discourse is structured around questions and answers. The denotation of an interrogative clause is a set of propositions or a partition of logical space, with each proposition or cell in the partition corresponding to a possible answer to the question (Hamblin, 1958; Groenendijk and Stokhof, 1984). If a question is uttered and accepted in the discourse, its denotation added to the QUD stack. Assertions play their standard role of adding information to the CG, but are only relevant if they answer a question on the QUD stack (i.e. their content corresponds to a cell or union of cells in the denotation of some question on the QUD stack).

On this view, questions play the role of directing what types of information will be added to the CG as the discourse proceeds. An impotent question is therefore one that does not change the accumulation of information in the CG. According to recent work, rhetorical questions (RQs) fit this description because they are questions whose answers are already part of the CG. In particular, Rohde (2006) and Caponigro and Sprouse (2007) have argued that rhetorical questions are only felicitous in contexts in which all discourse participants already know the answer to the question. This approach is not the only existing analysis of RQs[2], but it has greater empirical coverage that other analyses in that it accounts for the fact that RQs allow for answers, but do not require them, and allow for more than just negative answers.

RQs include questions that contain a strong NPI, are followed by *yet*, or are preceded by *after all*[3] (Caponigro and Sprouse, 2007):

(6) a. Who gave a damn when Paolo was in trouble?
b. Who helped Luca when he was in trouble? Yet he managed to become what he is now.
c. After all, who helped Luca when he was in trouble?

However, none of these features are necessary for a question to be an RQ. Rohde (2006) provides many examples of naturally occurring RQs that lack these features, including (7):

(7) Who would steal a newspaper?

Despite any explicit marking, examples like (7) can be identified as RQs because of the obviousness of their answers to all discourse participants.

A potential objection to treating RQs as impotent speech acts is that even if the answer to an RQ is known to all discourse participants, this does not guarantee that the answer is in the CG. Perhaps the answer to an RQ is a mutual belief, but not part of the CG, and the RQ plays the role of converting this mutual belief to common belief. This resembles a similar point raised about uninformative assertions in the previous section. There, I noted that uninformative assertions can be used even when their content was previously established as part of the CG, and it is possible to make a similar observation in the case of RQs.

But there is another reason to doubt that RQs always play the role of converting mutual belief to common belief. RQs are not assertions, at least according to most analyses, so they cannot directly add anything to the CG. Therefore, they cannot, in themselves, convert mutual belief to common belief. One could respond that RQs do not directly add propositions to the CG, but that they indirectly achieve this once their answers are asserted and enter the CG. But this fails to account for the fact that although RQs may be answered, they typically go unanswered. If an RQ goes unanswered,

---

[2] See Sadock (1971), Ladusaw (1979), Han (2002), and van Rooy (2003) for alternative analyses of RQs.

[3] As we saw above, *after all* may also mark uninformative assertions. In light of this, the marking of RQs with *after all* may be seen as additional evidence that the analysis of RQs as questions whose answers are already in the CG is correct.

we have no account of how its answer could enter the CG without already being in the CG.

Once again, if we think of the contribution of questions to discourse in terms of a QUD stack, questions direct the evolution of the information in the CG. But this view fails to account for the role of RQs, since their answers must already be part of the CG in order for them to be used felicitously.

2.3 Imperatives

As noted in the introduction, imperatives are associated with a wide range of speech acts. However, many of the canonical speech acts associated with imperatives are directive. That is, these speech acts have the aim of getting the addressee to perform some (usually non-linguistic) action. Directive uses of imperatives encompass many different speech acts (commands, requests, etc.), but differ from non-directive uses of imperatives such as well-wishes and curses (Condoravdi and Lauer, 2012). Most theories of imperative meaning offer similar accounts for different types of directive imperatives, but may offer a different analysis for imperatives used non-directively. For the sake of concreteness, I adopt an account of the effects of directive imperatives on discourse based on Portner's (2007), but the arguments put forth below can be reformulated using other theories of imperative meaning (Condoravdi and Lauer, 2012; Kaufmann, 2012).

I assume that the successful use of an imperative $p!$ has two effects: (i) $Must(p)$, which is interpreted in a standard Kratzerian manner (Kratzer, 1981; Kratzer, 1991), enters the CG; (ii) p is added to the addressee's To-Do List (TDL), a set of propositions detailing the addressee's commitments such that if p is on the addressee's TDL, the addressee is committed to acting in oder to bring about p.[4]

Adopting this view, an impotent use of the imperative $p!$ would involve a situation in which $Must(p)$ is already in the CG and the addressee is already committed to p. Imperatives involving the verbs *remember* and *forget* ("mnemonic verbs") meet these conditions. These verbs are traditionally classified as two-way implicatives (Karttunen, 1971). This classification has the consequence that, where defined, *X didn't remember/forgot to Y* generally implies that X did not Y. Thus, *remembering/not forgetting to Y* is a necessary condition for doing Y in most normal circumstances.

Now consider the following scenario. Barbara and Richard are colleagues in a linguistics department. Barbara utters (8a) to Richard on Monday, and utters (8b) to Richard on Wednesday.

(8) a. Send me a draft of your new paper once it's ready.
    b. Remember to send me a draft of your new paper once it's ready.

Let p and q be the propositions that Richard sends Barbara a draft and that Richard remembers to send Barbara a draft, respectively.

As a result of (8a), $Must(p)$ is added to the CG and p is added to Richard's TDL. Recall that in normal circumstances, remembering to perform an action is a necessary condition for performing that action, so we have $p \Rightarrow q$. Since we're assuming a standard Kratzerian account of modality, we also have $Must(p) \Rightarrow Must(q)$. Thus, the CG entails $Must(q)$ as a result of (8a). TDLs themselves are not closed under entailment, so p's presence on Richard's TDL does not entail that q is on Richard's TDL. But recall that TDLs are a way to formalize an agent's commitments. In virtue of (8a), Richard is committing to acting in such a way to bring about p. But since $p \Rightarrow q$, so long as Richard successfully acts in accordance with his commitment to bring about p, he will also bring about q. Although q is not on Richard's TDL as a result of (8a), it is hard to see how adding it to his TDL would have any effect on his behavior.

We've seen that (8a) causes Must(q) to be entailed by the CG and commits Richard to acting in such a way that he will bring about q. But these are exactly the effects that (8b) is supposed to have on the context, meaning that (8b) is impotent.

One may object that although in normal circumstances, remembering to perform an action is a necessary condition for performing that action, this is not always the case. For example, Richard might forget that Barbara would like to see a draft

---

[4] I depart from Portner's original proposal in two ways. First, Portner does not have $p!$ add $Must(p)$ to the CG directly. Rather, $p!$ adds p to some modal ordering source g with respect to which $Must(p)$ is interpreted. Portner intends this to have the effect of making $Must(p)$ true after $p!$ is uttered, but Condoravdi and Lauer (2012) point out that adding p to a relevant ordering source is not sufficient to achieve this effect. Instead, $Must(p)$ must be added to the CG directly. Second, on Portner's original proposal, a property related to p, rather than p itself, is added to the addressee's TDL. These changes from Portner's original proposal allow for a simplification of my argument, but the argument is not dependent on them.

of his paper, but still send her a draft by accidentally CCing her on an email that has the draft attached. Thus, the entailment relation $p \Rightarrow q$ may not hold, dissolving the arguments put forth above. Or alternatively, we could maintain that $p \Rightarrow q$ but have a view of modality or agents' commitments that prevents $p \Rightarrow q$ from making (8b) impotent.

Yet on this view, standard approaches to directive imperatives predict (8b)'s function to be that of specifying that Richard should remember to send Barbara a draft in addition to sending her a draft. This mischaracterizes the role of (8b), which is, intuitively, to remind Richard of his pre-existing commitment to send Barbara a draft. Moreover, if (8b) gave Richard a new commitment not already entailed by (8a), Barbara should be able to felicitously utter both (8a) and (8b) in succession on Monday:

(9) # Send me a draft of your new paper once it's ready and remember to send me a draft of your new paper once it's ready.

The redundancy of (9) shows that the mnemonic imperative in (8b) does not communicate anything that was not already communicated by (8a).

To recap, we've seen that imperatives involving mnemonic verbs do not function like normal directive imperatives. They do not add information to the CG or commit the addressee to an action they were not already committed to.

3 Awareness

We have seen that speech acts associated with different clause types may fail to have the effects normally associated with these speech acts. What, then, is the point of these impotent speech acts?

I propose that the utility of impotent speech acts lies in their ability to raise awareness or draw attention to issues that are already settled by the discourse (e.g. are already entailed by the CG, are already commitments of a discourse participant, etc.), but that discourse participants may be ignoring. Often, linguistic agents are idealized as being logically omniscient and perfectly rational, but such an idealization is not psychologically realistic. Real-world agents face cognitive limitations, including but not limited to, constraints on attention and awareness. In light of this reality, it would be unsurprising to find that speakers design have as one of their goals the manipulation of other agents' states of awareness.

Mnemonic imperatives provide a clear case of impotent speech acts whose purpose is to raise awareness. As noted above, the goal of these utterances is to remind the addressee of some pre-existing commitment, rather than to form some new commitment. Impotent assertions and questions can also be understood from this perspective. Following Barker and Taranto's observation that assertions of clarity are often uninformative, Bronnikov (2008) and Crone (2016) have analyzed clarity assertions as playing the role of drawing attention to inferences that can be made on the basis of information in the CG or raising awareness of propositions already entailed by the CG. Caponigro and Sprouse (2007) come to a similar conclusion regarding RQs, claiming that they "highlight" a proposition within the CG for reasons related to the structure of the discourse.

3.1 Awareness & Discourse Structure

Even if we take impotent speech acts to play the role of manipulating discourse participants' awareness states, we may still wonder why manipulating awareness would matter. Caponigro and Sprouse's claim about rhetorical questions serving purposes related to discourse structure provides one answer. By drawing attention to some issue that is already in the CG, impotent speech acts may establish this issue as a discourse topic worthy of further elaboration (Asher, 2004).

Drawing attention to an issue may also prove useful if a speaker leverages discourse coherence relations that listeners often infer. Kehler (2005) provides a comprehensive overview of such relations. To give just one example, listeners often infer from the successive assertion of S_1 and S_2 that S_1 causes S_2.

(10) As you already know, Trump terrifies me. I'm thinking of moving to Canada.

In (10), we easily infer that the speaker is thinking of moving to Canada *because* Trump terrifies the speaker, even though this is never explicitly asserted. The speaker achieves this effect by uttering an impotent speech act, which raises awareness of the common knowledge that Trump terrifies the speaker. The speaker then relies the listener's ability to infer the correct discourse relation to understand the full intended meaning of (10).

3.2 Awareness & Decision Making

Unawareness is also known to play a significant role in reasoning and decision making. Much work within economics and computer science, particularly following Fagin and Halpern (1987), has addressed the issue of decision making under unawareness. Unawareness has received less attention in linguistics, although Franke and de Jager (2011) provide a formal model for understanding the role of awareness in discourse.

A key feature of Franke and de Jager's approach is the recognition that unawareness may take one of two forms, with the difference having to do with the presence or absence of what Franke and de Jager call "implicit assumptions." To illustrate the notion of an implicit assumption, they use the example of Little Bo Peep searching for her keys throughout her apartment. Bo turns up empty-handed everywhere she looks, when her friend Little Jack Horner utters the following:

(11) Did you leave them in the car when you came in last night?

Bo slaps her forehead and immediately runs out to search for the keys in her car.

In this situation, Bo is initially unaware of the possibility that her keys are in the car, and her unawareness causes her to behave as if she knows that they keys are not in the car. After all, if she thought there was a slight chance that keys were in the car, she would have looked there after her other searches proved futile. Franke and de Jager say that an agent in such a state makes an implicit assumption; in this case, Bo implicitly assumes that her keys are not in the car.

An agent may be unaware of an issue without making an implicit assumption. To use an example from Yalcin (2011), an agent may not be considering whether it is currently raining in Topeka, Kansas. Such an agent is unaware of issue of whether it is raining in Topeka and cannot distinguish between possible worlds in which it is raining Topeka and those in which it is not. But this agent's behavior does not reflect an assumption that it is or is not raining in Topeka. Rather, the agent is unaware of an issue without making an implicit assumption.

As shown by the car keys example, implicit assumptions have an effect on agents' behavior. This provides yet another reason why it might be worth raising awareness of an issue. If an agent is making an implicit assumption about an issue, raising awareness of that issue may positively affect the agent's resolution of decision problems.

We can formalize this using the following model based on Franke and de Jager's.[5] Let $\mathcal{W}$ be a set of worlds, let $\mathcal{P} = \wp(\mathcal{W})$ be a set of propositions, and let $\mathcal{A}$ be a set of actions. We define for each agent α a background probability distribution over propositions $P_\alpha : \mathcal{P} \rightarrow [0, 1]$ and a utility function $U_\alpha : \mathcal{W} \times \mathcal{A} \rightarrow \mathbb{R}$. Awareness is modeled via an awareness state $\langle \mathfrak{U}_\alpha, \mathfrak{v}_\alpha \rangle$, where:

- $\mathfrak{U}_\alpha \subseteq \mathcal{P}$ is the set propositions of which α is unaware; $\mathfrak{U}_\alpha$ is closed under complement.
- $\mathfrak{v}_\alpha : \mathfrak{U}_\alpha \rightarrow \{\mathrm{T}, \mathrm{F}\}$ is a partial valuation function from unmentionable propositions to truth-values. This function encapsulates α's implicit assumptions. We require that if $\mathfrak{v}_\alpha(p)$ is defined, $\mathfrak{v}_\alpha(p) = \neg\mathfrak{v}_\alpha(\mathcal{W} \setminus p)$.

We next use the agent α's background probability distribution and awareness state to model α's probability distribution under unawareness P'_α:

$$P'_\alpha = P_\alpha(\cdot \mid \{w \in \mathcal{W} | \forall p \in \mathcal{P}(\mathfrak{v}_\alpha(p) = \mathrm{T} \rightarrow w \in p)\})$$

That is, an agent's probability distribution under unawareness is simply their background distribution conditioned on their implicit assumptions. An agent α acts by choosing the action with the highest expected utility given P'_α:

$$\mathrm{EU}_\alpha(\mathtt{a}) = \sum_{w \in \mathcal{W}} \mathcal{U}_\alpha(w, \mathtt{a}) \times P'_\alpha(\{w\})$$

An agent's awareness state can be modified by both linguistic events, such as Jack's utterance in (11), and non-linguistic events. Here, I focus only on how agents' awareness states are affected by other agents' utterances. In principle, it would be desirable to give necessary and sufficient conditions for a particular utterance u to make an agent aware of a proposition p. Formalizations of "attentive content" within the framework of Inquisitive Semantics (Ciardelli et al., 2011; Roelofsen, 2013) seem to have this goal in mind. Unfortunately, it is unlikely that such necessary and sufficient conditions can be given. To illustrate the

---

[5] Dekel et al. (1998) purport to show that standard possible worlds models, such as that proposed by Franke and de Jager, preclude non-trivial unawareness. However, Fritz and Lederman (2015) have shown that Dekel et al.'s result relies on several strong, psychologically implausible assumptions. Fritz and Lederman propose their own model of unawareness based on partitions of the set of all possible worlds, which is not dissimilar from Franke and de Jager's proposal.

difficulty, note that Jack could draw Bo's attention to the possibility of her keys being in her car by uttering the following:

(12) Sometimes I leave my keys in the car.

The literal content (12) is about *Jack's* keys and *his* car. It says nothing directly about Bo's keys or her car. Of course, we can understand the relevance of Jack's utterance in (12) to Bo's search given our knowledge of its context of utterance, but it is hard to see how we could derive its potential effect on Bo's awareness state from its literal meaning alone. Examples like (12) show that an utterance's ability to raise awareness of issues is highly context dependent. Because of this, when modeling scenarios involving changes to awareness states, we simply must stipulate the effects that particular utterances have on these states.

Let's consider how Franke and de Jager's model applies to impotent speech acts, focusing on the example using imperatives discussed above. Recall that on Monday Barbara requests that Richard send her a draft of his new paper (8a), and that on Wednesday she reminds him of this request (8b).

For the sake of simplicity, I abstract away from details regarding Richard's TDL, although in principle we could extend the notion of unawareness to TDLs or other formal devices used to model discourse contexts. So, the only effect of (8a) modeled here is the addition of *Richard must send Barbara a draft of his paper (given Barbara's wishes)* to the CG. We'll refer to this proposition as $Must(p)$. We only have two worlds in $\mathcal{W}$, w_1 and w_2, that differ only with respect to the truth-value of $Must(p)$: $Must(p) = \{w_1\}$ and $\neg Must(p) = \{w_2\}$. The set of possible actions that Richard consists in sending Barbara a draft or doing nothing: $\mathcal{A} = \{\texttt{send-draft}, \texttt{do-nothing}\}$.

On Wednesday, Richard's background model assigns a high, but non-maximal, probability to $Must(p)$. Even though Barbara's utterance in (8a) ensures that $Must(p)$ enters the CG on Monday, there is a small possibility that things could have changed by Wednesday. Perhaps Barbara has decided that Richard is a hack, and no longer wants to read any of his work. We'll assume that $P_{\mathbf{R}}(Must(p)) = 0.95$.[6]

We'll also assume that Richard is generally accommodating to Barbara's wishes, so that if she would like him to send a draft, sending the draft has a high utility. On the other hand, it is socially costly for Richard to not send a draft if Barbara would like him to. If she does not want to see the draft, sending it will incur a small cost, whereas doing nothing will be neutral.

$$U_{\mathbf{R}}(w, \mathtt{a}) = \begin{cases} 1 & \text{if } (w,\mathtt{a}) = (w_1, \texttt{send-draft}) \\ -1 & \text{if } (w,\mathtt{a}) = (w_1, \texttt{do-nothing}) \\ -0.25 & \text{if } (w,\mathtt{a}) = (w_2, \texttt{send-draft}) \\ 0 & \text{if } (w,\mathtt{a}) = (w_2, \texttt{do-nothing}) \end{cases}$$

Under full awareness, $\text{EU}_{\mathbf{R}}(\texttt{send-draft}) = 0.9375$ and $\text{EU}_{\mathbf{R}}(\texttt{do-nothing}) = -0.95$. Clearly, sending the draft is the right call. But there is a chance that Richard will forget about $Must(p)$, and if he does, he may behave as if Barbara had no desire to see his draft. In other words, he may operate with an implicit assumption that $Must(p)$ is false. This is a plausible assumption for him to make, since his default belief had (8a) never been uttered would likely have been that he had no commitment to send Barbara a draft.

To model this situation, we have $\mathfrak{U}_{\mathbf{R}} = \{Must(p), \neg Must(p)\}$ and $\mathfrak{v}_{\mathbf{R}}(Must(p)) = \mathrm{F}$. Now, when we look at Richard's probability distribution under unawareness, we have $P'_{\mathbf{R}}(Must(p)) = 0$. As a consequence, the expected utilities of each action change to the following: $\text{EU}_{\mathbf{R}}(\texttt{send-draft}) = -0.25$, $\text{EU}_{\mathbf{R}}(\texttt{do-nothing}) = 0$. Doing nothing is now the action with the greatest expected utility.

We are now in a position to explain Barbara's utterance in (8b). If she believes that Richard is making an implicit assumption that $Must(p)$ is false, making him aware of $Must(p)$ by uttering (8b) will change his behavior in such a way that it is more likely that he will send her a draft of his paper. For space reasons, I do not illustrate how impotent assertions and questions would be explained in this decision theoretic approach, but the basic idea is the same. If an agent makes an implicit assumption about some previously settled issue, overturning that assumption via an impotent speech act can have important consequences for the agent's behavior.

4 Why *Impotent* Speech Acts?

As we saw in the previous section, impotent speech acts are not the only speech acts that raise awareness. Rather, any utterance has the potential to raise awareness of its content or of related

---

[6] There is also a technical reason for taking $P_{\mathbf{R}}(Must(p))$ to be non-maximal. Eventually, we condition on $Must(p)$ being false, which is problematic if $P_{\mathbf{R}}(Must(p)) = 1$.

propositions. This raises the question of why a speaker would use an impotent speech act to raise awareness, rather than some other speech act. For example, suppose I am about to move to Canada and that this is known to my addressee. I would now like to invite my addressee to my big going-away party, and I begin the discourse by drawing attention to my upcoming move. Given the model presented in the previous section, either (13a) or (13b) would achieve the intended effect.

(13) a. As you know, I'm moving to Canada.
b. ? I'm moving to Canada.

Moreover, (13b) is a simpler expression, so we might expect it to be the preferred method of raising awareness of my plans for manner-related considerations. But this is not what we find; (13b) is marked if it is already established that my addressee knows about my move.

We can explain the preference for (13a) as follows. If we consider only the literal content of (13a) and (13b), both are uninformative in the context described. But in another sense, (13a) is *more* informative than (13b). This informativity derives from the general norm of language use that a speaker asserting p believes p. By this principle, a speaker who asserts (13a) communicates that the speaker takes the addressee to be knowledgeable about the speaker's plans to move to Canada. In contrast, (13b) does not communicate this. Thus, (13a) may be preferred for straightforward informativity-related reasons.

This reasoning can be taken one step further to consider the implicatures that would be generated by the assertion of (13b). If (13b) is in pragmatic competition with (13a) or some similar impotent assertion, then use of the less informative (13b) may generate the implicature that the speaker was not in a position to assert (13a). This would occur if the speaker did not take the addressee to be knowledgeable about the speaker's plans to move. Thus, use of (13b) may generate the undesirable implicature that the speaker takes the addressee to be unknowledgeable about the relevant information and that the speaker's intention was to inform the addressee, not simply to raise awareness of a piece of common knowledge.

At this point, one might object that before uttering (13b), the speaker must have already communicated their plans to the addressee. Therefore, the addressee should already know that the speaker knows that the addressee knows about the plans to move to Canada, and this knowledge is incompatible with the implicature that I say (13b) generates. The problem with this objection is that it relies on the assumption that linguistic agents are perfectly rational and are not affected by lapses in memory or attention. In the real world, a speaker may have forgotten that they previously told their addressee about their plans. In this case, a speaker would be predicted to use (13b), since (13a) would be false. Or perhaps the speaker remembers having tried to tell their addressee about the plans on a previous occasion, but thinks the addressee did not hear them or cannot retrieve the memory about these plans. In such contexts, a speaker would opt for (13b). But if the speaker does not think these conditions obtain, they would wish to avoid generating the implicature that they do.

As noted in §2.1, speakers may use expressions such as *As you know, p* non-literally for politeness reasons. Such uses relate to the pragmatic reasoning discussed here. As was just discussed, an outright assertion of p may implicate that the addressee is not knowledgeable about p. Such an implication may be highly face-threatening towards the addressee, and therefore impolite (Brown and Levinson, 1987). Thus, even when an addressee does not already know of the speaker's plans to move to Canada, a speaker may falsely utter (13a) if politeness considerations overrode concerns about truthfulness.

The reasoning employed to motivate the use of (13a) over (13b) can also explain the use of imperatives in the context involving Barbara and Richard. When Barbara wants to remind Richard to send her a draft on Wednesday, it is much more felicitous for her to use (14a), rather than (14b).

(14) a. Remember to send me a draft of your paper once it's ready.
b. ? Send me a draft of your paper once it's ready.

Again, both (14a) and (14b) would raise awareness of Barbara's desire to see Richard's draft, and (14b) is a simpler expression. But if Barbara has already uttered (8a) on a previous occasion, (14a) is much more natural than (14b).

We can explain the preference for (14a) by first noting that mnemonic verbs have a presuppositional component to their meaning (Karttunen, 1971; White, 2014). In the case of (14a), this presupposition is (roughly) that Richard is committed to sending Barbara a draft of his paper. In uttering

(14a), Barbara communicates that she believes this commitment can be presupposed. No such thing is communicated by (14b), which may generate the implicature that Barbara does not believe that Richard believes he has any such commitment.

I have sketched this reasoning at a high level without committing to any particular pragmatic theory. But the basic ideas are general enough that the basic reasoning should be easily implemented in any desired approach to pragmatics, e.g. classical Griceanism (Grice, 1975), a Neo-Gricean system (Levinson, 2000; Horn, 2004), or game-theoretic or Bayesian models of pragmatics (Franke, 2009; Frank and Goodman, 2012; Jäger, 2012). Across these approaches, what remains constant is the idea that impotent speech acts are informative about the speaker's beliefs about the addressee in a way that other speech acts are not.

5 Related Work & Conclusion

Across different clause types, we find speech acts that do not have the expected effects on the discourse. These impotent speech acts are nonetheless useful because of their ability to raise awareness of issues that discourse participants may be ignoring. At the same time, impotent speech acts communicate information about the speaker's beliefs about other discourse participants that are not communicated by alternative expressions.

The claim that impotent speech acts should be understood in terms of their awareness-raising potential joins a larger body of recent work focusing on discourse phenomena that are best understood in terms of their effects on the awareness states of interlocutors. This work includes Franke and de Jager's (2011) discussion of the effects of uninformative questions on agents' behavior, Rawlins's (2010) work on "conversational backoff," and Ciardelli et al.'s (2011) and Roelofsen's (2013) attention-based analyses of *might*.

Earlier work by Walker (1993) aligns even more closely to the phenomena discussed here.[7] Walker discusses "informationally redundant utterances" (IRUs), utterances whose informational content is completely hearer old. While IRUs may seem equivalent to impotent speech acts, there are an important distinctions between the two notions. First, in some cases it may be misleading to characterize impotent speech acts as having any informational content *per se* (e.g. RQs, mnemonic imperatives). Second, the redundancy of IRUs often follows from the fact that they are repetitions of information established earlier in a particular context. In contrast, impotent speech acts are often impotent across *all* contexts. For example, an assertion of *As you know, p* should always be uninformative whenever it is true.

Despite these differences, Walker's work on IRUs points to additional uses that impotent speech acts may have. Walker provides computational results that show the utility of raising awareness of old information with IRUs for resource-bounded agents. In addition, Walker notes that IRUs can help resolve uncertainty about old information and aid agents in drawing inferences. I leave for future work investigations of how impotent speech acts may have similar effects.

To close, I highlight two additional future directions for research on these topics. First, there is a great deal of work to be done to find out how different languages mark impotent speech acts. I have focused almost entirely on English data, the one exception being a note about German discourse particles. Undoubtedly, more cross-linguistic data will help further refine our understanding of impotent speech acts and their role in discourse. The German example may prove instructive in that we may find examples of discourse particles in other languages that similarly mark redundant information.

Second, raising awareness of issues is important both for the structure of discourse, as well as for decision making. Yet the formal model of awareness in discourse adapted from Franke and de Jager (2011) is only well-suited to capture the decision theoretic implications of raising awareness. Ultimately, it would be desirable to have a unified model of awareness as it relates to discourse structure as well as decision making.

Acknowledgments

My thanks go out to Dan Lassiter, Cleo Condoravdi, Chris Potts, anonymous reviewers for JerSem, anonymous reviewers and attendees of WCCFL 34, attendees of CUSP 8, and the members of the Stanford University SemPrag group for their support and feedback.

---

[7] My thanks to an anonymous reviewer for pointing out this connection.

References

Nicholas Asher. 2004. Discourse topic. *Theoretical Linguistics*, 30:163–201.

Chris Barker and Gina Taranto. 2003. The paradox of asserting clarity. In Pälvi Koskinen, editor, *Proceedings of the Western Conference on Linguistics (WECOL) 2002*, volume 14, pages 10–21, Fresno, CA. Department of Linguistics, California State University.

Chris Barker. 2009. Clarity and the grammar of skepticism. *Mind & Language*, 24(3):253–273.

George Bronnikov. 2008. The paradox of clarity: Defending the missing inference theory. In Tova Friedman and Satoshi Ito, editors, *Proceedings of SALT 18*, pages 144–157, Ithaca, NY. Cornell University.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge.

Ivano Caponigro and Jon Sprouse. 2007. Rhetorical questions as questions. In Estela Puig-Waldmüller, editor, *Proceedings of Sinn und Bedeutung 11*, pages 121–133.

Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. 2011. Attention! *Might* in Inquisitive Semantics. In E. Cormany, Satoshi Ito, and David Lutz, editors, *Proceedings of SALT 19*, pages 91–108.

Cleo Condoravdi and Sven Lauer. 2012. Imperatives: meaning and illocutionary force. In Christopher Piñon, editor, *Empirical Issues in Syntax and Semantics 9*, pages 37–58. CCSP.

Phil Crone. 2016. Asserting clarity and manipulating awareness. In *Proceedings of Sinn und Bedeutung 20*. To appear.

Eddie Dekel, Barton L. Lipman, and Aldo Rustichini. 1998. Standard state-space models preclude unawareness. *Econometrica*, 66(1):158–173.

Ronald Fagin and Joseph Y. Halpern. 1987. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1):39–76.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Michael Franke and Tikitu de Jager. 2011. Now that you mention it: Awareness dynamics in discourse and decisions. In Anton Benz, Christian Ebert, Gerhard Jäger, and Robert van Rooij, editors, *Language, Games, and Evolution*, pages 60–91. Springer-Verlag, Berlin.

Michael Franke. 2009. *Signal to Act*. Ph.D. thesis, University of Amsterdam.

Peter Fritz and Harvey Lederman. 2015. Standard state space models of unawareness. In *Proceedings of TARK XV: Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge*, Pittsburgh.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 3. Academic Press.

J. Groenendijk and M. Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam, Amsterdam.

Christine Gunlogson. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Ph.D. thesis, University of California, Santa Cruz.

Charles Hamblin. 1958. Questions. *Australasian Journal of Philosophy*, 36(3):159–68.

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

Laurence R. Horn. 2004. Implicature. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*. Blackwell Publishing, Oxford.

Gerhard Jäger. 2012. Game theory in semantics and pragmatics. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics. An International Handbook of Natural Language Meaning*, volume 3, pages 2487–2516. De Gruyter Mouton, Berlin.

Lauri Karttunen. 1971. Implicative verbs. *Language*, 47(2):340–358.

Magdalena Kaufmann and Stefan Kaufmann. 2012. Epistemic particles and performativity. In Anca Chereches, editor, *Proceedings of SALT 22*, pages 208–225, Chicago.

Magdalena Kaufmann. 2012. *Interpreting Imperatives*. Springer, Dordrecht.

Andrew Kehler. 2005. Discourse coherence. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*. Blackwell Publishing.

Angelika Kratzer. 1981. The notional category of modality. In Hans-Jürgen Eikmeyer and Hannes Rieser, editors, *Words, Worlds, and Contexts: New Approaches in Word Semantics*, pages 38–74. De Gruyter, Berlin and New York.

Angelika Kratzer. 1991. Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*, pages 639–650. de Gruyter, Berlin.

William A. Ladusaw. 1979. *Polarity Sensitivity as Inherent Scope Relations*. Ph.D. thesis, University of Massachusetts, Amherst, Amherst, MA.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge.

Paul Portner. 2007. Imperatives and modals. *Natural Language Semantics*, 15(4):3151–383.

Kyle Rawlins. 2010. Conversational backoff. In Nan Li and David Lutz, editors, *Proceedings of SALT 20*, pages 347–365, Ithaca, NY. CLC Publications.

Craige Roberts. 2012. Information structure. *Semantics and Pragmatics*, 5(6):1–69.

Floris Roelofsen. 2013. A bare bones attentive semantics for *might*. In Maria Aloni, Michael Franke, and Floris Roelofsen, editors, *The dynamic, inquisitive, and visionary life of* ϕ, $?\phi$, *and* $\Diamond\phi$: *a festschrift for Jeroen Groenendijk, Martin Stokhof, and Frank Veltman*, pages 190–215. Institute for Logic, Language, and Computation, Amsterdam.

Hannah Rohde. 2006. Rhetorical questions as redundant interrogatives. *San Diego Linguistic Papers*, 2:134–168.

Tania Rojas-Esponda. 2015. *Patterns and Symmetries for Discourse Particles*. Ph.D. thesis, Stanford University, Stanford.

Jerrold Sadock. 1971. Queclaratives. In *Proceedings from the Seventh Regional Meeting of the Chicago Linguistics Society (CLS 7)*, pages 223–331, Chicago. Chicago Linguistics Society.

Robert Stalnaker. 1978. Assertion. In Peter Cole, editor, *Syntax and Semantics 9: Pragmatics*. Academic Press, New York.

Robert C. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25:701–721.

Robert van Rooy. 2003. Negative polarity items in questions: Strength as relevance. *Journal of Semantics*, 20:239–273.

Marilyn A. Walker. 1993. *Informational Redundancy and Resource Bounds in Dialogue*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Aaron Steven White. 2014. Factive-implicatives and modalized complements. In Jyoti Iyer and Leland Kusmer, editors, *NELS 44: Proceedings of the* 44th *Meeting of the North East Linguistic Society*, Amherst. GLSA Publications.

Seth Yalcin. 2011. Nonfactualism about epistemic modality. In Andy Egan and Brian Weatherson, editors, *Epistemic Modality*, pages 295–332. Oxford University Press, New York.

Multi-layered analysis of laughter

Chiara Mazzocconi[2], **Ye Tian**[1], **Jonathan Ginzburg**[2,3]
[1]Laboratoire Linguistique Formelle (UMR 7110)
& [2]CLILLAC-ARP (EA 3967) & [3]Laboratoire d'Excellence (LabEx)—EFL
Université Paris-Diderot, Paris, France
`tiany.03@googlemail.com`

This paper presents a multi-layered classification of laughter in French and Chinese dialogues (from the DUEL corpus). Analysis related to the form, the semantic meaning and the function of laughter and its context provides a detailed study of the range of uses of laughter and their distributions. A similar distribution was observed in most of the data collected for French and Chinese. We ground our classification in a formal semantic and pragmatic analysis. We propose that most functions of laughter can be analyzed by positing a unified meaning with two dimensions, which when aligned with rich contextual reasoning, yields a wide range of functions. However, we also argue that a proper treatment of laughter involves a significant conceptual modification of information state account of dialogue to incorporate emotive aspects of interaction.

1 Introduction

Laughter is very frequent in everyday conversational interaction—(Vettin and Todt, 2005) suggest a frequency of 5,8/10 min of conversation. Although we can easily recognize laughter, it is not a homogeneous phenomenon. Laughter can take various forms and occur in a variety of contexts. Attempts to understand the nature of laughter go back as early as Aristotle, frequently intertwined with theories concerning humour. There have been many proposals on the laughter types yet little agreement on how laughter should be classified. We believe that one reason for the lack of agreement is that there are several layers relevant to the analysis of laughter. Different classification systems and even types within systems in fact often relate to different layers of analysis. In what follows, we will initially present a brief critical review of studies on laughter types. Building on this, we propose a multi-layered analysis of laughter, including a novel analysis of the meaning of laughter and attempt to describe its various uses. We then present our corpus study in sections 4 and 5. In section 7, on the base of our data observation, we will try to ground our classification in a formal semantic and pragmatic analysis within the KoS framework (Ginzburg, 2012).

2 Background

2.1 Existing taxonomies/classifications

Studies on laughter classification concern at least three areas: the sound, the context and the function[1]. Studies on the sound of laughter analyze phonetic, acoustic, para-linguistic, kinesic and anatomical features e.g.(Poyatos, 1993; Urbain and Dutoit, 2011; Trouvain, 2003; Provine and Yong, 1991, for example)), or propose constitutive elements of laughter (Kipper and Todt, 2003; Trouvain, 2003; Bachorowski and Owren, 2001; Campbell et al., 2005; Tanaka and Campbell, 2014; Nwokah and Fogel, 1993; Ruch and Ekman, 2001, for example). Due to space constraints and pertinence we will focus on reviewing analyses on contextual and functional classifications.

2.1.1 Contextual classifications

Studies on context of laughter investigate the stimuli (triggers) and the position of a given laughter event in relation to other components in conversation (e.g. speech and partner's laughter). Studies on laughter stimuli distinguish those that are "funny" (though that in itself is a tricky matter to characterize) and those that are not. It has been reported that contrary to 'folk wisdom', most laughters in fact follow a stimuli that is not "funny" (Coates, 2007; Provine, 2004).

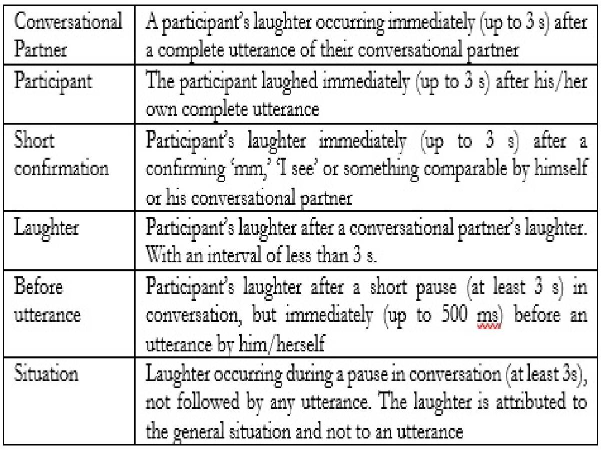
A second level of contextual analysis concerns the position of laughter in relation to laughter (or lack thereof) of a partner. With mildly differing parameters and timing thresholds, several authors distinguish between *isolated laughter* i.e. laughter not shortly preceded nor followed by

---

[1] There are also proposals on the causes of laughter e.g. (Morreall, 1983; Owren et al., 2003; Bachorowski and Owren, 2001)

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*,
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 97-107.

others' laughter, (Nwokah et al., 1994), *reciprocal/antiphonal/chiming in laughter* i.e., laughter that occurs immediately after partners laughter (Nwokah et al., 1994; Smoski and Bachorowski, 2003; Hayakawa, 2003), and *co-active/plural laughter* (Nwokah et al., 1994; Hayakawa, 2003). (Vettin and Todt, 2004) make an initial distinction between speaker and audience laughter. Then, they characterize the event preceding the laughter as being a complete sentence, a short confirmation, or a laughter bout. Combining these parameters, they obtain 6 mutually exclusive contexts for laughter to occur (see 1):



Conversational Partner	A participant's laughter occurring immediately (up to 3 s) after a complete utterance of their conversational partner
Participant	The participant laughed immediately (up to 3 s) after his/her own complete utterance
Short confirmation	Participant's laughter immediately (up to 3 s) after a confirming 'mm,' 'I see' or something comparable by himself or his conversational partner
Laughter	Participant's laughter after a conversational partner's laughter. With an interval of less than 3 s.
Before utterance	Participant's laughter after a short pause (at least 3 s) in conversation, but immediately (up to 500 ms) before an utterance by him/herself
Situation	Laughter occurring during a pause in conversation (at least 3s), not followed by any utterance. The laughter is attributed to the general situation and not to an utterance

Figure 1: Vetting and Todt, 2004 - Context classification

2.1.2 Functional classifications

This is the area where debate is quite unresolved. Many taxonomies have been proposed; some contain binary types and others contain dozens. The most problematic issue is that very often, taxonomies have within them a mixture of types of function and types regarding triggers.

(Szameitat et al., 2009) distinguishes between physical (tickling) and emotional laughter (including joy, taunts, and *schadenfreude*). While (Poyatos, 1993) bases its classification on the social functions that laughter might have. He defines laughter as a *paralinguistic differentiator* (one that allows the differentiation of physiological and emotional states and reactions among interlocutors). He distinguishes at least eight social functions: affiliation, aggression, social anxiety, fear, joy, comicality and ludicrousness, amusement and social interaction, self-directedness. (Shimizu et al., 1994) identifies three types of laughter: laughter due to pleasant feeling, sociable laughter, and laughter for releasing tension. (Hayakawa, 2003) distinguishes three non-mutually-exclusive functions: laughter for joining a group, balancing laughter for releasing tension, laughter as a concealer (to soften or evade). A yet different classification comes from (Campbell et al., 2005; Reuderink et al., 2008), where four laughter types are distinguished on the basis of perceptual analyses of their characteristics: hearty, amused, satirical, social.

2.1.3 Weaknesses of existing classifications

A common issue with most taxonomies, as has been mentioned before, is that they contain types that relate to different layers of analysis. For example, in (Poyatos, 1993)'s taxonomy, affiliation (e.g., agree) is roughly the illocutionary act performed by a laughter, while joy is a feature of the laughter trigger. Apart from that, at least three issues can be raised.

Contextual classification: (Vettin and Todt, 2004) use exclusively timing parameters (i.e., what precedes and what follows) to support claims about laughter eliciting situations. However, their classification runs into problems in the way it deals with the referentiality of laughter. In Figure 2 we schematize some possible patterns observed in our corpus when conducting a detailed analysis of each laughter in relation to its laughable. Laughter can refer both to events that precede or follow it, but also to events or utterances with which it overlaps. Timing parameters are not optimal as a means for inferring the referent of laughter given that significant time misalignment can occur between the laughter and the laughable, namely their lack of adjacency.

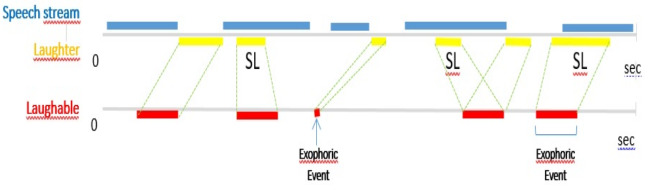


Figure 2: Temporal misalignment speech stream, laughter and laughable

Unfunniness: The proposal from (Provine, 1993) that laughter is not usually related to "humourous stimuli" is made by assuming what a laughter is about is what immediately precedes the laughter. As we have already pointed out, there is much freedom in the alignment between laughable and laughter, so a laugh can be about not the preceding utterance but the utterance before, or an upcoming utterance. Moreover, even if the laughable *is* the preceding utterance, funniness rarely resides simply in the utterance itself, but is most frequently in the relation between that utterance and the context, or can reside in the enriched content of the utterance mot necessarily accessible to an extraneous listener. Therefore, it is misleading to conclude what laughter is about by analyzing merely what immediately precedes laughter.

Acoustically-based classification: Studies such as (Campbell et al., 2005; Tanaka and Campbell, 2014) classified the function of laughter on acoustic parameters only. (Tanaka and Campbell, 2014) asked participants to listen to the laughter bouts and judge whether it is a "mirthful" or "politeness" laughter. In the first instance we would like to point out that the names of the categories "mirthful" and "politeness" do not belong to the same level of analysis: one can feel "mirthful", but cannot feel "politeness"; and the two categories are not mutually exclusive, i.e., one can politely laugh while feeling mirthful, and one can impolitely laugh without feeling mirthful. We believe that laughters with similar acoustic features can have different functions in different contexts. We will move a first attempt to test this issue in our data, specifically whether the function of laughter can be predicted by context *and* form-based measurements, deferring a more detailed analysis of phonetic aspects to further studies.

3 A multi-layer analysis of laughter

We have argued that the confusion in laughter type classification comes from not distinguishing different levels of analysis. An additional intrinsic problem for previous analyses is that they did not attempt to integrate their account with an explicit semantic/pragmatic module on the basis of which content is computed.[2] The sole recent exception to this, as far as we are aware, is the account of (Ginzburg et al., 2015), which sketches an information state–based account of the meaning and use of laughter in dialogue. We take that account as our starting point, though that account has a number of significant lacunae which we point to here and (some of) which we briefly sketch means of plugging in section 7. The purpose of the current study is to test a new method for laughter analysis whereby each laughter episode is described by means of the following: its context of occurrence both in relation to the laughable, to other's laughter and the other's or laugher's own speech[3]; the nature of the laughable; its pragmatic use (whether laughter is used in its literal or ironical meaning); the amount of arousal perceived by the listener; and finally, in the function that it serves in the specific context of occurrence.

The account of (Ginzburg et al., 2015) views laughter essentially as an event anaphor. They associate two basic meanings with laughter, one involving the person laughing expressing her *enjoyment* of the laughable l, the other expressing her perception of l as being *incongruous*. These meanings, combined with a dialogical reasoning theory, Breitholtz and Cooper's enthymatic approach (see e.g., (Breitholtz, 2014)), allow one to deduce a potentially unlimited set of functions that laughter can exhibit. For instance, seriousness cancellation (of an assertion or query), scare quotation, and acknowledgment.

The account focuses on the laughter stimulus or trigger, i.e., the laughable. One question to raise here is whether incongruity and enjoyment are the only two dimensions to distinguish the person laughing's relation to the laughable. Certain uses we see below suggest, arguably, the need for a third possible relation pertaining to *ingroupness* or *sympathy*.

Be that as it may, the account due to (Ginzburg et al., 2015), abstracts away from a significant dimension of laughter, namely *arousal*. In line with (Morreall, 1983) we think that laughter effects a "positive psychological shift". Thus, an additional dimension we identify is one which relates to arousal. This can go from very low to extremely high, and different amplitudes in the shift can depend on the trigger itself and on the individual current information/emotional state. It is important to point out that laughter does not signal that the speaker's current emotional state is positive, merely that there was a shift which was positive. The speaker could have a very negative baseline emotional state (being very sad or angry) but the recognition of the incongruity in the laughable or its enjoyment can provoke a positive shift (which could be very minor). The distinction between the overall emotional state and the direction of the shift explains why laughter can be produced when one is sad or angry.

We therefore claim that the "literal" meaning

---

[2] This is not the case for some theories of humour. For example, (Raskin, 1985) offers a reasonably explicit account of incongruity emanating from verbal content. However he did not attempt to offer a theory of laughter in conversation.

[3] See (Nwokah et al., 1999; Kohler, 2008; Trouvain, 2001; Menezes and Igarashi, 2006) for detailed descriptions of acoustic features of speech-laughs.

conveyed by a laughter (more or less genuinely) is that a stimulus y has triggered in the laugher a positive arousal shift of the value x. Like language it can be used ironically, intending to convey exactly the opposite of its literal meaning i.e., the stimulus y totally didnt trigger in me a positive shift in the arousal of any value. A more detailed analysis of ironic laughter is a topic for future study.

What about function? We distinguish the functions of laughter from its form, its meaning and its triggers, in contrast with previously proposed classifications (see section 2.1.3). As we mentioned above, (Ginzburg et al., 2015) sketch how some functions can be derived from the meanings they posit in conjunction with a theory of dialogical reasoning. However, they do not propose a systematic repertory of possible functions. Building on previous work, we conducted a detailed overview of the possible functions that laughter could serve in interaction. We believe that an efficient way to partition them is to differentiate two big classes—*cooperative* functions that promote the continuation of interaction (e.g., show enjoyment, show agreement, and softening) and *non-cooperative* functions that damage the flow of the interaction (e.g. mocking, showing disagreement)[4]. Following are some examples from our corpus exemplifying this—**film script, border control, dream apartment** are names of the tasks the participants were engaged in, further described in section 4; the laughter serves the function given in capitals and lasts throughout the text surrounded by < laughter > and < /laughter >:

1. SHOW-ENJOYMENT (**film script**) A: there is one one of my buddies stupid as he is who who put a steak on the border of the, of the balcony B: < laughter > you have weird buddies! < /laughter >
2. SMOOTHING: second laughter of B (**border control**) A: You are dealing with my visa? Then it will be very easy right? < laughter/ > B: < laughter/ > But we have to follow the rules. I have to < laughter/ > ask you some questions.
3. SHOW-AGREEMENT (**dream apartment**) A: and then in the evening we can cook a very good pasta! B: < laughter/ > yes! why not?
4. BENEVOLENCE-INDUCTION (**film script**) B: actually we need to think about what we say when we hang up the phone? hi how are you? A: so then uh: < laughter > so that's going well or not? < /laughter >
5. MARKING-FUNNINESS (**film script**) uh:: oh < laughter > it is something < /laughter > uh < laughter >that happened< /laughter > to a buddy < /laughter >it is< /laughter > in fact, his chick and one of our buddies were playing (and + and) and playing they splashed some ice tea on him and we thought that he had pissed himself.

In what follows, we attempt to validate this account on the basis of a cross linguistic corpus study. We then sketch a formal theory that combines the various dimensions, stimulus, arousal, and function.

4 Material and Method

4.1 Material (corpus)

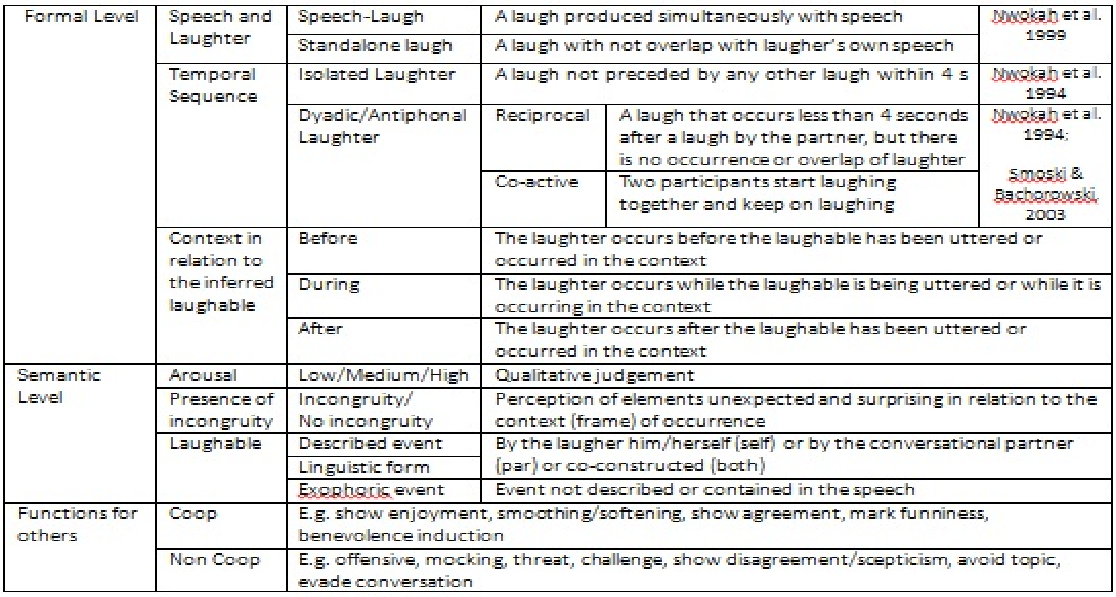
We analyzed a portion of the DUEL corpus (citation suppressed for anonymity). The corpus consists of 10 dyads/ 24 hours of natural, face-to-face, loosely task-directed dialogue in French, Mandarin Chinese and German. Each dyad conversed in three tasks which in total lasted around 45 minutes. The three tasks used were **"dream apartment"**: the participants are told that they are to share a large open-plan apartment, and will receive a large amount of money to furnish and decorate it. They discuss the layout, furnishing and decoration decisions; **"film script"**: The participants spend 15 minutes creating a scene for a film in which something embarrassing happens to the main character; and **"border control"**: one participant plays the role of a traveler attempting to pass through the border control of an imagined country, and is interviewed by an officer. The traveler has a personal situation that disfavours him/her in this interview. The officer asks questions that are general as well as specific. In addition, the traveler happens to be a parent-in-law of the officer. The corpus is transcribed in the target language and glossed in English. Disfluency, laughter, and exclamations are annotated. The current paper presents analysis of a portion of the DUEL corpus (Hough et al., 2016): two dyads both in French and Chinese (3 tasks x 2 pairs x 2 languages), having a total of 657 laughter events analysed in relation to their laughable over a total of 160mins.

4.2 *Audio-video* coding of laughter

Coding was conducted by the first and second authors: each video was observed until a laugh occurred. The coder detected the exact onset and

---

[4] The distinction between smoothing/softening on the one hand and benevolence induction on the other lies in whether the speaker is trying to induce agreement (benevolence induction), or to reduce intrusion (smoothing). A helpful way to look at this distinction is with reference to the notion of positive and negative politeness (Brown and Levinson, 1987).



Formal Level	Speech and Laughter	Speech-Laugh		A laugh produced simultaneously with speech	Nwokah et al. 1999
		Standalone laugh		A laugh with not overlap with laugher's own speech	
	Temporal Sequence	Isolated Laughter		A laugh not preceded by any other laugh within 4 s	Nwokah et al. 1994
		Dyadic/Antiphonal Laughter	Reciprocal	A laugh that occurs less than 4 seconds after a laugh by the partner, but there is no occurrence or overlap of laughter	Nwokah et al. 1994; Smoski & Bachorowski 2003
			Co-active	Two participants start laughing together and keep on laughing	
	Context in relation to the inferred laughable	Before		The laughter occurs before the laughable has been uttered or occurred in the context	
		During		The laughter occurs while the laughable is being uttered or while it is occurring in the context	
		After		The laughter occurs after the laughable has been uttered or occurred in the context	
Semantic Level	Arousal	Low/Medium/High		Qualitative judgement	
	Presence of incongruity	Incongruity/ No incongruity		Perception of elements unexpected and surprising in relation to the context (frame) of occurrence	
	Laughable	Described event		By the laugher him/herself (self) or by the conversational partner (par) or co-constructed (both)	
		Linguistic form			
		Exophoric event		Event not described or contained in the speech	
Functions for others	Coop	E.g. show enjoyment, smoothing/softening, show agreement, mark funniness, benevolence induction			
	Non Coop	E.g. offensive, mocking, threat, challenge, show disagreement/scepticism, avoid topic, evade conversation			

Figure 3: Laughter coding parameters

offset in Praat, and conducted a multi-layer analysis previously illustrated. Reliability was assessed by having a Masters student as a second coder for 10% of the material observed. Percentage agreements between the two coders for french data averaged 86.6%, with an overall Krippendorff α (Krippendorff, 2012) across all tiers of 0.652. The value is very negatively affected by the layer regarding the presence or absence of incongruity where one of the coders almost never coded a situation where no incongruity was perceived and the almost absence of one value is "strongly punished" by α. The discrepancy could also be accounted for by errors due to the coder. When excluding that tier α is 0.706. For the Chinese data, the percentage of agreement across all tiers averaged 90.5% with αbeing 0.752. In the Chinese coding the factor more responsible for the discrepancy observed is arousal. Acknowledging the very subjective measure that we are at the moment relying on i.e., personal perceptual judgment, we plan to use more objective acoustic and behavioural measures in future investigations.

Identification of a laughter episode A laugh was identified using the same criteria as (Nwokah et al., 1994), based on the facial expression and vocalization descriptions of laughter elaborated by (Apte, 1985) and (Ekman and Friesen, 1975). Following (Urbain and Dutoit, 2011) we counted laughter offset (final laughter in-breath inhalation) as part of the laughter event itself, thus resulting in laughter timing longer than other authors (Bachorowski and Owren, 2001; Rothgänger et al., 1998). All laughter events were categorised according to different parameters: formal and contextual aspects, semantic meaning and functions. Coding criteria were elaborated in order to capture the difference, stressed in previous sections, between form, meaning, and functions of laughter production in dialogical interaction (Table 1). In the current study we restrict our observations about the aspects pertaining to form to the contextual distribution and positioning of a laugh in relation to others' laughter, the laughable and laugher's own speech[5].

5 Results

5.1 General patterns

5.1.1 Frequency and duration

Laughter was in general very frequent. In the French data, there were 430 laughter events (lasting a total of 13.3 minutes) in 77 minutes of dialogue, giving a frequency of 56 laughter events per 10 minutes or 17% of the time. In the Chinese data, there were 215 laughter events (lasting a total of 6 minutes) in 85 minutes of dialogue, giving a frequency of 26 laughter events per 10 minutes or 7.2% of the time. A Z-test on the proportion of laughter minutes shows that laughter is marginally more frequent in French than in Chinese (z=1.9, p=0.05). Whether this is a language/cultural difference or an inter-subject one will be tested in the future with more data. There were higher propor-

---

[5] Hypothesis and discussion of data about different behaviour across tasks is deferred to a future study when a larger set of data will be available

tions of speech-laughter in Chinese (47%) than in French (33%), χ^2= 4.9, p= 0.03.

5.1.2 Dyadic laughter

The distributions of isolated, reactive and coactive laughter do not differ across tasks. There is more antiphonal laughter in French than in Chinese. Collapsing reactive and coactive laughters into antiphonal/dyadic laughters, these account for 44% of all laughter events in French and 36% in Chinese, showing that participants frequently join in in another's laughter. The mean of transitional probability of antiphonal laughter in relation to the participant laughter behaviour is very similar between languages (fr: 43.5% sd 5.5, ch: 42.75% sd 24.97).

Type	Ch.no.	Ch.%	Fr.no.	Fr.%
Reactive	38	18%	107	25%
Coactive	39	18%	80	19%
Total Antiphonal	77	36%	187	44%
others	138	64%	241	56%

Table 1: Percentage of antiphonal laughter

5.2 Laughable and relative position of laughter

5.2.1 Laughable

The distribution of laughable is nearly identical in Chinese and French, with half being a self described event, and around 40% being an event described by the partner, or jointly described by both participants. Around 10% are exophoric and there were very few laughs that were only about the linguistic form or content. The task did not make a significant difference to the distribution.

laughable	Ch no.	Ch %	Fr no.	Fr %
de_self	118	55%	221	52%
de_par	67	31%	160	37%
de_both	7	3%	13	3%
ex	21	10%	31	7%
ling	2	1%	3	1%

Table 2: Laughable types distribution

5.2.2 Laughter-laughable alignment

Laughter can occur before (cataphoric), during or after (anaphoric) the laughable (see Table 3). Unlike lexical anaphora, laughs sometimes occur at the same time as the laughable. As illustrated in Figure 1, we found that there are big variations in the alignment of laughter and laughable. Some laughter events span from before the laughable until after the laughable. Some laughter events are more than one utterance away from the laughable[6].

In relation to laughables, when the laughter occurs after the laughable, there are equal numbers of self described events and other described events. When the laughter occurs during or before the laughter, there are more self described events than other described events.

Ch							
Cntxt	de_par	de_slf	de_both	ex	ling	Ttl	\%
aft	49	55	4	6	2	116	57%
dur	15	49	2	13		79	39%
bef	1	5	0	2		8	4%
Fr							
aft	145	143	6	13	3	310	74%
dur	13	66	4	16		99	24%
bef	0	8	1	1		10	2%

Table 3: Position of laughter in relation to laughable

5.3 Meaning and function: arousal, presence of incongruity and function

5.3.1 Perceived arousal

The majority of the laughters had low arousal in both languages. High arousal laughters were rare. Task did make a difference. Laughters in the more serious border control task were 100% low arousal in French and 85% low arousal in Chinese. Arousal correlates with laughter duration: mean(low)= 1.11s, mean(mid)= 2.55s, mean(high)= 4.6s.

Arousal	Ch no.	Ch %	Fr no.	Fr %
low	165	77%	265	62%
mid	47	22%	162	38%
high	3	1%	2	0.40%

Table 4: Level of Arousal percentages

5.3.2 Presence of incongruity

The majority of the laughs were perceived to communicate an appraisal of incongruity (85% for both languages). Non-incongruity laughs were perceived to communicate ingroupness with the hearer. In Chinese, there is a higher proportion of non-incongruity laughs in the border control task, while in French the distribution was consistent across tasks.

5.3.3 Functions

The distribution of functions are surprisingly consistent between French and Chinese (see figure 4), with the most frequent being *show enjoyment*, followed by *smoothing/ softening, show agreement, mark funniness* and *benevolence induction*. Clustering analysis on all tiers shows that the latter two functions have similar distributions. Less frequent functions include self-mocking, apology, show sympathy and showing appreciation (to thank).

5.4 Interactions across tiers

We are interested in how the tiers interact with each other, and to what extent functions can be

---

[6] Due to such variability, we leave this tier out of the regression analysis and will investigate it in more detail in future studies analysing both the freedom in laughter-laughable alignment as well as its limits and constraints.

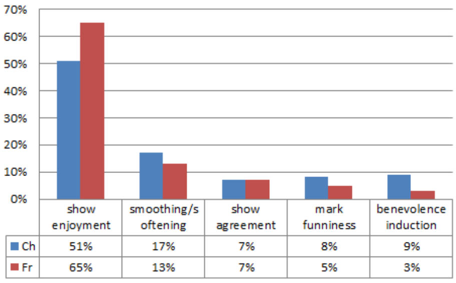


Figure 4: Function distribution

predicted by form and context tiers. Due to similar distributions we collapsed function ”smoothing/softening” with ”benevolence induction”. To study the relations among tiers, we performed multiple factor analysis, which converts a set of possibly correlated variables into a smaller set of uncorrelated variables. Figure 5 plots the correlation of each tier in relation to the two dimensions that explain the most amount of variance (x and y axes). In Chinese, antiphonal and speech laugh contribute to the same dimension, roughly independent from arousal. The type of laughable (and to a less degree incongruity) contribute to both dimensions. Function only correlates with the first dimension. In French, arousal and antiphonal contribute to the same dimension, roughly independent from speech laugh. The type of laughable and incongruity contribute to both dimensions. Therefore the main difference between the languages is that in Chinese, it is arousal which doesn’t explain the variances in function; in French, it is speech laugh.

We then performed multinomial logistic regression analysis, trying to predict the function (specifically the odds ratio of one function over another) from speech laugh, antiphonal, arousal and laughable. Figure 6 plots the distribution of functions against four tiers. In both languages, *show agreement* and *show enjoyment* are often antiphonal laughters, and they have low proportions of laughables from self. In Chinese, *mark funniness/ridiculousness* has a very distinct signature from the other functions, being almost exclusively speech laugh and having a laughable from self. In French, *mark funniness/ridiculousness* is close to *benevolence induction* apart from arousal (the former has higher arousal). Table 5 shows that in Chinese, the factors antiphonal, laughable and speech laugh have significant effects in functions, duration has a marginally significant effect (after adjusting p for multiple comparisons), and arousal doesn’t have an effect. In French, antiphonal, laughable, and arousal have significant effects in functions, while speech laugh has no significant effect.

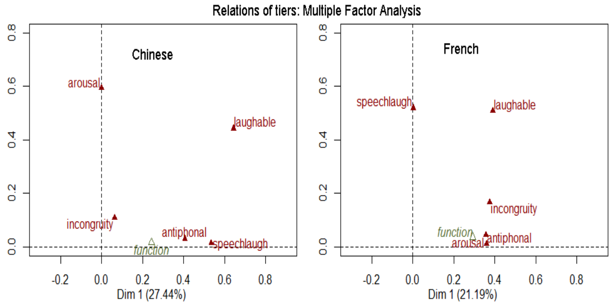


Figure 5: Relation of tiers

Tiers	value	bnvlnce/ enjoy	mrk funny/ enjoy	agree enjoy	mrk funny/ bnvlnce	agree bnvlnce	agree/ mrk funny
Chinese							
speech-laugh	coeffcnt	-.30	2.08*	.50	2.38**	.81	-1.58
	p-adjst	1	.02	1	.01	1	.57
antiphnl/ coactive	coeffcnt	-1.19*	-16.58***	.28	-15.75***	1.47	17.60***
	p-adjst	.05	.00	1	.00	.24	.00
mid/high-arousal	coeffcnt	-.39	.99	.53	1.38	.92	-.46
	p-adjst	1	1	1	.72	1	1
laughable-self	coeffcnt	.94	56.56***	-1.16	23.22***	*-2.10*	-17.98***
	p-adjst	.27	.00	1	.00	.08	.00
duration	coeffcnt	*-.72*	-.66	-.66	.06	.06	.00
	p-adjst	.08	.26	.75	1	1	1
French							
speech-laugh	coeffcnt	-.14	-.17	.12	-.03	.26	.30
	p-adjst	1	1	1	1	1	1
antiphnl coactive	coeffcnt	1.63***	1.75***	.38	.12	*-1.24*	-1.36
	p-adjst	.00	.00	1	1	.09	.16
mid/high-arousal	coeffcnt	-1.67***	-.60	-1.93***	1.07	-.26	-1.33
	p-adjst	.00	.90	.00	.29	1	.23
laughable-self	coeffcnt	1.77**	1.68*	-1.48**	-.09	-3.25***	-3.15***
	p-adjst	.00	.02	.00	1	.00	.00

Table 5: Multinomial logistic regression results: coefficients of log odds and *p* value (adjusted for multiple comparisons) comparing each pair of functions

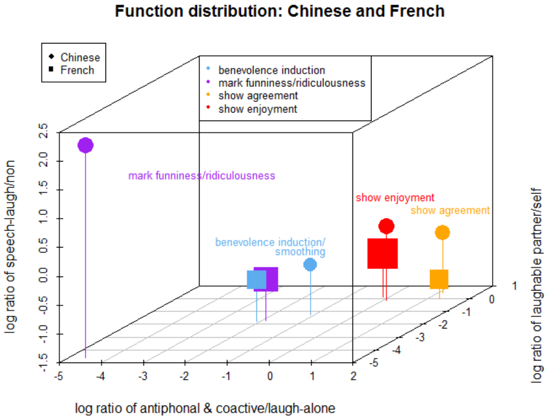


Figure 6: Distribution of functions in relation to speech laugh, dyadic laugh, laughable, arousal and duration. The x, y and z axes represent log ratios of dyadic over non-dyadic laugh, speechlaugh over non-speechlaugh and laughable from partner over self. The size of dots represents the average arousal.

6 Discussion

Our multi-layered analysis of laughter in dialogue investigated the contextual forms (frequency, duration, speech-laughter, and laughter co-occuring with partner’s laughter), laughable (type of laughable, position of laughter in relation to laughable, meaning (incongruity and arousal) and function.

Probably due to the cooperative and (to some extent) funniness oriented corpus, we found higher frequencies of laughter (in French and Chinese respectively 26 and 56 laughter bouts over 10 minutes of interaction) than reported elsewhere (e.g., (Vettin and Todt, 2004)'s 5.8 (2.5)/10min)), and within that value we also reported a higher proportion of speech laughter over stand alone laughter (40%) than previously (e.g. (Nwokah et al., 1999)'s mean of 18,6%, even though they reported a variance up to 50%.). However, duration wise, our results are similar to previous results, both for stand alone laughter (Petridis et al., 2013; Truong and Van Leeuwen, 2007; Nwokah et al., 1999; Bachorowski and Owren, 2001) and speech laugh (e.g.(Nwokah et al., 1999). We found a higher percentage both of reactive and co-active laughter compared to e.g., (Nwokah et al., 1994) (8%) and (Smoski and Bachorowski, 2003) (34%). The values are nevertheless consistent between the French and Chinese samples, having an overall mean of 43.12 transitional probability of a participant to laugh antiphonally in relation to his partner.

In terms of laughables, there are more self-described events than partner described events, suggesting that speakers laugh more than the audience. Most laughables are described events; exophoric laughables are less frequent and linguistic laughables are rare. More than half of the laughters follow the laughable, but a significant amount occur during the laughable, a few occur before the laughable. In terms of meaning, we perceived that around 85% of the laughters communicate an appraisal of incongruity, and most laughter have low arousal. In terms of the laughter's effect or function, We identified four most frequent types in our data: *show enjoyment* (most frequent), *smoothing/benevolence induction, show agreement*, and *mark funniness*. The four functions have distinct distributions in measurements from form and laughable layers. The functions seem to be characterized by a cluster of layers rather than from a single one.

7 The Varieties of Laughter: interfacing with grammar and emotional state

In this section we sketch a formal semantic and pragmatic treatment of laughter that can accommodate the results in section 5. In section 3 we pointed to certain lacunae that (Ginzburg et al., 2015) faces. We briefly sketch some solutions, leaving to a more extended version a more detailed treatment.

On the approach developed in KoS, information states comprise a private part and the dialogue gameboard that represents information arising from publicized interactions. In addition to tracking shared assumptions/visual space, Moves, and QUD, the dialogue gameboard also tracks **topoi** and **enthymemes** that conversational participants exploit during an interaction (e.g., in reasoning about rhetorical relations.)(Ginzburg et al., 2015). Here topoi represent general inferential patterns (e.g., *given two routes choose the shortest one*) represented as functions from records to record types and enthymemes are instances of topoi (e.g., *given that the route via Walnut street is shorter than the route via Alma choose Walnut street*). An enthymeme belongs to a topos if its domain type is a subtype of the domain type of the topos.

(Ginzburg et al., 2015) posit distinct, though quite similar lexical entries for enjoyment and incongruous laughter. For reasons of space in (1) we exhibit a unified entry with two distinct contents. (1) associates an enjoyment laugh with the laugher's judgement of a proposition whose situational component l is *active* as enjoyable; for incongruity, a laugh marks a proposition whose situational component l is *active* as *incongruous*, relative to the currently maximal enthymeme under discussion.

(1)

$$\left[\begin{array}{l}
\text{phon} : \texttt{laughterphontype} \\
\text{dgb-params} : \left[\begin{array}{l}
\text{spkr} : \text{Ind} \\
\text{addr} : \text{Ind} \\
\text{t} : \text{TIME} \\
\text{c1} : \text{addressing(spkr,addr,t)} \\
\text{MaxEud} = \text{e} : \text{(Rec)RecType} \\
\text{p} = \left[\begin{array}{l} \text{sit} = \text{l} \\ \text{sit-type} = \text{L} \end{array}\right] : \text{prop} \\
\text{c2} : \text{ActiveSit(l)}
\end{array}\right] \\
\text{content}_{enjoyment} = \text{Enjoy(spkr,p)} : \text{RecType} \\
\text{content}_{incongruity} = \text{Incongr(p,e,}\tau\text{)} : \text{RecType}
\end{array}\right]$$

(1) makes appeal to a notion of an *active situation*. This pertains to the accessible situational antecedents of a laughter act, given that (Ginzburg

et al., 2015) proposed viewing laughter as an eventive anaphor. However, given the significant amount of speech laughter, this notion apparently needs to be rethought somewhat, viewing laughter in gestural terms. This requires interfacing the two channels, a problem we will not address here, though see (Rieser, 2015) for a recent discussion in the context of manual gesture. Given the enjoyment meaning and the topos *If X is enjoying that X/Y said that p, then X agrees that p*, (Ginzburg et al., 2015) obtain as a consequence that enjoyment laughter can be used as a positive feedback signal. We think that this can be extended to yield also the function of *benevolence induction* via the topos *if X is enjoying Y's presence, X does not want to have a disagreement with Y.*

(Ginzburg et al., 2015) explicate incongruity in terms of a clash between the enthymeme triggered by the laughable and a topos which the enthymeme is supposed to instantiate. On the basis of this they explicate seriousness cancellation in an utterance u as (mock) self-repair. The laugher relies on the enthymeme 'If I'm saying u, then I don't mean it.' This clashes with the sincerity topos 'If A says p, then A means p'. One can extend this to smoothing in an interaction between A and B as arising from a clash between the enthymeme if A is manifestly pleasant to B, A need not wish to be overly intimate with B and the topos if an individual X is manifestly pleasant to Y, X wants to be open to Y.

The dialogue gameboard parameters utilised in the account of (Ginzburg et al., 2015) are all 'informational' or utterance related ones. However, in order to deal with notions such as arousal and psychological shift, one needs to introduce also parameters that track appraisal (see e.g., (Scherer, 2009)). For current purposes, we mention merely one such parameter we dub *pleasantness* that relates to the appraisal issue—in Scherer's formulation—*Is the event intrinsically pleasant or unpleasant?*. We assume this parameter is scalar in value, with positive and negative values corresponding to varying degrees of pleasantness or unpleasantness.

This enables us to formulate conversational rules of the form 'if A laughs and pleasantness is set to k, then reset pleasantness to k + $\theta(\alpha)$', where α is a parameter corresponding to arousal. We provide a more precise formulation in an extended version of this paper.

8 Conclusions and Further Work

This paper presents a multi-layered classification of laughter based on a detailed corpus study of French and Chinese dialogues taken from the DUEL corpus. Data from the form/context layers show that laughter can occur before, during or after the laughable, which can be a described event, an exophoric event, or a metalinguistic stimuli. The freedom in time alignment between laughter and laughable demonstrates that analyzing what precedes laughter on the surface is unreliable as a means for determining what laughter is about. Data from the meaning layer show that in our corpus, laughter, with varying degrees of arousal, can communicate an appraisal of incongruity, the enjoyment of an event, or the feeling of ingroupness with the partner. The simple meaning of laughter, when combined with rich contextual reasoning, can have various effects or functions in interaction. The most frequent ones in our corpus are *show enjoyment, smoothing/benevolence induction, mark funniness and show agreement*. These types are not distinguishable by any *single* form or context layer measurement, but rather by a cluster of them (for example, benevolence induction and smoothing laughters are mostly stand-alone, low arousal laughter, when the laughable is partner produced). Cross-linguistically, the distributions of most layers of analysis are very similar between French and Chinese, suggesting tentatively that laughter is not heavily shaped by linguistic features. Based on our data, we ground the analysis in a formal framework, treating laughter as gestural event anaphora, and proposing the incorporation of emotional appraisal into the dialogue gameboard.

There is much further work to be done on all fronts addressed here. This includes a more accurate analysis of acoustic features, and those pertaining to laughables; on the formal front further integration of information state dialogue analysis with appraisal models coming from cognitive psychology and AI.

Acknowledgments

We would like to thank three anonymous reviewers for JerSem for their very helpful comments. We acknowledge the support of the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) and the Disfluency, Exclamations, and Laughter in Dialogue (DUEL) project within the *projets franco-allemand en sciences humaines et sociales* funded by the ANR and the DFG.

References

Mahadev L Apte. 1985. *Humor and laughter: An anthropological approach*. Cornell Univ Pr.

Jo-Anne Bachorowski and Michael J Owren. 2001. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science*, 12(3):252–257.

Ellen Breitholtz. 2014. Reasoning with topoi–towards a rhetorical approach to non-monotonicity. In *Proceedings of the 50th anniversary convention of the AISB, 1st–4th April 2014, Goldsmiths, University of London*.

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

Nick Campbell, Hideki Kashioka, and Ryo Ohara. 2005. No laughing matter. In *Ninth European Conference on Speech Communication and Technology*.

Jennifer Coates. 2007. Talk in a play frame: More on laughter and intimacy. *Journal of Pragmatics*, 39(1):29–49.

Paul Ekman and Wallace V Friesen. 1975. Unmasking the face: A guide to recognizing emotions from facial cues.

Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*, University of Amsterdam.

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

Haruko Hayakawa. 2003. The meaningless laughter: Laughter in japanese communication. *Unpublished PhD Thesis, University of Sydney, Australia*.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, 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*.

Silke Kipper and Dietmar Todt. 2003. The role of rhythm and pitch in the evaluation of human laughter. *Journal of Nonverbal Behavior*, 27(4):255–272.

Klaus J Kohler. 2008. speech-smile,speech-laugh,laughterand their sequencing in dialogic interaction. *Phonetica*, 65(1-2):1–18.

Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

Caroline Menezes and Yosuke Igarashi. 2006. The speech laugh spectrum. *Proc. Speech Production, Brazil*.

John Morreall. 1983. *Taking laughter seriously*. SUNY Press.

Evangeline Nwokah and Alan Fogel. 1993. Laughter in mother-infant emotional communication. *Humor: International Journal of Humor Research*.

Evangeline E Nwokah, Hui-Chin Hsu, Olga Dobrowolska, and Alan Fogel. 1994. The development of laughter in mother-infant communication: Timing parameters and temporal sequences. *Infant Behavior and Development*, 17(1):23–35.

Eva E Nwokah, Hui-Chin Hsu, Patricia Davies, and Alan Fogel. 1999. The integration of laughter and speech in vocal communicationa dynamic systems perspective. *Journal of Speech, Language, and Hearing Research*, 42(4):880–894.

Michael J Owren, Drew Rendall, and Jo-Anne Bachorowski. 2003. Nonlinguistic vocal communication. *Primate psychology*, pages 359–394.

Stavros Petridis, Brais Martinez, and Maja Pantic. 2013. The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202.

Fernando Poyatos. 1993. The many voices of laughter: A new audible-visual paralinguistic approach. *Semiotica*, 93(1-2):61–82.

Robert R Provine and Yvonne L Yong. 1991. Laughter: A stereotyped human vocalization. *Ethology*, 89(2):115–124.

R. R. Provine. 1993. Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology*, 95(4):291–298.

Robert R Provine. 2004. Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science*, 13(6):215–218.

V. Raskin. 1985. *Semantic mechanisms of humor*, volume 24. Springer.

Boris Reuderink, Mannes Poel, Khiet Truong, Ronald Poppe, and Maja Pantic. 2008. *Decision-level fusion for audio-visual laughter detection*. Springer.

Hannes Rieser. 2015. When hands talk to mouth. gesture and speech as autonomous communicating processes. *SEMDIAL 2015 goDIAL*, page 122.

Hartmut Rothgänger, Gertrud Hauser, Aldo Carlo Cappellini, and Assunta Guidotti. 1998. Analysis of laughter and speech sounds in italian and german students. *Naturwissenschaften*, 85(8):394–402.

Willibald Ruch and Paul Ekman. 2001. The expressive pattern of laughter. *Emotion, qualia, and consciousness*, pages 426–443.

Klaus R Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351.

A Shimizu, N Sumitsuji, and M Nakamura. 1994. Why do people laugh.

Moria Smoski and Jo-Anne Bachorowski. 2003. Antiphonal laughter between friends and strangers. *Cognition & Emotion*, 17(2):327–340.

Diana P Szameitat, Kai Alter, André J Szameitat, Dirk Wildgruber, Annette Sterr, and Chris J Darwin. 2009. Acoustic profiles of distinct emotional expressions in laughter. *The Journal of the Acoustical Society of America*, 126(1):354–366.

Hiroki Tanaka and Nick Campbell. 2014. Classification of social laughter in natural conversational speech. *Computer Speech & Language*, 28(1):314–325.

Jürgen Trouvain. 2001. Phonetic aspects of speech-laughs. In *Proceedings of the 2nd Conference on Orality and Gestuality*, pages 634–639.

Jürgen Trouvain. 2003. Segmenting phonetic units in laughter. In *Proc. 15th International Conference of the Phonetic Sciences, Barcelona, Spain*, pages 2793–2796.

Khiet P Truong and David A Van Leeuwen. 2007. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158.

Jérôme Urbain and Thierry Dutoit. 2011. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In *Affective Computing and Intelligent Interaction*, pages 397–406. Springer.

Julia Vettin and Dietmar Todt. 2004. Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115.

Julia Vettin and Dietmar Todt. 2005. Human laughter, social play, and play vocalizations of non-human primates: An evolutionary approach. *Behaviour*, 142(2):217–240.

Extralinguistic State Localization in Service of Turn Generation in Task-Oriented Dialogue

Petr Babkin
Cognitive Science Department / 110 8th St
Rensselaer Polytechnic Institute / Troy, NY 12180
babkip@rpi.edu

Abstract

The tremendous role of context in understanding natural language dialogue has been amply emphasized in the literature. Alas, in much research to date, context is defined simply as preceding linguistic material within some window. In real life, however, linguistic content amounts to only a fraction of contextual information that helps humans to act appropriately in a conversation. In fact, in some cases it is non-linguistic cues that are most informative e.g., in certain stereotypical situations such as the famous restaurant script. This study, explores the notion of context as latent extralinguistic state underlying task-oriented dialogue. This view is put to test of deriving a coherent task-relevant dialogue turns in the face of arbitrarily ablated input. The paper outlines the approach to be presented as a poster along with preliminary results.

1 Introduction

It is no secret that a better informed decision is bound to lead to a better outcome. When it comes to natural language processing, effective feature engineering is often attributed much success, in many cases, offsetting the merit of the actual algorithms that utilize them. With the growing complexity of NLP tasks, the heuristics, too, become more sophisticated, capturing the decision's increasing dependence on the context in which the input is observed. Alas, in much of NLP research, context is limited to features that are directly available from linguistic input. Such reliance on a single source of heuristics assumes error-free input, which is not always the case[1]. This study explores the capabilities of extra-textual heuristics by artificially encouraging the exploitation of pragmatic context over the observed input, through the use of ablation. The hypothesis is the more ablated the input[2] — the more the system has to rely on pragmatic reasoning to compensate for the deficiency. In other words, a good sense of the situation may enable one to come up with a correct answer without necessarily understanding the question.

2 Domain and representation

Task-oriented dialogue appears to be a promising domain, being a rich source of goal-based heuristics that could support pragmatic reasoning. Specifically, the Cards corpus, with its clear goal structure and highly goal-oriented linguistic content, appears well suited for modeling of this sort (Potts, 2012). Analysis of a sample of dialogues from the corpus revealed that each speech act is not simply conditioned on its preceding utterances *per se* but also depends upon a) a persistent extralinguistic state that is maintained via speech acts, and b) goal-directed implications of this state. Therefore, the problem of response generation is preconditioned on the following two subproblems:

- inference of the current extralinguistic state from the observed language input,
- selection of the desirable state and generation of the state-inducing linguistic output.

In order to capture the goal dynamics underlying language communication (herein hypothesized to be necessary for handling imperfect input), a variant of the state-space representation was used with two modifications. First, the standard definition of states as sets of domain-specific predicates and truth values was replaced with a more abstract notion of states of common ground (CG)

---

[1] Error propagation is one good analogy of a system's fragility because of the unrealistic expectations about the quality of upstream information (Caselli and Postma, 2015).

[2] To isolate the effects of ablation and bypass issues unrelated to it, semantic meaning representations were used as input rather than raw text.

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 108-109.

(Clark, 2006) — points in time when certain facts become shared knowledge (e.g., through verbalization by one of the agents). Second, domain-specific actions are assumed latent and state transitions are modeled solely as knowledge dependencies among the states. The choice of this proxy representation can be justified by a) the primary need for heuristics to aid language understanding rather than to help actually solve an associated planning problem and b) such a representation can be derived from the language material irrespective of the subject domain[3].

3 Model

For the reasons of space, much detail is omitted from the model description and the purpose of this section is limited to providing a high-level overview.

As noted in the previous section, the agent's choice of a response depends not directly on the input speech but on the unobservable state, which, in turn, is conditioned on both the observed input and the previous state:

$$o_{t+1} \leftarrow f(s_t | o_t, s_{t-1}) \tag{1}$$

This naturally brings us to hidden state models, such as HMM, where emitted states correspond to observed utterances and the hidden states are the underlying situations[4]. Such a model of course needs to be extended for there is not a one-to-one correspondence among observable and hidden states. For example, in Figure 1, the predicate $cur\text{-}cards(a,\ 5s)$[5] is added to the common ground as a result of an exchange between the conversants rather than a single utterance. While this compositionality could be modeled as a joint distribution, it appears reasonable to employ a feature-based grammar instead. A binary result returned by the corresponding recognizer effectively replaces the coefficient from the emission matrix in the state probability equation. In order to account for ablation, this value also needs to be weighed based on the likelihood of the ablation instance. The data flow in the model is summarized by the following stages.

$$\{cur\text{-}cards(a,\ 5s)\} \left\langle \begin{array}{l} \dfrac{B\text{: } you\ have\ 5\ spades}{\mathbf{suggest}(b,\ cur-cards(a,\ 5s))} \\ \\ \dfrac{A\text{: } yep}{\mathbf{acknowledge}(a)} \end{array} \right.$$

Figure 1: A fact grounding adjacency pair.

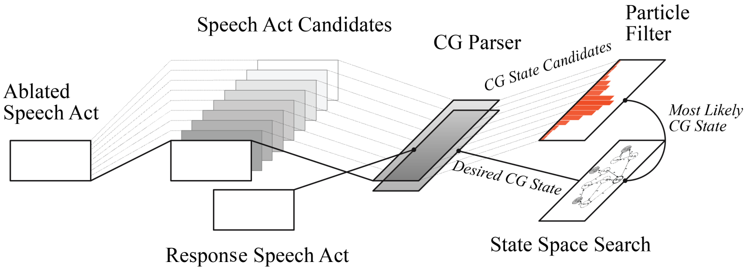


Figure 2: Model architecture.

1. For the observed speech act, alternatives increasingly dissimilar to the original are generated by enumerating feature values up until a set cutoff probability threshold.
2. The candidates are then mapped to their corresponding states (if any) by the CG parser/recognizer.
3. The resulting CG states along with their weights are passed to the particle filter, which outputs the belief distribution over CG states.
4. The most likely current CG state is used to compute the desirable CG state via breadth-first search.
5. The grammar is used again to induce a speech act that would expand the desired CG state.

References

Tommaso Caselli and Marten Postma. 2015. When it's all piling up: investigating error propagation in an NLP pipeline. In *WNACP 2015*, Passau, Germany.

Herbert Clark. 2006. Context and Common Ground. *Concise Encyclopedia of Philosophy of Language*, pages 105–108.

Niels Kasch and Tim Oates. 2010. Mining script-like structures from the web. In *Proceedings of the First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 34–42, Stroudsburg, PA.

Pat Langley, Ben Meadows, Alfredo Gabaldon, and Richard Heald. 2014. Abductive understanding of dialogues about joint activities. *Interaction Studies*, 15(3):426–454.

Christopher Potts. 2012. Goal-driven answers in the Cards dialogue corpus. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, Somerville, MA.

---

[3] which in turn opens up an exciting possibility of generating models of novel domains automatically cf. (Kasch and Oates, 2010)

[4] It is important to note that hidden states in this case are not true values of observed inputs (emissions) common for noisy channel model, but extralinguistic states comprised of domain predicates.

[5] For the sake of brevity, the notation as in (Langley et al., 2014) is used for speech acts and domain predicates.

Structural focus and discourse structure in Hungarian narratives

Kata Balogh
Heinrich-Heine-Universität Düsseldorf
`katalin.balogh@hhu.de`

Abstract

In my presentation I investigate and exemplify the discourse-related triggers of the preverbal focus position in Hungarian narratives. Inspired by the approaches of Riester (2015; 2016), Büring (2003) and Roberts (2012), I propose a QUD-based analysis and explanation of the licensing condition of the Hungarian preverbal focus position in narratives.

The research presented here is part of a larger project[1] investigating and modeling the interaction of morphosyntax and the conceptual background of information structure [InfS], the local common ground [CG] or discourse context. The project investigates the influence of InfS on morphosyntax from a cross-linguistic perspective in three unrelated, non-configurational languages: Tagalog, Hungarian and Lakhota.

InfS in Hungarian

In non-configurational languages, information packaging considerations often determine the choice of a marked morpho-syntactic structure. The choice of a marked construction signals a certain informational structural organization that is only felicitous if it is licensed by the given discourse context (local CG), the shared knowledge at a given point of the discourse (dialogue or narration). Languages differ in how far morphosyntactic structure is influenced by information structure and CG considerations. A rigid syntax language like English does not have syntactic means to signal the narrow focus or the aboutness topic. The discourse-configurational language (É. Kiss, 1995), Hungarian, however, reflects information structure at the syntactic level.

Hungarian shows verb-initial word order in the unmarked case (1), and has special structural positions for the sentence topic and the narrow focus of the utterance; topics being sentence initial, while the narrow focus standing in the immediate preverbal position (2).

(1) Meg-látogatta Péter Mari-t.
prt$_{meg}$-visited Peter Mary-ACC
'Peter visited Mary.'

(2) Péter MARI-T látogatta meg.
Peter Mary-ACC visited prt$_{meg}$
'Peter visited [Mary]^F.'

The structural focus position in Hungarian is often analyzed in semantic terms, featuring the phenomena of identification, predication and exhaustivity (e.g. É.Kiss 2006, Szabolcsi 1994). In my presentation I propose a pragmatic approach (see also e.g. Wedgwood 2007), in line with, e.g., Riester (2016; 2015), Vellema & Beaver (2015) and Roberts (2012).

Drawing on various corpus data[2] exemplify discourse-related triggers and licensing conditions of the preverbal focus position in Hungarian narratives, and propose a QUD-based analysis and explanation of these licensing conditions.

The focus position

As Riester (2015) points out, narratives are less expected to provide the basis of exploring information structural phenomena, since narratives are often structured along a temporal line. However, the corpus data from different narratives show interesting uses of the preverbal focus position, and provide a good basis to investigate triggers of focusing in terms of discourse structure. Consider, e.g., the following utterance in the given context:

---

[1] CRC 991 Project D04: The Role of Information Structure in Sentence Formation and Interpretation: A Frame-based Approach; PIs: Prof.Dr. Robert Van Valin, Jr. (working on Lakhota) and Dr. Anja Latrouite (working on Tagalog).

[2] Translations of the the Hunger Games books and self elicited data from recordings of the Frog Stories (by Mercer Mayer).

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 110-111.

(3) *When we are ready at the market, we go to the back door of the mayor's house to sell the half of the strawberry. We know how much he likes it and he gives the price we ask.*

Madge,	a	polgármester	lánya	nyit	ajtót.
Madge	the	mayor	daughter.POSS	opens	door.ACC

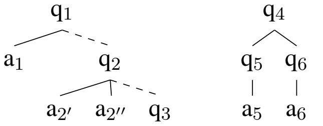
'[Madge, the mayor's daughter]^{FOC} opens the door.'

This example shows one instance where unexpectedness play a role in licensing the focus construction. From the local CG the expectation is that the mayor opens the door, and the focused constituent expresses unexpectedness. The underlying QUD is the constituent question *Who opened the door?* licensed by the global CG: the previous sentence introduces a selling situation, from which opening the door (the background) is inferable.

Analysis and goals

In my analysis I take focus as a pragmatic notion, (see e.g. Roberts, 2012; Vellema and Beaver, 2015), being an answer of the current Question Under Discussion. According to this view, the function of focus is to help determine the current QUD. The syntactically marked narrow focus construction in Hungarian determines the actual QUD being the corresponding wh-question. This current question (and thus indirectly the focus construction) must be licensed by the underlying context. Two aspects of the CG are both relevant for the licensing conditions: (1) the local discourse context (local CG) and the situational context (global CG) or background knowledge.

The local discourse context is structured and represented as a discourse-tree (d-tree) extended by an annotation schema for indicating the *focus structure* (focus, focus domain, (not-)at-issue content, aboutness topic), as well as the *thematic structure* (discourse topics). In my analysis I adopt the static d-trees from Büring (2003) and Riester (2016). The nodes in the d-tree represent the discourse moves: internal nodes represent the QUDs while the terminal nods indicate the answers. The structure of the d-tree is given by increasingly specific questions, the sub-question relation has no strict entailment relation to the preceding QUD. Sub-questions are either entailed by a previous question (e.g. q$_5$ and q$_6$), or dependent on the immediately preceding answer (e.g. q$_2$ and q$_3$).



Riester (2016) claims that the QUD-structure of the discourse is driven by multiple constraints, like: (i) there must be congruence between the actual QUD and its answer and (ii) the implicit QUD must be maximally given (or salient).

As narrow focus in Hungarian indicates its immediate QUD, licensing the use of the marked syntactic construction of the preverbal focus position is on the one hand determines by the licensing on the current QUD in the given discourse context. However, licensing the focus position is also influenced by different means, like, e.g. unexpectedness. In my analysis I also investigate what aspects besides the QUD-structure license the preverbal focus position. The following issues will be explored: (i) whether the focus constituent always contains new information, (ii) whether it is an element of a contrast set, (iii) what are the requirements for the background: pre-mentioned, presupposed or expected (conventionally or situationally inferable), and (iv) whether the focused constituent serves as the newly introduced topic.

References

Daniel Büring. 2003. On D-trees, Bfans and B-accents. *Linguistics & Philosophy*, 26(5).

Katalin É. Kiss, editor. 1995. *Discourse Configurational Languages*. Oxford University Press.

Katalin É. Kiss. 2006. Focussing as predication. In V. Molnar and S. Winkler, editors, *The Architecture of Focus*. Mouton de Gruyter, Berlin.

Arndt Riester. 2015. Analyzing questions under discussion and information structure in a Balinese narrative. In *Proceedings of the Second International Workshop on Information Structure of Austronesian Languages*, ILCAA, Tokyo University.

Arndt Riester. 2016. Constructing QUD trees. manuscript.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6).

Anna Szabolcsi. 1994. All quantifiers are not equal: The case of focus. *Acta Linguistica Hungarica*, 42.

Leah Vellema and David Beaver. 2015. Question-based models of information structure. In C. Féry and S. Ishihara, editors, *The Oxford Handbook of Information Structure*. Oxford University Press.

Daniel Wedgwood. 2007. Identifying inferences in focus. In K. Schwabe and S. Winkler, editors, *On Information Structure, Meaning and Form*. John Benjamins, Amsterdam/Philadelphia.

Are you mocking me or are you laughing with me?

Ellen Breitholtz
Dept of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
`ellen.breitholtz@gu.se`

Kristina Lundholm Fors
Dept of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
`kristina.lundholm@gu.se`

Abstract

This pilot study explores the influence of a set of semantic-pragmatic and phonetic acoustic parameters on the perception of laughter. The results suggest that voiced and unvoiced laughter are associated with different types of situations. There are also indications that the perceived meaning of laughter can be modified by modification of the context in which the laughter appears.

1 Introduction

Laughter does not necessarily equate to joy, since laughter can be used to express a range of emotions. When someone laughs we can immediately tell whether the person laughing does so kindly or maliciously. However, it is not clear to what extent this judgement is based on the actual sound of the laughter and to what extent it is based on the context in which the laughter occurs. In this paper we will describe and investigate how the phonetic-acoustic properties and the pragmatic context of laughter influence how a laughter event is perceived. Specifically, we are interested in the type of laughter called ”hånskratt” in Swedish, which may be translated as mocking or jeering laughter – what makes us perceive a laugh as mocking? In this paper we present the result of a pilot study where subjects were asked to match laughter of various phonetic-acoustic quality to various situations where laughter would be expected.

2 Background

There are several subtypes of laughter, such as song-like laughter, snort-like laughter and voiced and unvoiced laughter (Bachorowski et al., 2001). It has been shown that humans are adept at distinguishing between positive and negative laughter (Devillers and Vidrascu, 2007), and also that voiced laughter elicits much more positive emotions than unvoiced laughter (Bachorowski and Owren, 2001). Thus we may hypothesise that the voicing is a factor in determining whether a laughter event is mocking or not. Many studies have investigated various aspects of the pragmatic function of laughter, for example OConnell and Kowal (2005), Holmes (2006), and Adelswärd (1989). These studies show that laughter not only expresses joy, but also performs other social and communicative functions. Fewer studies have been carried out that focus on a precise analysis of the semantic contribution a dialogue participant makes by laughing, and how the perceived meaning is affected by particular contextual parameters. Recently there has been some work into this issue: for example Ginzburg et al. (2014) aim at creating formal models capable of accounting for laughter and laughterful utterances, and there is also recent work on the semantics of other types of non-verbal dialogue contributions that is relevant to this.

3 Aim and hypothesis

Our aim is to find out how a set of semantic-pragmatic and phonetic-acoustic parameters affect the perception of laughter, and ultimately to integrate these parameters in a semantic model of dialogue. We are particularly interested in which features are characteristic of mocking laughter, and whether semantic-pragmatic features or phonetic-acoustic features have the strongest influence on the perception of an instance of laughter as mocking.

4 Method

Samples of spontaneous laughter were obtained by letting two subjects read jokes to each other. They were also given three hypothetical scenarios which could elicit mocking laughter, and asked how they would laugh in that situation. Finally, the subjects

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 112-113.

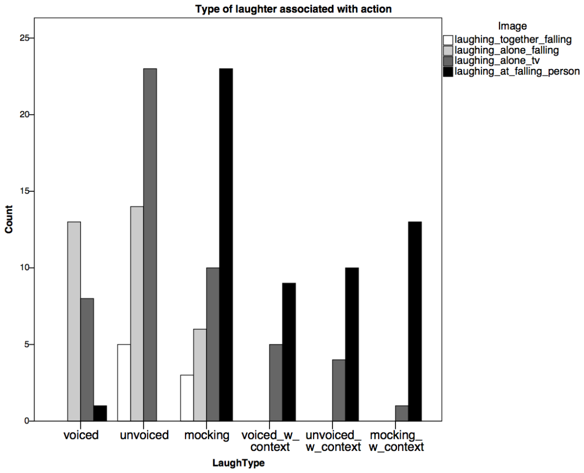
were also explicitly asked to give an example (by imitation) of their concept of mocking laughter.

Out of the samples collected, 9 were used in the study: 3 voiced samples, 3 unvoiced samples, and three samples of what was given as an example of mocking laughter. The laughter samples were all taken from the same speaker, who is a male 34 year old native speaker of Swedish. In addition, the samples were paired with the same speaker uttering ”de va rätt åt dig” (”that serves you right”). This speech sample was uttered spontaneously by the speakers during the discussion about mocking laughter.

Four two-part image series were created in Adobe Illustrator, depicting stick figures in the following situations: 1) one person falling over, and laughing about it, 2) two persons falling over, and laughing about it, 3) one person watching a clown on tv, and laughing, and 4) one person falling over, and another person pointing and laughing at that person, while the person that fell over looks sad.

With each set of images, the participants – two females and five males, all native speakers of Swedish – heard a laughter and were told to click on the image that they thought best corresponded with the laughter. The laughter sound files were presented in a randomised order, using the experiment software PsychoPy (Peirce, 2007). Each subject heard each sound file twice.

5 Results



Fisher’s exact test showed that laughter types differ significantly by image (p <0.001). Voiced laughter is primarily associated with the two persons laughing together after falling over, while unvoiced laughter is primarily associated with the person laughing alone at something funny on tv. Mocking laughter is linked to the person laughing at someone falling over. Adding the utterance ”de va rätt åt dig” (”that serves you right”) lead to subjects associating the voiced and unvoiced laughter samples with the image of the person being laughed at for falling over.

6 Discussion

The results of this pilot study suggest that voiced and unvoiced laughter are associated with different types of laughter-inducing situations. However, unvoiced laughter was not identified primarily as mocking laughter, but rather seems to be perceived as the laughter of someone who is laughing by them-self. Further, our results indicate that the perceived meaning of laughter can be modified by the context in which the laughter appears.

References

Viveka Adelswärd. 1989. Laughter and dialogue: The social significance of laughter in institutional discourse. *Nordic Journal of Linguistics*, 12(02):107–136.

Jo-Anne Bachorowski and Michael J Owren. 2001. Not All Laughs are Alike: Voiced but Not Unvoiced Laughter Readily Elicits Positive Affect. *Psychological Science*, (3):252–257.

Jo-Anne Bachorowski, Moria J Smoski, and Michael J Owren. 2001. The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, 110(3):1581.

Laurence Devillers and Laurence Vidrascu. 2007. Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Interdisciplinary Workshop on The Phonetics of Laughter*, page 37.

Jonathan Ginzburg, Ye Tian, Pascal Amsili, Claire Beyssade, Barbera Hemforth, Yannick Mathieu, Claire Saillard, Julian Hough, Spyridon Kousidis, and David Schlangen. 2014. The disfluency, exclamation and laughter in dialogue (duel) project. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt), Posters*.

Janet Holmes. 2006. Sharing a laugh: Pragmatic aspects of humor and gender in the workplace. *Journal of Pragmatics*, 38(1):26–50.

Daniel C OConnell and Sabine Kowal. 2005. Laughter in Bill Clintons My life (2004) interviews. *Pragmatics*, 15(2/3):275–299.

Jonathan W Peirce. 2007. PsychoPy: Psychophysics software in python. *Journal of neuroscience methods*, 162(1):8–13.

PentoRob: A Puzzle-Playing Robot for Dialogue Experiments

Julian Hough and David Schlangen
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literature
Bielefeld University
`firstname.lastname@uni-bielefeld.de`

Abstract

We present a simple puzzle-playing interactive robot, PentoRob, which allows investigation into real-time, real-world dialogue. The dialogue control framework consists of a combination of interactive Harel statecharts and the Incremental Unit framework. We outline its architecture and potential use cases for dialogue and human-robot interaction.

1 Introduction

In embodied dialogue systems research, there is a need for simple robots that do not require heavy mechanical maintenance or robotics experts when developing functionality of interest. Here we present a system to fulfil these needs: *PentoRob*, a simple pick-and-place robot controlled by an incremental dialogue framework.

2 PentoRob

PentoRob is a puzzle-playing robot which manipulates Pentomino pieces– see Fig. 1. Its dialogue control consists of Harel statecharts (Harel, 1987) and the Incremental Unit framework (Schlangen and Skantze, 2011), and is implemented with the dialogue toolkit InproTK (Baumann and Schlangen, 2012). Here we describe its components in terms of input information or Incremental Units (IUs), processing, and output IUs.

Hardware For the robotic arm, we use the ShapeOko2,[1] a heavy-duty 3-axis CNC machine, which we modified with a rotatable electromagnet, whereby its movement and magnetic field is controlled via two Arduino boards. The sensors are a webcam and microphone.

Incremental Speech Recognizer (ASR) We use Google's web-based ASR API which packages hypotheses into individual *WordIUs*. While its incremental performance is not as responsive as more inherently incremental local systems such as Kaldi or Sphinx-4, this does not incur great costs for many interesting applications.

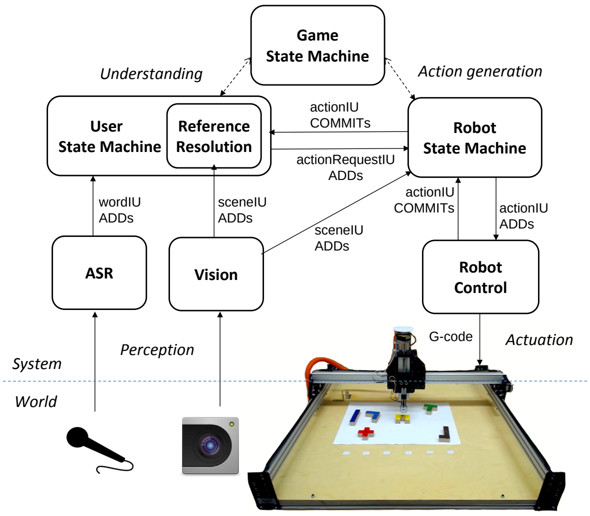


Figure 1: PentoRob's architecture.

Computer Vision (CV) We use OpenCV in a Python module to track objects in the camera's view. This information is relayed to InproTK from Python via the Robotics Service Bus (RSB),[2] which outputs IDs and positions of objects detected along with their low-level features (e.g., RGB/HSV values, x,y coordinates, number of edges, etc.), converting these into $SceneIU$s which the reference resolution module consumes and the *Robot State Machine* uses for obtaining the positions of objects it plans to grab.

Reference resolution (WAC) The reference resolution component consists of a Words As Classifiers (WAC) model (Kennington and Schlangen, 2015) trained on real-world objects using low-level vision features from $SceneIU$s.

---

[1] `http://www.shapeoko.com`.

[2] `https://code.cor-lab.de/projects/rsb`.

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,* edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA, July 16-18, 2016, pages 114-115.

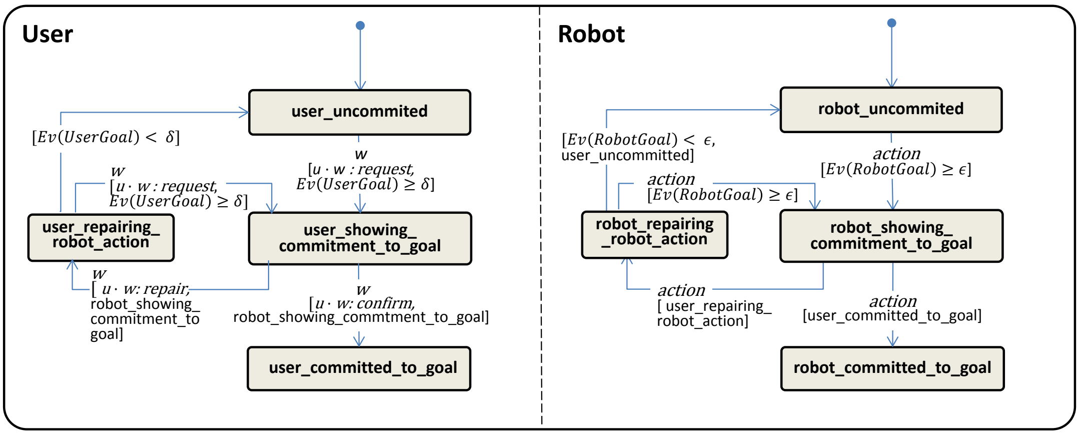


Figure 2: PentoRob's Interactive Statechart with two parallel, concurrent states

During application, as a referring expression is recognised, each classifier for the words in the expression are applied to the puzzle pieces in view, which after normalisation, results in a probability distribution over pieces.

User and Robot State Machines For dialogue control, we use an *Interactive Statechart*– see Fig. 2. Rather than comprising a single dialogue state, there are concurrent states for each agent in the interaction with their own variables. The *User State Machine* has access to the estimated current user goal $UserGoal$ and a *strength-of-evidence* function $Ev(UserGoal)$, both of which can be defined by the designer. In our domain $UserGoal$ is the taking of the most likely object according to WAC's output distribution given the utterance u so far and the Ev function as the probability value of the highest ranked object in WAC's distribution over its second highest rank. If $UserGoal$ is changed or instantiated, a new *ActionRequestIU* is made available in its right buffer with the goal.

The *Robot State Machine* gets access to its transition conditions involving the user through the ActionRequestIUs. Through a simple planning function, a number of ActionIUs are cued to achieve the goal. It sends these as RSB messages to the PentoRob control module and once confirmed, via RSB, that the action has begun, the ActionIU is *committed*. For estimation of its own state, the robot state has a strength-of-evidence function $Ev(RobotGoal)$ defined by the designer.

PentoRob control module The module controlling the robotic actuation of the ShapeOKO arm is a Python module with an Arduino board G-code interface. This sends RSB feedback messages to the Robot State Machine to the effect that actions have been successful or unsuccessful.

3 Use cases

We are currently experimenting with achieving more fluidity for grounding behaviour in human-robot interaction. The statechart in Fig. 2 has parameters δ and ϵ which are thresholds for the $Robot$ and $User$ that must be reached by the Ev functions to show sufficient evidence of each agent's goal. Early results show that lower thresholds allow more fluid grounding behaviour, while higher thresholds are 'safer' for task success. We are planning a series of related experiments. Other areas where our setup could be used is learning grounded semantics for verbs.

Acknowledgments

We thank Oliver Eickmeyer, Casey Kennington and Livia Dia for contributions to software and hardware. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG), and the DFG-funded DUEL project (grant SCHL 845/5-1).

References

Timo Baumann and David Schlangen. 2012. The inprotk 2012 release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*. ACL.

David Harel. 1987. Statecharts: A visual formalism for complex systems. *Science of computer programming*, 8(3).

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

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1).

Generating Surplus Content in a Q/A-Setting

Sebastian Reuße and **Ralf Klabunde**
`{reusse,klabunde}@linguistics.rub.de`
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum

Jon Stevens and **Anton Benz**
`{benz,stevens}@zas.gwz-berlin.de`
ZAS Berlin

Abstract

We present a framework for question-answering systems which employ pragmatic reasoning based on domain-level goals inferred from users' queries. Inferred user goals are used to anticipate how generated replies are interpreted. This allows the system to predict when surplus content can be felicitously included for the sake of discourse efficiency, and in cases where surplus content is strictly required to avoid supplying misleading answers.

1 Background

The system presented here is part of the *PragSales* project, which set out to develop a Q/A-system architecture for imperfectly cooperative domains. This paper illustrates an application to the real-estate sales domain, where prospective tenants looking for an apartment to let interact with an automated real-estate agent. Interaction proceeds based on customer's queries, posed in the form of Y/N-questions for attributes of the flat being discussed. Within this setting, the system needs to 1. ensure the *felicity* of generated utterances, and 2. keep interactions *efficient* and non-repetitive. We propose a theoretically grounded framework which is able to address both of these concerns from a unified perspective.

2 Architecture

Domain Model. A model of a real-estate sales domain was implemented using an adapted version of *PyKE*, Horn logic theorem prover. A range of predicates were defined to represent a variety of flat-intrinsic attributes, such as size and pricing, as well as extrinsic attributes pertaining to a flat's surroundings, such as distances to public transit stops, schools and grocery stores.

User Model. We posit that users issue queries in order to ascertain whether a flat fulfills one or more of their underlying *goals*, i. e., sets of preconditions on desirable flats. We formalize these as sets of Horn-logic implications, where a goal term is fulfilled when one or more conjunctions of attribute predicates uphold. E. g., we might define that some customers are seeking a place which allows them to enjoy some time in the sun, and that such a goal may be realized in either of two ways:

```
sunTan ← flat(e) ∧ hasGarden(e)
sunTan ← flat(e) ∧ hasBalcony(e)
```

Goal Inference. A user goal underlying each particular query is inferred by performing a Bayesian update on a flat prior over user goals g, given an observed query term a. Each query term is deemed an a priori equally likely option of finding out about its superordinate goal.

Dialogue Management As opposed to a canonical Q/A-setting, our application setting allows for queries to be posed in an iterative fashion, simulating a dialogical interaction between a sales agent and a customer. To this end, we implemented a minimal dialogue manager using the *PyTrindiKit* toolkit.

3 Anticipating Pragmatic Inference

Whereas canonical Q/A-systems create a relatively constrained interpretation context, the continuous fashion in which users interact with our application induces a context that makes it plausible for users to assume the sales agent to be aware that their overt queries are motivated by implicit goals. Consequently, replies trigger a pragmatically enriched interpretation of system responses (van Rooy, 2003). Specifically, a game-theoretic analysis of our domain predicts two interrelated types of pragmatic inference (Stevens et al., 2014), both of which a dialogue move engine should take into account when generating replies:

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 116-117.

1. Users assume the system to supply the maximally beneficial reply, so that unrealized, better answers are assumed to be negated. This allows No-responses to be addressed indirectly, or in the form of *No, but...* replies, which acts as a license to supply unrequested, but beneficial information.
2. As a corrolary, plain Yes/No-responses can trigger misleading implicatures which an automated system may want to avoid.

In the following, we briefly outline how both of these concerns can be addressed within our framework.

Increasing Dialogue Efficiency. Based on our supposition that what is relevant for the purposes of the conversation is determined by implicit user goals, rather than by overt query terms, representing these goals within the scope of the system allows generating *indirect* response types which give *unrequested, but goal-relevant information*, as in (1-b), or which leave out an overt *No*, thereby increasing dialogue efficiency, as shown in evaluations by Stevens et al. (2015).

(1) Q.Does the flat feature a balcony?
  a. Well, it features a garden.
    $\rightsquigarrow$ It does not feature a balcony.
  b. No, but there's a garden.

Since, given the user's underlying goal, a *garden* is a valid alternative to a *balcony*, our system can predict that (1-a/b) are coherent responses to the original query, which may be preferred for the sake of discourse efficiency.

Blocking Undesirable Inferences. In some contexts, the assumption that answer interpretation is strengthened based on common knowledge about the user's underlying goals allows us to predict answers which, while factually true, implicate false propositions. In (2-a), a literal *No* response negates, by implicature, not only an overt query term, but all manifestations of the underlying goal, since a cooperative or self-interested seller would have included any relevant, true alternatives in his response.

(2) Q.Does the flat feature a balcony?
  a. No. $\rightsquigarrow$ It's not good for getting a tan.
  b. No, but if you're looking to get a tan, there's a garden.

In addition to literal responses, our system generates contrastive responses, such as (2-c), which block the undesirable inference. We define the underlying contrastive message operator $but(m_1, m_2)$ to be formally licensed when: 1. both m_1 and m_2 are sufficient preconditions of some user goal D, 2. the user issues a query for m_1, 3. m_1 fails to uphold, and 4. m_2 upholds.

Vice versa, a literal *Yes* response triggers an expectation for the superordinate user goal to be fulfilled, as in (3-a). This expectation would mislead the user when this is not actually the case.

(3) Q.Does the flat feature a balcony.
  a. Yes. $\rightsquigarrow$ The balcony is a place to work up a tan.
  b. Yes, though it faces north.

In cases such as this, our system is able to generate concessive replies such as (3-b), using a concession operator $although(m_1, m_2)$ which is licensed when: 1. both m_1 and m_2 are necessary preconditions of some user decision problem D, 2. the user issues a query for m_1, 3. m_1 upholds, and 4. m_2 fails to uphold.

4 Discussion

We believe that the way in which our framework allows formalizing the use conditions of a variety of linguistic means, encompassing both indirect and augmented answers, as well as contrastive resp. concessive replies, shows that the theory of van Rooy (2003) forms a workable base on top of which pragmatic phenomena can be handled within the scope of a Q/A-setting.

References

Stevens, J. S., Benz, A., Reuße, S., Klabunde, R., & Raithel, L. (2015). Pragmatic query answering: results from a quantitative evaluation. In C. Biemann, S. Handschuh, A. Freitas, F. Meziane, & E. Métais (Eds.), *Natural language processing and information systems* (pp. 110–123). Lecture Notes in Computer Science. Springer.

Stevens, J. S., Benz, A., Reuße, S., Laarmann-Quante, R., & Klabunde, R. (2014). Indirect answers as potential solutions to decision problems. In *Proceedings of the 18th workshop on the semantics and pragmatics of dialogue* (pp. 145–153).

van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*(6), 727–763.

Aligning Intentions: Acceptance and Rejection in Dialogue.

Julian J. Schlöder
ILLC / Amsterdam
`j.j.schloder@uva.nl`

Antoine Venant
IRIT / Toulouse
`antoine.venant@irit.fr`

Nicholas Asher
IRIT / Toulouse
`asher@irit.fr`

Abstract

We present a novel framework for a formal pragmatics and show applications with a special focus on the dynamics of agreement and disagreement. We are particularly interested in *intentions*. We circumvent the notoriously difficult task of axiomatizing agents' *internal* intentions by reducing *externalized* intentions to commitments to preferred futures. We give a formally precise account of both Stalknakerian rejections and more general rejections of arbitrary speech acts.

Many accounts of pragmatic reasoning and utterance interpretation refer to the speakers' intentions (Cohen et al. 1990 is a collection of related work). For example, the model of Asher and Lascarides 2013 verifies the following chain of inferences to govern agreement in a cooperative dialogue between A and B:[1]

$$C_A p \mathrel{|\!\sim} C_A I_A C_B p \mathrel{|\!\sim} I_B C_B p$$
$$\text{by (}\mathtt{ISC}\text{) } C_A\varphi > C_A I_A C_B \varphi$$
$$\text{and (}\mathtt{Co}\text{) } C_A I_A \varphi > I_B \varphi.$$

That is, by asserting that p, the speaker A establishes the commitment $C_A p$. By `Intent to Share Commitment (ISC)`, A intends that B make the same commitment, $C_A I_A C_B p$. Then, `Cooperativity (Co)` infers that B adopts this intention, $I_B C_B p$. In her next move, B is expected to fulfill this intention by making an agreement move. The two axioms formalise two truisms about dialogues: that speakers want to be agreed with, and that, in cooperative settings, intentions are mutually adopted (Clark, 1996). A problem with such accounts is that the notion *intention* is notoriously nebulous. Intentions are inherently private to the interlocutors. Hence it is hard to give a motivated semantics to operators like I_A above, *i.e.*, to say when $I_A\varphi$ is true and, if it is, what grounds its truth.

The first contribution of our work reduces propositions about dialogue intentions to propositions that have truth-conditions grounded in the external world. Informally, an intention for us restricts a space of possible futures. That is, we identify intentions with futures to which an agent commits. To this end, we integrate the language of *temporal modal logic* (where $\Diamond$ means 'eventually') into an action model. We then can say that to-intend-that-α is the *committment* that α-will-happen, *i.e.*, in the language of Asher and Lascarides, $I_A\alpha$ is $C_A\Diamond\alpha$. On first glance, this seems to overstate the matter in two ways: (i) one surely does not externalise (by the public commitment operator C_A) all internal intentions; and (ii) merely intending something is strictly weaker than claiming that it will happen. A brief summary of our counterarguments is as follows.

We concede (i), but argue that all intentions that are relevant to the dialogue are externalised: wherever required for utterance interpretation, they are presupposed or inferred and—as such—on the public record with the utterance. The overstatement (ii) is in fact not as severe as it seems, if such commitments to futures are only inferred by non-monotonic inference. For individual actions α, it is clear that '*A intends* α' non-monotonically entails that '*A will do* α.'[2] If one voices an intention to a (linguistic) *joint* action, one can generally expect that cooperative interlocutors will partake.[3] Of course, these inferences are not mono-

---

[1] $\mathrel{|\!\sim}$ is defeasible inference, $>$ a default conditional, $C_A\varphi$ means that the speaker A is publicly *committed* to the formula φ and $I_A\varphi$ means that the speaker A *intends* to establish a state that brings about φ.

[2] Note that this separates intentions from mere desire. For example, I might desire to go to an expensive restaurant, but if I'm pressed for money, I will not intend it.

[3] For example, one speaks an utterance with the intent to be understood. Then, in speaking an utterance one also *com-*

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 118-119.

tonic. Therefore we make use of default logic in our formalisation of intentions in dialogue.

We model temporal logic as KT4 modal logic and state the following basic coordination principles for commitments and temporal operators ($>$ denotes a default conditional).

(a) $C_A\varphi > C_A\neg\Diamond\neg C_A\varphi$. Commitments are not intended to be broken (*i.e.*, to be kept).

(b) $C_A\neg C_A\varphi > C_A\neg\Diamond C_A\varphi$. Withhold judgements are intended to be kept.

To do pragmatics in this language we formalise two truisms: interlocutors intend to reach eventual agreement, and if intentions align they are actualised.

(c) $C_A(\Diamond\Box(C_A\varphi \leftrightarrow C_B\varphi))$ and
$C_B(\Diamond\Box(C_A\varphi \leftrightarrow C_B\varphi))$
(Eventual agreement).

(d) $((C_A\Diamond\varphi) \wedge (C_B\Diamond\varphi)) > ((C_A\varphi) \wedge (C_B\varphi))$
(Alignment of intentions).

Underlying our analysis is a sophisticated analysis of speech acts that models them as actions in a dynamic logic: For A to make an assertion that p or to agree to p is to make an action that changes the commitment structure of A, *i.e.*, C_Ap. We view speech acts as such actions and therefore can model the temporal operators as ranging over sequences of action-induced model transitions. In this model, we (i) generalise the Asher-Lascarides model above, (ii) give a formalisation of Stalnakerian rejection, and (iii) generalise on (i) and (ii) to formalise the acceptance and rejection of arbitrary speech acts.

For (i), we verify that $C_Ap \mathrel{|\!\sim} C_A\Diamond C_Bp$, *i.e.*, that A intends that B agree. This corresponds to (`ISC`).[4] To *reject* an assertion, according to Stalnaker 1978, is to refuse to accept it; this distinguishes it from previous analyses of *correction* where a speaker asserts a revised version of an assertion. For (ii), we formalise the action *'B rejects p'* as effecting $C_B\neg C_Bp$ (*i.e.*, to commit to not accepting) and derive $C_B\neg C_Bp \mathrel{|\!\sim} C_B\Diamond\neg C_Ap$ (*i.e.*, B wants A to retract). However, a rejection also presupposes that the rejected proposition is *understood*; B understanding A is C_BC_Ap (Venant et al., 2014). We integrate this into our action structure and verify that it does not interfere with the above derivations.

In addition, *any* speech act can be rejected, *i.e.*, refused to be taken up (Austin, 1962). To our knowledge, there is no extant formal model for this. Under an Austinian conception of felicity, speech acts are made to *intentionally* bring forth a change; Clark 1996 characterises linguistic acts as *joint actions* in general. Thus, the rejection of a speech act is to refuse participating in its intended action. We account for this as follows: If A makes a speech act with intended effect φ, our model sees this as $C_A\Diamond\varphi$ (as shown above, our model derives that the effect of an assertion is to project acceptance, and the effect of rejecting-an-assertion is to prompt retraction). For B to take φ up is to adopt the intention, *i.e.*, $C_B\Diamond\varphi$. Upon such uptake, φ is realised by both speakers according to (d). Conversely, for B to reject a speech act with effect $\Diamond\varphi$ is to commit to negating the effect, *i.e.*, $C_B\neg\Diamond\varphi$. This generalises the narrower Stalnakerian conception of rejection described above. The effect of A asserting p is $\Diamond C_Bp$, and B rejecting this is $C_B\neg\Diamond C_Bp$. Our axioms verify that $C_B\neg\Diamond C_Bp \mathrel{|\!\sim} C_B\Diamond\neg C_Ap$—the same as B rejecting by $C_B\neg C_Bp$.

Applying this framework to further pragmatic phenomena, most notably to non-cooperative dialogue settings, we can show that different degrees of cooperativity correspond to structural properties of commitment structures. Also, we can equip our logic with a full model theory, giving truth conditions to statements about intentions. This is done using the *conversations as infinite games* framework of Asher and Paul 2013. In the model theory, a conversation is an unbounded sequence of speech acts; basic commitments (*i.e.*, commitments to propositions) are monotonic consequences of these speech acts, and temporal operators partition the space of sequence continuations.

References

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6:1–58.

Nicholas Asher and Soumya Paul. 2013. Infinite games with uncertain moves. In *1st SR Workshop,* p. 25–32.

John L. Austin. 1962. *How to do Things with Words.* Clarendon Press.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

P. R. Cohen, J. Morgan, and M. Pollack, editors. 1990. *Intentions in Communication.* MIT Press.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics (Syntax and Semantics 9).* Academic Press.

Antoine Venant, Nicholas Asher, and Cedric Degremont. 2014. Credibility and its attacks. In *SemDial 18,* p. 154–162.

---

mits that it *can* be understood and thus that it, eventually, *will* be understood.

[4] $C_B\Diamond C_Bp$ (analogous `Co`), follows if commitments satisfy axiom 5: $\neg C_A\neg p \rightarrow C_A\neg C_A\neg p$.

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

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

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

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

Abstract

We present a multi-modal dialogue system for interactive learning of perceptually grounded word meanings from a human tutor. The system integrates an incremental, semantic, and bi-directional grammar framework – Dynamic Syntax and Type Theory with Records (DS-TTR[1], (Eshghi et al., 2012; Kempson et al., 2001)) – with a set of visual classifiers that are learned throughout the interaction and which ground the semantic/contextual representations that it produces (c.f. Kennington & Schlangen (2015)) Our approach extends Dobnik et al. (2012) in integrating perception (vision in this case) and language within a single formal system: Type Theory with Records (TTR (Cooper, 2005)). The combination of deep semantic representations in TTR with an incremental grammar (Dynamic Syntax) allows for complex multi-turn dialogues to be parsed and generated (Eshghi et al., 2015). These include clarification interaction, corrections, ellipsis, and utterance continuations (see e.g. the dialogue in Fig. 1).

1 Architecture

The system is made up of two key components – a vision system and the DS-TTR parser/generator. The vision system classifies a (visual) situation, i.e. deems it to be of a particular type, expressed as a TTR Record Type (RT) (see Fig. 1). This is done by deploying a set of binary attribute classifiers (Logistic Regression SVMs with Stochastic Gradient Descent (Yu et al., 2015)) which ground the simple types (atoms) in the system (e.g. 'red', 'square'), and composing their output to construct the total type of the visual scene. This representation then acts not only as (1) the non-linguistic context of the dialogue for DS-TTR, for the resolution of e.g. definite references and indexicals, see Hough & Purver (2014); but also (2) the logical database from which answers to questions about object attributes are generated. Questions are parsed and their logical representation acts directly as a query on the non-linguistic/visual context to retrieve an answer (via *type checking* in TTR, itself done via *unification*, see Fig. 1). Conversely, the system can generate questions to the tutor (Yu et al., 2016b) about the attributes of objects based on the entropy of the classifiers that ground the semantic concepts, e.g. those for colour and shape. The tutor's answer then acts as a training instance for the classifiers (basic, atomic types) involved - see Fig. 1 for a screenshot.

2 Learning via Incremental Dialogue

Interaction with a human tutor enables systems to take initiative to seek the particular information they need by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 1). For example, a robot could ask questions to learn the colour of a "square" or to request to be presented with more "red" things to improve performance. Furthermore, such systems could allow for meaning negotiation in the form of clarification interactions with the tutor.

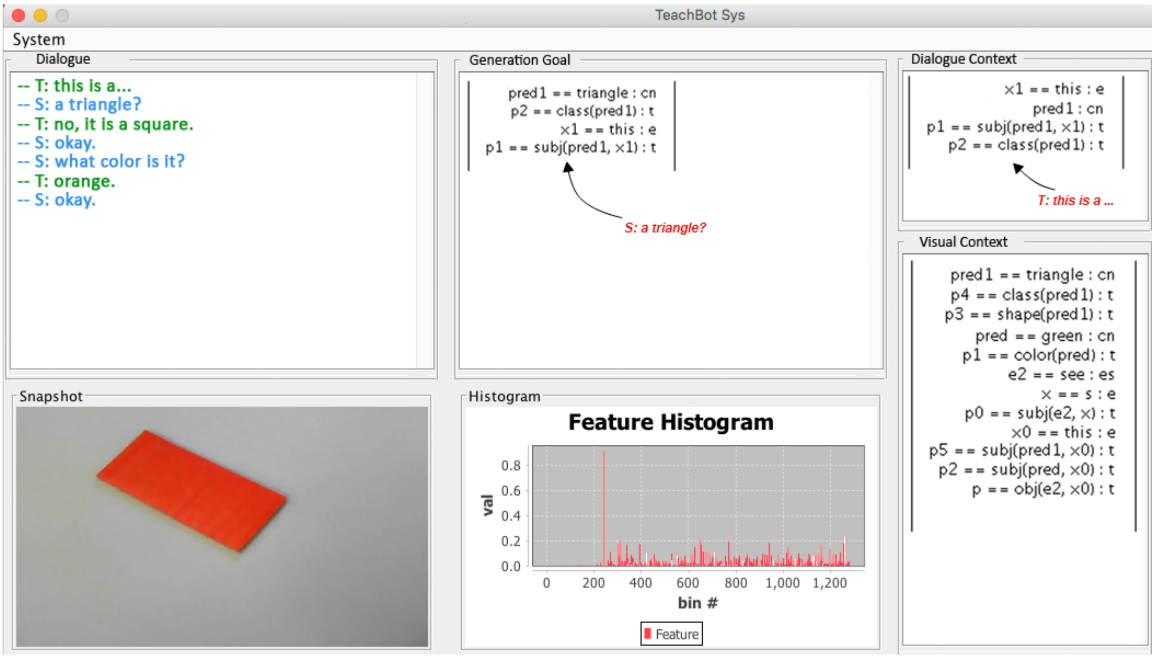
Dialogue with the tutor continuously provides semantic information about objects in the visual scene which is then fed to online classifiers in the form of training instances. Conversely, the system can utilise the DS-TTR grammar and its existing knowledge about the world, encoded in its classifiers, to make reference to and formulate questions about the different attributes of objects identified in the visual scene.

We will show an interactive demonstration of this system, illustrating how questions, answers

---

[1] Download at `sourceforge.net/projects/dylan/`

Appears in *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue,*
edited by Julie Hunter, Mandy Simons and Matthew Stone. New Brunswick, NJ USA,
July 16-18, 2016, pages 120-121.



-- T: this is a...
-- S: a triangle?
-- T: no, it is a square.
-- S: okay.
-- S: what color is it?
-- T: orange.
-- S: okay.

```
       pred1 == triangle : cn
        p2 == class(pred1) : t
               x1 == this : e
p1 == subj(pred1, x1) : t
```

S: a triangle?

```
         x1 == this : e
            pred1 : cn
p1 == subj(pred1, x1) : t
  p2 == class(pred1) : t
```

T: this is a ...

```
 pred1 == triangle : cn
 p4 == class(pred1) : t
p3 == shape(pred1) : t
     pred == green : cn
   p1 == color(pred) : t
         e2 == see : es
              x == s : e
    p0 == subj(e2, x) : t
          x0 == this : e
p5 == subj(pred1, x0) : t
 p2 == subj(pred, x0) : t
    p == obj(e2, x0) : t
```

Figure 1: Incremental, visually grounded dialogue in the Concept Learning System. T= tutor, S=system

and object descriptions are derived and generated incrementally by the Concept Learner in real-time. Work in progress addresses: (1) optimising the Learner dialogue strategy (Yu et al., 2016a); (2) data-driven, incremental dialogue management at the lexical level.

Acknowledgements

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

References

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

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *Proc. CSLP*.

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.

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proc. IWCS*.

Julian Hough and Matthew Purver. 2014. Probabilistic type theory for incremental dialogue processing. In *EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88.

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

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proc. ACL-IJCNLP*.

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

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2015. Comparing attribute classifiers for interactive language grounding. In *Proceedings of ENMLP workshop on Vision and Language*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016a. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *(under review)*.

Yanchao Yu, Oliver Lemon, and Arash Eshghi. 2016b. Comparing dialogue strategies for learning grounded language from human tutors. In *Proceedings of SEMDIAL*.

---

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

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