Universität Potsdam

David Schlangen | Raquel Fernández (Eds.)

brandial'06

Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10)

Potsdam, Germany — September 11-13, 2006

Preface

With *brandial'06* we celebrate the 10th installment of the Semantics and Pragmatics of Dialogue workshop series.

The series started in Munich in 1997, at the initiative of Gerhard Jäger and Anton Benz, and from then on wandered the European continent: Twente (Twendial'98), Amsterdam (Amstelogue'99), Gothenburg (Götalog'00), Bielefeld (BiDialog'01), Edinburgh (Edilog'02), Saarbrücken (DiaBruck'03), Barcelona (Catalog'04), Nancy (Dialor'05). Focussing more on formal semantics and pragmatics in the early editions, the workshop has over the years widened its scope to include work from many other disciplines concerned with building models of dialogue—Artificial Intelligence and Computational Linguistics contributing work on artificial conversational agents, and Psycholinguistics observations on natural conversational agents—and has become the premiere forum to present this kind of work.

brandial'06 continues this trend with high-quality papers and posters from all of these fields. The papers, together with invited presentations by James Allen, Elisabeth André, Manfred Krifka, and Emanuel Schegloff, and a Special Session on "Visual Attention and References to the Visual Situation" with its own invited presentation by John Kelleher and Geert-Jan Kruijff, give a good overview of the current state of dialogue research, which we hope will form the basis of lively discussion and productive further work.

We received 56 submissions. Each was reviewed by at least two experts. From the submissions, 20 were accepted as regular papers for oral presentation (including those in the Special Session and one reserve paper), and 12 as posters (of which 11 will be presented). A further call for posters and demos generated 11 submissions, of which 9 will be presented.

The reviewers did a fantastic job, producing many very detailed reviews that not only made the task of selecting papers easy, but also hopefully will help all authors, those of the accepted papers as well as of those we couldn't include, in their future work. Thank you to Jan Alexandersson, Ellen Bard, Johan Bos, Justine Cassell, Matthew Crocker, Paul Dekker, Simon Garrod, Jonathan Ginzburg, Pat Healey, Rodger Kibble, Alistair Knott, Jörn Kreutel, Ivana Kruijff-Korbayova, Staffan Larsson, Alex Lascarides, Oliver Lemon, Colin Matheson, Nicolas Maudet, Philippe Muller, Yukiko Nakano, Massimo Poesio, Matt Purver, Hannes Rieser, Michael Strube, Takenobu Tokunaga, and David Traum! The Special Session, chaired by Massimo Poesio and Hannes Rieser, had a separate committee, which we also thank: Sarah Brown-Schmidt, John Kelleher, Pia Knoeferle and Geert-Jan Kruijff.

About the name *brandial'06*: as the list of previous workshops shows, tradition demands some, preferably witty, combination of location name and the word "dialogue". After an initial proposal generated too much amusement among our American focus group, we settled on the slightly bland, but phonetically pleasant *brandial* (a combination of the name of the federal state where Potsdam is located, Brandenburg, and "dialogue").

Whatever the name, the quality of the workshop itself is more important, and if *brandial* turns out to be good, then that is due in no small measure to generous financial support from the Deutsche Forschungsgemeinschaft (German Research Council, DFG) and the EU-"Marie Curie programme"-

financed DEAWU ("dealing with uncertainty") research group at the University of Potsdam; support from the University of Potsdam in general (which let us use the venue, the Neues Palais lecture hall in Park Sanssouci, for free) and the Linguistics Department and the Chair for Applied Computational Linguistics, Manfred Stede, in particular; help during the workshop by our student assistants; endorsement by the ACL Special Interest Groups SIGdial and SIGsem; and, last but not least, tireless work by our secretary Annett Eßlinger, who handled many of the local arrangements.

David Schlangen and
Raquel Fernández

Potsdam, August 2006

Organisers and Programme Committee Chairs:

David Schlangen, University of Potsdam
Raquel Fernández, University of Potsdam

Invited Speakers:

James Allen (University of Rochester, USA)
Elisabeth André (University of Augsburg, Germany)
Manfred Krifka (Humboldt University and ZAS, Germany)
Emanuel Schegloff (University of California LA, USA)

Programme Committee:

Jan Alexandersson (DFKI, Saarbrücken)
Ellen Bard (University of Edinburgh)
Johan Bos (Università di Roma La Sapienza)
Justine Cassell (Northwestern University)
Matthew Crocker (Universität des Saarlandes)
Paul Dekker (University of Amsterdam)
Simon Garrod (University of Glasgow)
Jonathan Ginzburg (King's College London)
Pat Healey (Queen Mary University of London)
Rodger Kibble (Goldsmiths University of London)
Alistair Knott (University of Otago)
Jörn Kreutel (Semantic Edge and University of Potsdam)
Ivana Kruijff-Korbayova (Universität des Saarlandes)
Staffan Larsson (Gothenburg University)
Alex Lascarides (University of Edinburgh)
Oliver Lemon (University of Edinburgh)
Colin Matheson (University of Edinburgh)
Nicolas Maudet (Université Paris-Dauphine)
Philippe Muller (Université Paul Sabatier)
Yukiko Nakano (RISTEX Japan)
Massimo Poesio (University of Essex)
Matthew Purver (CSLI Stanford)
Hannes Rieser (Universität Bielefeld)
Michael Strube (EML Research)
Takenobu Tokunaga (Tokyo Institute of Technology)
David Traum (ICT, University of Southern California)

Special Session on Visual Attention and Reference to the Visual Situation:

Sarah Brown-Schmidt (University of Illinois at Urbana)
John Kelleher (Dublin Institute of Technology)
Pia Knoeferle (Universität des Saarland)
Geert-Jan Kruijff (DFKI, Saarbrücken)
Massimo Poesio (University of Essex) *chair*
Hannes Rieser (Universität Bielefeld) *chair*

Conference Programme

Monday, September 11, 2006

09:00–09:15	Opening Remarks
09:15–10:15	*A Robust Dialogue Agent for Collaborative Problem Solving* James F. Allen
10:15–10:45	*Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems* Norbert Pfleger and Jan Alexandersson
10:45–11:15	Coffee Break
11:15–11:45	*Understanding Student Input for Tutorial Dialogue in Procedural Domains* Myroslava O. Dzikovska , Charles B. Callaway , Matthew Stone and Johanna D. Moore
11:45–12:15	*A multi-speaker dialogue system for computer-aided language learning* Peter Vlugter and Alistair Knott
12:15–13:30	Lunch
13:30–14:00	*Correction and acceptance by contrastive focus* Elena Karagjosova
14:00–14:30	*Modelling Correction Signalled by “But” in Dialogue* Kavita Thomas
14:30–15:15	Posters/Demo Session 1
15:15–15:45	Coffee Break
15:45–16:15	*From Complex to Simple Speech Acts: a Bidimensional Analysis of Illocutionary Forces* Claire Beyssade and Jean-Marie Marandin
16:15–16:45	*Question Intonation and Lexicalized Bias Expression* Yuri Hara
16:45–17:45	*Question-Answer Congruence and the Proper Representation of Focus* Manfred Krifka

Tuesday, September 12, 2006

09:00–09:30	*Identifying Reference to Abstract Objects in Dialogue* Ron Arstein and Massimo Poesio
9:30–10:00	*Formal Semantics of Iconic Gesture* Alex Lascarides and Matthew Stone
10:00–10:30	*SDRT and Multi-modal Situated Communication* Andy Lücking, Hannes Rieser and Marc Staudacher
10:30–11:00	Coffee Break
11:00–12:00	*Engaging in a Multimodal Dialogue with Embodied Conversational Agents along the Virtuality Continuum* Elisabeth André
12:00–13:30	Lunch
13:30–14:15	Poster/Demo Session 2
	Special Session
14:15–14:45	*Handling Spatial Reference in Visually-Situated Dialogs* Gert-Jan M. Kruijff and John D. Kelleher
14:45–15:15	*Measuring and Reconstructing Pointing in Visual Contexts* Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser and Marc Staudacher
15:15–15:45	Coffee Break
15:45–16:15	*Verbal or Visual? How Information is Distributed across Speech and Gesture in Spatial Dialog* Kirsten Bergmann and Stefan Kopp
16:15–16:45	*Cross Recurrence Quantification Analysis of indefinite anaphora in Swedish dialog. An eye-tracking pilot experiment* Philip Diderichsen
16:45–17:15	*Multi-modal Integration for Gesture and Speech* Andy Lücking, Hannes Rieser and Marc Staudacher
18:15–21:30	Conference Dinner

Wednesday, September 13, 2006

09:30–10:00	*Acquiring Words across Generations: Introspectively or Interactively?* Zoran Macura and Jonathan Ginzburg
10:00–10:30	*Clarifying Spatial Descriptions: Local and Global Effects on Semantic Co-ordination* Gregory J. Mills and Patrick G.T. Healey
10:30–11:00	Coffee Break
11:00–12:00	*On "Uh" and "Uhm" and some of the Things They are Used to Do* Emanuel Schegloff
12:00–13:00	Lunch
13:30–14:00	*Goal-oriented Dialog as Collaborative Subordinated Activity Involving Collaborative Acceptance* Sylvie Saget and Marc Guyomard
14:00–14:30	*Scorekeeping in an Uncertain Language Game* David DeVault and Matthew Stone
15:15–15:45	Coffee Break
15:45–16:15	*Modeling Anaphora in Informal Mathematical Dialogue* Magdalena Wolska and Ivana Kruijff-Korbayova
16:15–16:30	Closing Remarks

Reserve Paper:

Dialogue Pressures and Syntactic Change
Ruth Kempson and Ronnie Cann

Poster/Demo Session 1

Route Instruction Dialogues with a Robotic Wheelchair
Thora Tenbrink, Hui Shi and Kerstin Fischer

Non-verbal Information in Communication: A Study of Interaction in a Tourist-information Setting
Ichiro Umata, Sadanori Ito, Shoichiro Iwasawa, Noriko Suzuki, Tomoji Toriyama, Naomi Inoue, Kiyoshi Kogure and Kenji Mase

Compliments in Virtual Dialogs
Lilia Mironovschi

An Algorithm for Context Update Mechanisms in Cooperative Dialogue
Roser Morante and Simon Keizer

ISA - Corpus Annotation with OWL
Stephanie Becker, Thomas Kleinbauer and Stephan Lesch

Browsing Meetings: Automatic Understanding, Presentation and Feedback for Multi-Party Conversations
Patrick Ehlen, Stephane Laidebeure, Niekrasz Niekrasz, Matthew Purver, John Dowding and Stanley Peters

Poster/Demo Session 2

Scene-Sentence Integration: Incremental Effects of Mismatch and Scene Complexity
Pia Knoeferle and Monica Rodriguez

Perspective Guides Interpretation of Questions, Declarative Questions and Statements in Unscripted Conversation
Sarah Brown-Schmidt, Christine Gunlogson, Duane Watson and Michael T. Tanenhaus

Prosodic Marking of Contrasts in Information Structure
Markus Guhe, Markus Steedman, Ellen Bard and Max Louwerse

Evaluation of an Information State-Based Dialogue Manager
Antonio Roque, Hua Ai and David Traum

Dialogue management for cooperative, symmetrical human-robot interaction
Mary Ellen Foster

Hierarchical Reinforcement Learning of Dialogue Policies in a development environment for dialogue systems: REALL-DUDE
Oliver Lemon, Xingkun Liu, Daniel Shapiro and Carl Tollander

Dico: A Multimodal Menu-based In-vehicle Dialogue System
Jessica Villing and Staffan Larsson

Poster/Demo Session 3

Accommodating Social Relationships
Joris Hulstijn

The Effect of Multiple Modalities in Dialogue Act Annotation
Crystal Nakatsu and Chris Brew

Building a Multi-Lingual Interactive Question-Answering System for the Library Domain
Manuel Kirschner

Referring Expressions and Local Coherence of Discourse in the Parallel Corpus of English and Japanese Map Task Dialogues
Etsuko Yoshida

How Speakers Represent Size Information in Referential Communication for Knowing and Unknowing Recipients
Judith Holler and Rachel Stevens

Programming by Voice: Enhancing Adaptivity and Robustness of Spoken Dialogue Systems
Kallirroi Georgila

Table of Contents

A Robust Dialogue Agent for Collaborative Problem Solving
James F. Allen 1

Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems
Norbert Pfleger and Jan Alexandersson 2

Understanding Student Input for Tutorial Dialogue in Procedural Domains
Myroslava O. Dzikovska, Charles B. Callaway, Matthew Stone and Johanna D. Moore 10

A multi-speaker dialogue system for computer-aided language learning
Peter Vlugter and Alistair Knott 18

Correction and acceptance by contrastive focus
Elena Karagjosova 26

Modelling Correction Signalled by "But" in Dialogue
Kavita Thomas 34

From Complex to Simple Speech Acts: a Bidimensional Analysis of Illocutionary Forces
Claire Beyssade and Jean-Marie Marandin 42

Question Intonation and Lexicalized Bias Expression
Yuri Hara 50

Question-Answer Congruence and the Proper Representation of Focus
Manfred Krifka 55

Identifying Reference to Abstract Objects in Dialogue
Ron Arstein and Massimo Poesio 56

Formal Semantics of Iconic Gesture
Alex Lascarides and Matthew Stone 64

SDRT and Multi-modal Situated Communication
Andy Lücking, Hannes Rieser and Marc Staudacher 72

Engaging in a Multimodal Dialogue with Embodied Conversational Agents along the Virtuality Continuum
Elisabeth André 80

Handling Spatial Reference in Visually-Situated Dialogs
Geert-Jan M. Kruijff and John D. Kelleher 81

Measuring and Reconstructing Pointing in Visual Contexts
Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser and Marc Staudacher 82

Verbal or Visual? How Information is Distributed across Speech and Gesture in Spatial Dialog
Kirsten Bergmann and Stefan Kopp 90

Cross Recurrence Quantification Analysis of indefinite anaphora in Swedish dialog. An eye-tracking pilot experiment
Philip Diderichsen 98

Multi-modal Integration for Gesture and Speech
Andy Lücking, Hannes Rieser and Marc Staudacher 106

Acquiring Words across Generations: Introspectively or Interactively?
Zoran Macura and Jonathan Ginzburg 114

Clarifying Spatial Descriptions: Local and Global Effects on Semantic Co-ordination
Gregory J. Mills and Patrick G.T. Healey 122

On "Uh" and "Uhm" and some of the Things They are Used to Do
Emanuel Schegloff 130

Goal-oriented Dialog as Collaborative Subordinated Activity Involving Collaborative Acceptance
Sylvie Saget and Marc Guyomard 131

Scorekeeping in an Uncertain Language Game
David DeVault and Matthew Stone 139

Modeling Anaphora in Informal Mathematical Dialogue
Magdalena Wolska and Ivana Kruijff-Korbayová 147

Dialogue Pressures and Syntactic Change
Ruth Kempson and Ronnie Cann 155

Route Instruction Dialogues with a Robotic Wheelchair
Thora Tenbrink, Hui Shi and Kerstin Fischer 163

Non-verbal Information in Communication: A Study of Interaction in a Tourist-information Setting
Ichiro Umata, Sadanori Ito, Shoichiro Iwasawa, Noriko Suzuki, Tomoji Toriyama, Naomi Inoue, Kiyoshi Kogure and Kenji Mase 165

Compliments in Virtual Dialogs
Lilia Mironovschi 167

An Algorithm for Context Update Mechanisms in Cooperative Dialogue
Roser Morante and Simon Keizer 169

ISA - Corpus Annotation with OWL
Stephanie Becker, Thomas Kleinbauer and Stephan Lesch 171

Browsing Meetings: Automatic Understanding, Presentation and Feedback for Multi-Party Conversations
Patrick Ehlen, Stéphane Laidebeure, John Niekrasz, Matthew Purver, John Dowding and Stanley Peters 173

Scene-Sentence Integration: Incremental Effects of Mismatch and Scene Complexity
Pia Knoeferle and Monica Rodriguez 175

Perspective Guides Interpretation of Questions, Declarative Questions and Statements in Unscripted Conversation
Sarah Brown-Schmidt, Christine Gunlogson, Duane G. Watson and Michael T. Tanenhaus 177

Prosodic Marking of Contrasts in Information Structure
Markus Guhe, Markus Steedman, Ellen Gurman Bard and Max M. Louwerse 179

Evaluation of an Information State-Based Dialogue Manager
Antonio Roque, Hua Ai and David Traum 181

Dialogue management for cooperative, symmetrical human-robot interaction
Mary Ellen Foster 183

Hierarchical Reinforcement Learning of Dialogue Policies in a development environment for dialogue systems: REALL-DUDE
Oliver Lemon, Xingkun Liu, Daniel Shapiro and Carl Tollander 185

Dico: A Multimodal Menu-based In-vehicle Dialogue System
Jessica Villing and Staffan Larsson 187

Accommodating Social Relationships
Joris Hulstijn 189

The Effect of Multiple Modalities in Dialogue Act Annotation
Crystal Nakatsu and Chris Brew 191

Building a Multi-Lingual Interactive Question-Answering System for the Library Domain
Manuel Kirschner 193

Referring Expressions and Local Coherence of Discourse in the Parallel Corpus of English and Japanese Map Task Dialogues
Etsuko Yoshida 195

How Speakers Represent Size Information in Referential Communication for Knowing and Unknowing Recipients
Judith Holler and Rachel Stevens 197

Programming by Voice: Enhancing Adaptivity and Robustness of Spoken Dialogue Systems
Kallirroi Georgila and Oliver Lemon 199

A Robust Dialogue Agent for Collaborative Problem Solving
(invited presentation)

James F. Allen
Department of Computer Science
University of Rochester
Rochester, NY 14627
`James@cs.rochester.edu`

While plan- or agent- based models provide interesting formalisms for producing dialogue agents, most robust dialogue system use much simpler technology based on finite state machines and/or slot filling models. Such simpler models allow the designers to encode strong expectations that enable robust processing, however the complexity of the tasks they can represent is quite limited. We are developing dialogue agents that not only can execute hand-coded tasks, but can also learn new tasks from the user through combined demonstration and dialogue. All system behavior, including its dialogue capabilities, is driven by the execution of declaratively specified tasks that model the collaborative problems solving process in a domain-independent manner. We will describe this model and explore how it contributes to producing robust dialogue behavior on a number of different levels. I will also show running examples of our system as it learns and executes new procedures that involve finding and using information on the world wide web.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, page 1.

Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems*

Norbert Pfleger and **Jan Alexandersson**
DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
{pfleger,janal}@dfki.de

Abstract

We present an extension to a comprehensive context model that has been successfully employed in a number of practical conversational dialogue systems. The model supports the task of multimodal fusion as well as that of reference resolution in a uniform manner. Our extension consists of integrating implicitly mentioned concepts into the context model and we show how they serve as candidates for reference resolution.

1 Introduction

The basis for any spoken interaction between two or more interlocutors is common ground. The common ground of two people is the sum of their mutual beliefs about the conversation and the surrounding world. As indicated in (Clark, 1996), when strangers meet they start building up presuppositions about the background knowledge of the other, thereby mutually updating and extending their common ground. Consequently, many referring expressions are only meaningful in the particular context in which they are uttered.

When we consider the course of interactions, it becomes clear that for virtually every contribution the context is extended with more or less related concepts. For a hearer, the process of relating new information to concepts already *known* is vital to the comprehension of a discourse. The basic mechanisms of this process—called *bridging*—is discussed in (Clark, 1977). In addition to direct references, he identifies, for instance, *indirect* references by association. Consider the following example taken from (Clark, 1977):

> “`I looked into the room. The ceiling was very high.`”

During the processing of the first sentence, a hearer with profound knowledge about rooms will most likely not only activate the concept *room* per se but also associated concepts (e. g., *ceiling*, *wallpaper*, etc.). The degree of activation is influenced by factors such as the hearer's personal interests, recency of the acquired knowledge etc. But also the situational context (current location, time, weather conditions, etc.) is vital. For instance, for a mobile dialogue system in a tourist scenario it will be necessary to activate—and deactivate—buildings and streets while the user is moving around.

The aim of our work is the incorporation of these findings into a module for reference resolution for a multimodal conversational agent. Here, we are focusing in particular on the correct interpretation of named entities and definite noun phrases whose referents have not been explicitly mentioned but are part of the implicit context. Key to our approach is the integration of a long term memory (LTM) modeling the complete knowledge of an agent. Next to this LTM is a working memory (WM) that realizes a comprehensive context model. However, as we will argue in this paper, some processes in the human LTM have direct impact on the organization and structure of the contextual model. To this end, we integrate a structure resembling the human long term memory into our discourse model. The LTM represents the complete knowledge a discourse participant of a particular social role and status is supposed to *know*.

The paper is organized as follows: In the next

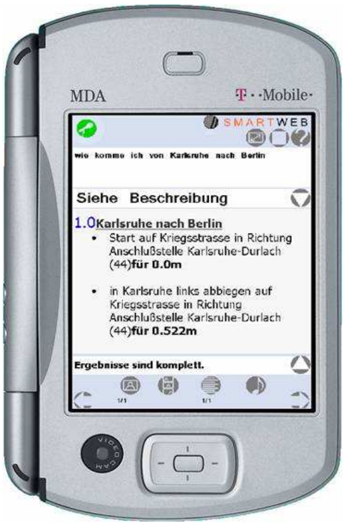
---

[*] This research is funded by the German Ministry of Research and Technology (BMBF) under grant 01IMD01A and by the EU 6th Framework Program under grant FP6-506811 (AMI). The responsibility lies with the authors.

section we will give a brief overview of the system within which this approach is being developed. Then we describe our context model in section 3. In section 4 we will detail how references to implicitly activated concepts are resolved in this framework.

2 System Context

In this paper, we use the Question-Answering system SmartWeb (Reithinger et al., 2005). However, our model including the approach described here is used in other systems as well. SmartWeb is a mobile, multimodal interface to the semantic web. The user is able to ask open-domain questions to the server-based dialog system via a smartphone. Primary input modalities are speech and pen input that operate in a continuous recognition mode. There are also a camera based on-/off-view detection and a prosody-based on-/off-talk detection that prevent the system from processing user contributions that were not intended to address the system. Figure 1 depicts an example configuration of the system's graphical user interface.



wie komme ich von Karlsruhe nach Berlin

Siehe Beschreibung

1.0 Karlsruhe nach Berlin

- Start auf Kriegsstrasse in Richtung Anschlußstelle Karlsruhe-Durlach (44)**für 0.0m**
- in Karlsruhe links abbiegen auf Kriegsstrasse in Richtung Anschlußstelle Karlsruhe-Durlach (44)**für 0.522m**

Ergebnisse sind komplett.

Figure 1: Graphical user interface of the SmartWeb system.

An important aspect of SmartWeb is its context-aware processing strategy. All recognized user actions are processed with respect to their situational and discourse context. A user is thus not required to pose separate and unconnected questions. In fact, they might refer directly to the situation, e. g., , "*How do I get to Berlin from here?*", where *here* is resolved via GPS information, or to previous questions (e. g., "*And in 2002?*" in the context of a previously posed question "*Who won the Fifa World Cup in 1990?*"). The interpretation of user contributions with respect to their context is performed by a component called *Fusion and Discourse Engine*—FADE (Pfleger, 2005). FADE integrates the verbal and nonverbal user contributions into a coherent multimodal representation and enriches it with contextual information, e. g., resolution of referring and elliptical expressions.

The basic architecture of FADE consists of two interweaved processing layers (see figure 2): (i) a production rule system—PATE—that is responsible for the reactive interpretation of perceived monomodal events, and (ii) a discourse modeler—DiM—that is responsible for maintaining a coherent representation of the ongoing discourse and for the resolution of referring and elliptical expressions. Both processing layers operate on the same working memory. This paper deals with processes that are related to the discourse modeler of FADE, for a comprehensive overview of FADE see (Pfleger, 2005).

2.1 Knowledge Representation

All knowledge in the SmartWeb system is represented by means of a single system-wide ontology *SWINTO*. This ontology integrates central concepts of SUMO (Niles and Pease, 2001) and DOLCE (Gangemi et al., 2002) and covers a set of sub-ontologies like a sportevent ontology, a navigation ontology, a media representation ontology, a linguistic ontology, a discourse ontology, etc. All data exchanged within SmartWeb's dialogue component is represented by means of RDF (resource description framework; see `http://www.w3.org/RDF/`) instances of concepts of the SWINTO ontology.

In this paper we focus on the sportevent subontology. This ontology covers a great number of instances describing the Fifa World Cups since 1954: knowledge about all players and countries that are somehow related to the World Cups, all the games including events like goals, yellow and red cards, etc.

2.2 Sample Dialogues

We will discuss two sample interactions with the SmartWeb system: The first example illustrates how our approach deals with implicit knowledge while the second shows how the physical context influences the activation process and thus the resolution of referring expressions.

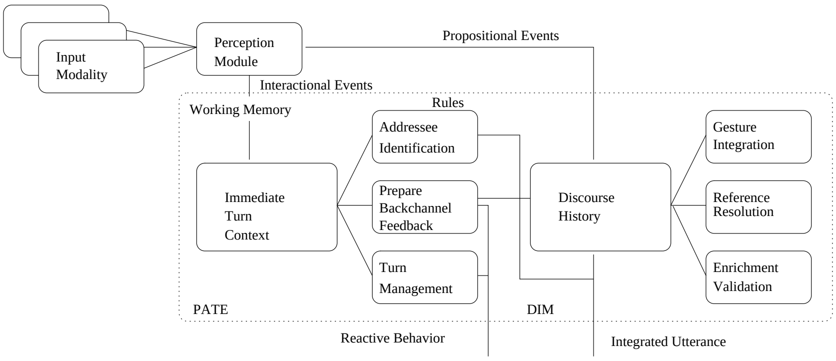


Figure 2: A schematic overview of FADE and its basic functionality.

In example 1 the user's first question sets the context for the interpretation of the second one. A possible context-free interpretation of the second contribution is *"How many goals did Ziege score in his entire career?"*. However, in the given context, the intended interpretation is rather *"How many goals did Ziege score in this particular game?"*.

(1) **User:** *"Wie ist das Spiel Deutschland gegen die USA bei der WM 2002 ausgegangen?"*
(*"How did the game Germany against USA during the World Cup 2002 end?"*)

System: *1:0.*

User: *Wieviele Tore hat Ziege geschossen?*
(*"How often did Ziege score?"*)

The intended interpretation is settled by the fact that the football player "Ziege" was a member of the German team and participated in that particular game. Thus, what we need is to establish a link between the instance representing the FOOTBALLPLAYER Ziege, which is embedded in an under-specified FIELDMATCHFOOTBALLPLAYER and the already activated instance of the class FIELDMATCHFOOTBALLPLAYER of the game mentioned in the previous user turn (see instance *Ziege_1149* in figure 3). This link can only be established if not only the game itself is activated but also related concepts like all the players that participated in that game, the location where the game took place, etc. are activated as well.

Example 2 shows the need for making concepts accessible not only through relatedness to verbally mentioned concepts but also to graphical—or even physical—objects.

(2) **User:** *"Wer war im Finale der WM 2002?"*
(*"Who was in the World Cup final in 2002?"*)

System: [Displays pictures of the two finalists France and Brazil]
(The user selects a picture of the Brazilian team and looks at it)

User: *"Wie heißt der Torwart?"*
(*"What's the goalkeeper's name?"*)

The user refers with *the goalkeeper* to a person that has not been explicitly introduced into the discourse yet but is visible on a picture of the Brazilian team (which is in the current focus of attention of the user). Again, for the resolution of the referring expression we need access to concepts that are somehow related to the pictures.

3 A Comprehensive Context Model

The architecture of our context model is centered around the idea of two interweaved memory representations: (i) a working memory (WM) where the actual processing of contributions takes place and (ii) a long term memory (LTM) that serves as the central knowledge repository of the system. Vital for the work described here is that every concept has a certain amount of activation in the codomain $[0, 1]$ defining its accessibility.

The LTM and WM are separated by a threshold (see figure 3). All instances whose activation is below the threshold belong to the LTM and vice versa. Thus, the LTM comprises all the instance knowledge of the system that is not directly accessible and the WM comprises all those concepts that have been explicitly and implicitly mentioned in the previous discourse. An increase in activation is not only triggered by verbal reference to a knowledge base entry but also by nonverbal ac-

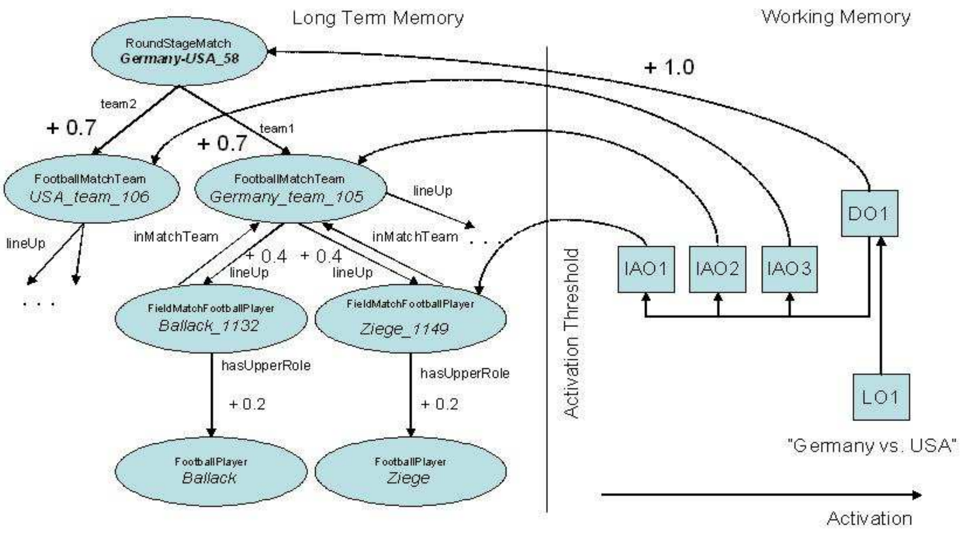


Figure 3: Basic organization of the context model.

tions (e. g., pointing and iconic gestures, etc.) or by physical presence.

3.1 Representation of Knowledge Chunks

All data of the SMARTWEB system is represented by means of instances of ontological concepts. This representation, however, is not sufficient for our purpose as we need a more expressive representation in order to represent the activation of the individual instances. To this end we take a slightly different view on the data and convert every incoming structure into a typed feature structure (TFS; see (Carpenter, 1992)) like representation. Basically, this extended TFS-based representation has the same expressive power as the RDF instances but supports activation values that are directly associated with an individual instance and supports unification and overlay (Alexandersson and Becker, 2003; Pfleger et al., 2002). For a more detailed description of this extended TFS implementation see (Kempe, 2004).

3.2 Organization of the Working Memory

Following the *three-tiered discourse representation* of (LuperFoy, 1991), our discourse modeler comprises two main layers: (i) a modality object layer—extending its linguistic layer and (ii) a concept layer—extending its discourse layer. The third layer—the knowledge base or belief system—corresponds in our approach to the LTM.

3.2.1 Modality Layer

The objects of the modality layer provide information about the surface realizations of objects at the concept layer that have been introduced into the discourse. Thus, modality objects describe the circumstances that caused the increase in activation of their corresponding concept object. The modality layer consists of three classes of objects reflecting the modality by which the corresponding working memory element was referenced: (i) *linguistic actions*, (ii) *nonverbal actions*, and (iii) *physical events*.

Linguistic Actions Linguistic actions resemble the linguistic objects of (LuperFoy, 1992). They comprise information about the surface realization of a concept like lexical information (the lemma used to reference a concept), syntactical information (e. g., number, gender or case), its realization time, or the type of reference (e. g., definite/indefinite, deictic/anaphoric/partial anaphoric). Each Linguistic Action is linked to exactly one instance of the knowledge base and when this link is established, the referenced object's activation is increased. Linguistic Actions are of particular importance for the resolution of referring expressions as they provide the linguistic information needed to identify co-references on the linguistic level.

Nonverbal Actions Nonverbal Actions represent the nonverbal behavior of the interlocutors that contribute to the propositional content of the utterance (e. g., pointing gestures, iconic gestures, emblematic gestures, but also gaze behavior, or drawings, etc). Nonverbal Actions comprise information about the type of nonverbal action, its start and end time. Nonverbal Actions facilitate the resolution of deictic expressions (e. g., *"What's the name of that [pointing gesture] player?"*).

Physical Events Physical Events describe the appearance or disappearance of objects in the physical environment. They comprise information about the type of the event, when it happened, and about the spatial properties of that object (including its relative position to other objects in the scene).

3.2.2 Concept Layer

Objects at the concept layer provide the link to the concepts of the LTM. Each object at the concept layer (if completely disambiguated) represents a unique instance of a concept of the LTM whose activation value exceeds the threshold. We distinguish three types of objects at the concept layer: (i) Discourse Objects, (ii) Implicitly Activated Objects, and (iii) Physical Objects.

Discourse Objects Discourse Objects are containers for concepts that were directly mentioned during the preceeding discourse. They comprise a unified representation of the semantic information gathered so far. In case a Discourse Object is completely resolved, its unified representation is replaced by a link to the corresponding concept of the LTM. Additionally, it contains a set of links to objects at the modality layer; every time a Discourse Object is mentioned, a new link is added.

Implicitly Activated Objects Implicitly Activated Objects are objects that are related to a Discourse Object. If a Discourse Object accesses a concept in the LTM, the activation of concepts related to it is increased by a dynamic factor which depends on the activation of the superordinated concept and the *strength* of the relation between them. The spreading of activation is a recursive process (see section 3.3).

Implicitly Activated Objects may appear in the WM when their corresponding Discourse Object appears. This happens in case their activation exceeds the threshold. Also, the activation of Implicitly Activated Objects decreases faster than that of Discourse Objects. Consequently they are only accessible for a short time.

Physical Objects Physical Objects represent objects that can be perceived in the visual environment. If a Physical Object is explicitly activated through the mentioning of a Discourse Object, it can serve as a referent for a referring expression. Physical Objects are not only part of the discourse layer but they are also part of a superordinate structure representing the complete physical surroundings by modeling the relations between the physical objects located in a scene (e. g., the grey building is on the left of the blue building, etc.). In our implementation, we treat objects displayed on the screen as Physical Objects.

3.3 Organization of the Long Term Memory

As in the human LTM, the objects represented in our LTM are connected through relations. These relations have also a strength associated with them that defines the proximity between the two connected objects. The left part of figure 3 depicts a small excerpt of such a semantic network. At the bottom of this figure there is a concept representing the German football player *Michael Ballack*. This concept exhibits several connections with other concepts like teammates, or the German national team that participated in the game against the USA in the 2002 World Cup.

3.3.1 Lookup and Retrieval of Concepts of the LTM

The concepts of the LTM are stored directly in the type hierarchy, i. e., every type also provides a storage position for its concepts. This means a concept of Type A is stored directly at that type and can be retrieved in turn via that type. Thus, lookup and retrieval of concepts of the LTM is comparatively cheap since the type of the search pattern already restricts the search space to a restricted subset of the complete knowledge base. The actual matching is done by unifying the search pattern with the individual concepts stored for that type. Note that since the lookup is typically based on an under-specified concept, it might return more than one match. In such a case the reference resolution algorithm must deal with this ambiguity.

3.3.2 Activation Propagation

An important aspect of our LTM is that every object has an activation value defining its accessibility. The higher the activation value, the easier it is to access the object (i. e., to retrieve the object from the complete set of knowledge). To account for the activation of neighboring concepts that can be observed in human interactions, the activation of a knowledge chunk is passed on to its associated chunks by a process called *spreading activation*. Spreading activation doesn't only mean that each connected object receives part of the activation of its neighbors but also that it spreads its own activation on to its own neighbors.

An important aspect of a spreading activation model is that activation may spread not only to directly related concepts but also from those concepts to concepts further away in the memory network. This is called the *multi-step* assumption as opposed to the *one-step* assumption that predicts spreading activation only between directly related concepts. In an experimental study, (Sharifian and Samani, 1997) found evidence that also supports the multi-step assumption and the assumption that the activation reduces as it traverses intermediate concepts.

Another aspect of a spreading activation network is the amount of activation that is passed on from one concept to another. Typically, this is controlled by means of strength values associated with the connections (slots) between concepts.

Both the reduction of activation passed from one node to the next and the strength values of connections are important features that influence the behavior of such a network. Therefore, we are currently in the process of developing an empirical method to gather this information for specific domains (see section 5).

3.3.3 Activation Decay

In order to reflect the processes of the human memory, the activation of a concept fades out in time. This means that the longer a concept has not been referenced, the lower its activation will be. Eventually, when the concept's activation is below the threshold, the corresponding object at the concept layer of the working memory will disappear so that the object is no longer directly accessible. However, the activation of an instance will never get below its basic activation.

The three different objects at the concept layer exhibit different intensities in activation decay. Implicitly Activated Objects show the most rapid decay, followed by Discourse Objects. The activation of a Physical Object normally remains on a level that is above the threshold where objects disappear from the working memory.

3.3.4 Current Settings

In our current implementation, the basic activation is $BA = 0.2$. An explicitly mentioned concept receives an increase in activation of 0.7—the connection strength. Our spreading activation algorithm uses this number for multiplying the activation of the related concepts and stops when the result is below the threshold. The threshold is currently set to 0.4. Note, that some concepts might receive an activation between the basic activation and the threshold, in which case they remain slightly more activated in the LTM. Eventually, they might, due to repeated mentioning of related concepts, qualify for the WM. Clearly, these numbers are nothing but heuristics; but as indicated in section 6 we strive for more natural and elaborated numbers.

In the current implementation, the search for referents in the WM is cheap since there are—in our experiments so far—well below hundred. For lookups in the LTM, even in the case of 100k concepts, we have a response time of less than 250 ms.

4 Activation-Based Reference Resolution

Our reference resolution approach differs from standard approaches for reference resolution in one major aspect, see, e. g., (Jurafsky and Martin, 2000) for an overview. A standard reference resolution algorithm initially computes a candidate list of potential antecedents. However, since our context model is self-organizing with regard to the activation of the concepts, there is no need to compute this list because it is always accessible and ordered. Thus, our algorithm takes the existing list and tries to narrow it down by using the linguistic features of the referring expression thereby looking for compatible semantic representations. Note that due to the decay in activation, the concept layer comprises only those candidates that have either been mentioned recently or multiple times. Since we focus here on the resolution of named entities and NPs that might refer to implicitly activated objects, the description of the algorithm will be focused on the resolution of these references.

Given such a reference, the algorithm traverses the objects of the Concept Layer several times until a match is found. In the first run it assumes a reference to an explicitly mentioned discourse object or a physical object. Only if there was no matching object (i. e., an object whose syntactic information is compatible and whose semantic representation unifies with the referring expression) it starts a second run through the list, now focusing on Implicitly Activated Objects. This search continues until the algorithm encounters an Implicitly Activated Object whose semantic representation is unifiable with the one of the referring expression. Finally, if no matching object has been found, there is a third run assuming a discourse reference which we will not discuss here due to spatial restrictions.

4.1 Revisiting Example 1

In the second contribution, the user mentions the named entity *Ziege* that had not been introduced until then. However, the previous reference to the game Germany against USA activated the German team of that game which in turn activated all players that were members of that team. Among these activated players is also our target referent, an instance describing the FIELDMATCHFOOTBALLPLAYER “Ziege” (as depicted in figure 3).

Because of the intra-sentential context of the second user utterance (*“How often did Ziege* **score***?”*), the speech analysis component of SMARTWEB will come up with an interpretation for the named entity “Ziege” where the instance of FOOTBALLPLAYER is enclosed by an unresolved instance of a FIELDMATCHFOOTBALLPLAYER (see the figure 4). Given this interpretation of the named entity, the reference resolution algorithm of FADE fails to find a matching Discourse Object or Physical Object in the Concept Layer. In the second run FADE encounters the implicitly activated instance of the FIELDMATCHFOOTBALLPLAYER *Ziege_1149* that is unifiable with the semantic interpretation of the named entity.

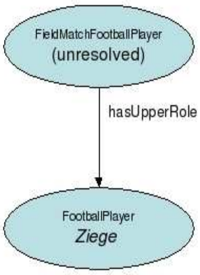


Figure 4: Analysis result for the named entity *Ziege*.

5 Discussion

The application of associative networks and spreading activation for the identification of named entities or the resolution of lexical ambiguity has a long research tradition in psychology and artificial intelligence, see e. g., (Kintsch, 1988; Hirst, 1988). However, in practical dialogue systems this type of contextual information has, to our knowledge, not been applied yet.

Indeed, there exists a number of comprehensive models for the resolution of referring expressions, e. g., (LuperFoy, 1991; LuperFoy, 1992; Allen et al., 2000; Allen et al., 2001). However, all of these models lack the inclusion of implicitly activated concepts into their model of the ongoing discourse.

Our model is best compared to that of (Allen et al., 2000; Allen et al., 2001). There, an architecture for implementing interactive conversational agents is described. We relate our work to their discourse and reference module. In their discourse module, five types of information are present: (i) salient entities, (ii) preceeding utterance, (iii) turn status, (iv) discourse history, (v) discourse obligations.

In contrast to the TRIPS architecture, in our module the salient entities (i) are extended with the activated referents which allow for an interpretation already in the discourse module. Moreover, our context model (see section 3) includes a rich discourse history ((iv), (ii)) where preceeding utterances—along with information such as speech act—is just one part. To a certain extent, their model is able to deal with implicitly mentioned concepts too, but whereas our model utilizes the ontology, their model relies on the plans.

6 Conclusion and Future Work

We have presented a cognitively motivated comprehensive discourse model that mimics the behavior of humans by means of a Long Term Memory (LTM) and a Working Memory (WM). We have shown how *explicitly* mentioned objects are activated and how their activation exceeds a threshold—the edge of consciousness—transfered from the LTM to the WM. Once a concept is in the WM, its activation decreases as time goes by until its activation falls below the threshold and then vanishes from there again. Focus, of this paper has been to show that by using spreading activation, the activation of *implicitly* mentioned concepts increases and when their activation exceeds the threshold that they are transferred into the WM. This enables the interpretation of natural utterances as humans produce them.

6.1 Future Work

As pointed out, our current implementation of the spreading activation process is based on handmade numbers and this, of course, is not feasible in the long run. We therefore recently started to work on an empirical method for measuring the strength of connections between concepts for a given domain. Currently, we aim at a combined experiment that will provide not only information about the strength of connections but also about the frequency measures for particular instances that can be used to compute the basic activation of instances. The ultimate goal is to define a set of experiments and post-processing steps so that we will be able to automatically extract the connection strengths between related concepts.

Moreover, we are currently investigating to what extent it is possible to apply some kind of online-learning functionality for adjusting the connection strength and the introduction of completely new connections/associations between previously unrelated concepts. If, for example, in the course of interactions two unrelated concepts *A* and *Z* appear frequently in the same context, these two concepts will be connected. This means that the mentioning of concept *A* will in the future activate concept *Z*.

Future work will also include the incorporation of implicitly activated concepts that are part of plans or scripts.

References

Jan Alexandersson and Tilman Becker. 2003. The formal foundations underlying overlay. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 22–36, Tilburg, The Netherlands.

James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering special issue on Best Practices in Spoken Language Dialogue Systems Engineering*, 6(3), December.

James Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of Intelligent User Interfaces 2001 (IUI-01)*, Santa Fe, NM, January 14–17.

Bob Carpenter. 1992. *The logic of typed feature structures*. Cambridge University Press, Cambridge.

Herbert H. Clark. 1977. Bridging. In Philip N. Johnson-Laird and Peter C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 169–174. Cambridge University Press, Cambridge, England.

Herbert H. Clark. 1996. *Using language*. The Press Syndicate of the University of Cambridge.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. 2002. Sweetening ontologies with DOLCE. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, volume 2473 of *Lecture Notes in Computer Science*, pages 166–182, Siguenza, Spain, Oct. 1–4.

Graeme Hirst. 1988. Resolving lexical ambiguity computationally with spreading activaiton and Polaroid Words. In *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*, pages 73–107. Morgan Kaufmann Publishers, San Mateo, CA.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Lanuage Processing*. Prentice Hall.

Benjamin Kempe, 2004. *PATE – a production rule system based on activation and typed feature structure elements*. http://www.dfki.de/˜kempe/pate.pdf.

Walter Kintsch. 1988. The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. *Psychological Review*, 95(2):163–182.

Susann LuperFoy. 1991. *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions*. Ph.D. thesis, University of Texas at Austin, December.

Susan LuperFoy. 1992. The Representation of Multimodal User Interface Dialogues Using Discourse Pegs. In *Proceedings of the* 30th *Annual Meeting of the Association for Computational Linguistics - (ACL-92)*, pages 22–31, Newark, USA.

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.

Norbert Pfleger, Jan Alexandersson, and Tilman Becker. 2002. Scoring functions for overlay and their application in discourse processing. In *KONVENS-02*, Saarbrücken, September – October.

Norbert Pfleger. 2005. Fade - an integrated approach to multimodal fusion and discourse processing. In *Proceedings of the Doctoral Spotlight Session of the International Conference on Multimodal Interfaces (ICMI'05)*, pages 17–21, Trento, Italy.

Norbert Reithinger, Simon Bergweiler, Ralf Engel, Gerd Herzog, Norbert Pfleger, Massimo Romanelli, and Daniel Sonntag. 2005. A look under the hood: design and development of the first smartweb system demonstrator. In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*, pages 159–166, New York, NY, USA. ACM Press.

F. Sharifian and R. Samani. 1997. Hierarchical spreading of activation. In F. Sharifian, editor, *Proceedings of the Conference on Language, Cognition, and Interpretation*, pages 1–10. IAU Press.

Understanding student input for tutorial dialogue in procedural domains

Myroslava O. Dzikovska*, **Charles B. Callaway***, **Matthew Stone**†,*, **Johanna D. Moore***
*Human Communication Research Centre, University of Edinburgh,
Edinburgh, EH8 9LW, United Kingdom
{mdzikovs,ccallawa,jmoore}@inf.ed.ac.uk
† Department of Computer Science, Rutgers University, Picataway, NJ, 08845-8020
mdstone@cs.rutgers.edu

Abstract

We present an analysis of student language input in a corpus of tutoring dialogue in the domain of symbolic differentiation. Our focus on procedural tutoring makes the dialogue comparable to collaborative problem-solving (CPS). Existing CPS models describe the process of negotiating plans and goals, which also fits procedural tutoring. However, we provide a classification of student utterances and corpus annotation which shows that approximately 28% of non-trivial student language in this corpus is not accounted for by existing models, and addresses other functions, such as evaluating past actions or correcting mistakes. Our analysis can be used as a foundation for improving models of tutoring dialogue.

1 Introduction

In domains from mathematics to maintenance, human tutors often instruct students by coaching them through procedures they must learn. There seems to be a natural analogy between this activity and people's collaborative problem solving (CPS) when they jointly pursue real-world goals. For example, in both cases, interlocutors must talk through what they have accomplished so far, what to do next, and how to do it. The analogy has prompted tutoring researchers such as Rickel et al. (2001) to model procedural tutoring in terms of existing models of CPS.

This research program promises to enrich tutorial dialogue systems by leveraging a rich research tradition (Grosz and Sidner, 1986; Grosz and Kraus, 1996; Lochbaum, 1998; Blaylock and Allen, 2005). But it assumes that students and tutors use the same kinds of utterances to make the same kinds of moves as found in previously studied collaborative domains. In this paper, we offer an empirical assessment of this assumption.

We report an analysis of a corpus of dialogues for tutoring a mathematics procedure, symbolic differentiation, which has been annotated reliably for a coarse categorization of student behavior. Our analysis suggests that students and tutors work together to maintain a shared understanding of the procedure being carried out and learned, so in broad terms, the CPS model of procedural tutoring is a good one. However, we find that about 28% of student utterances are not covered explicitly by existing models of procedural tutoring and CPS. CPS research seems to have overlooked much of the talk that interlocutors use to reach common ground not just about what they are *going to do* but about what they *have done*. Such moves may be particularly frequent in procedural tutoring because it aims at student understanding, and must accommodate the errors students make while learning.

Our work is informed by the long-term goal of building interactive natural language dialogue systems that reproduce the increased effectiveness of one-on-one human tutoring over classroom instruction (Bloom, 1984). Crucially, from this perspective, the new moves we identify not only occur frequently, but also seem to offer useful information about what the student has learned and what feedback they need — information that would not be available just from student actions or even moves in earlier CPS models. So we advance the development of richer formal models as a challenge for future theoretical and applied research in tutorial dialogue.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 10–17.

We begin by describing in Section 2 our corpus and the domain model. In Section 3 we discuss our annotation scheme, how it relates to the Shared Plan model (Grosz and Sidner, 1986; Grosz and Kraus, 1996) as well as to specific implementations of CPS for tutoring, particularly that of Rickel et al. (2001), and the role different student inputs play in tutoring. In Section 4 we discuss the corpus annotation and analysis, concluding with related and future work in Section 5.

2 Corpus and Tutoring System

2.1 Corpus Collection

The task domain we used in this study is tutoring symbolic differentiation. The task involves applying a set of differentiation rules (the chain rule, the product rule, the sine rule etc.), and the algebraic transformations to bring the result to a normal form. We analyzed a corpus of interactions between students and experienced tutors in this domain, which we are using to study tutorial dialogue and to build a tutoring system.

The data consists of 19 transcripts of 30-minute tutoring sessions conducted via a chat interface. During the session, the tutor gave the student a sequence of problems on using the chain rule until they ran out of time. The student did not propose or choose the problems.[1] Two experienced mathematics instructors (as tutors) and 14 first-year mathematics or science undergraduate students who were learning differentiation in a calculus course at the University of Edinburgh were paid to participate.

The data collection environment separated students from tutors physically. They could only exchange utterances via a chat interface where interlocutors could send each other text messages. Complex mathematical expressions could be entered using a special editor, and text and formulas could be intermixed. The tutor could observe the student's actions in real-time on a second screen. Students and tutors were trained to use the interfaces prior to the data collection session.

The resulting corpus consists of 19 dialogues (5 students returned twice) and contains 1650 utterances (with textual and symbolic parts), 5447 words and 559 formulas.

2.2 Domain Model

Based on the corpus analysis, we built a model of differentiation with the chain rule which instantiates possible actions in our domain. Our model consists of 5 basic steps: 1) rewriting the function to a recognizable form which can be matched against a differentiation rule; 2) differentiating the outer layer; 3) differentiating the inner layer; 4) combining the results by multiplying; and 5) simplifying the resulting expression. To support tutoring, the actions applicable at each step may take the form of either applying a correct differentiation rule, or else using an incorrect (buggy) rule (Brown and Burton, 1978).

This domain model integrates both correctly executed actions and errors as possible steps in problem-solving. However, when a student enters a formula, it may be ambiguous with respect to which of the steps in the domain model it matches, especially in the presence of errors. Moreover, the students do not always know all the necessary steps. For example, they frequently do not realize that the resulting expression after step 4) needs to be simplified. Student language may provide additional important clues to resolve such ambiguities, as discussed in Section 3.

We implemented a prototype system, BEEDIFF (Callaway et al., 2006) with a domain reasoner which supports the task model described above (Zinn, 2006) without language input. The domain reasoner acts as a plan recognizer by fitting submitted formulas with the task model. One of the goals of our corpus analysis is to identify the types of student utterances which occur frequently in our corpus, for example, help requests, and in the subsequent analysis to identify the appropriate pedagogical and dialogue strategies to use in response in our dialogue system.

3 Annotation Scheme

The goal of our corpus analysis is to identify student language in tutorial dialogue consistent with existing CPS models, as well as the language which is not accounted for directly. In our comparison we focus on the Shared Plan model of discourse (SP model) (Grosz and Sidner, 1986) and its implementation in the COLLAGEN PACO tutoring agent (Rickel et al., 2001). PACO implements Lochbaum's (1998) intention recognition algorithm which is based on the SP model. The implementation supports a subset of an artificial

---

[1] This restriction was not explicitly part of the experimental design, but in practice students did not attempt to choose or negotiate on problems to solve.

T: Differentiate $sqrt(x^3 - 9x)$

S: $(x^3 - 9x)^{-1}$
is this equal to the question (**help-specific**)

T: No. Remember that $\sqrt{(z)} = z^{1/2}$ Can you rewrite the question now?

S: would that be $(x^3 - 9x)^{1/2}$ (**task-progression**)

Figure 1: A sample dialogue with student utterances containing language with our annotations.

CPS language proposed by Sidner (1994), which gives an account of beliefs and intentions which occur in utterances modeled by the SP model.

Blaylock and Allen (2005) provide another CPS model of task-oriented dialogue. The model is broadly compatible with the SP model, but focuses primarily on interlocutors' collaboration in negotiating tasks and resources amid task execution. It does not directly address how explanations and assessment of completed steps fits into the dialogue, and has not been extended to tutoring, hence our main focus on the SP model and PACO.

The classes of utterances implemented in PACO are agreeing and disagreeing, proposing tasks, actions and goals, indicating that a task has been accomplished, asking about or proposing task and action parameters, asking and proposing how tasks should be accomplished, and asking what to do next. By contrast, our classification differentiates the following kinds of student behavior: help requests; queries about next steps; comments on cognitive state (*i.e.* explaining what they are doing, or stating their knowledge or beliefs related to the problem); progress evaluation; and more general dialogue behavior related to agreement and clarification. Our model is deliberately coarse-grained in order to explore the properties of the dialogue. The categories thus indicate directions for future specification and implementation, rather than moves that could be directly formalized in a detailed model of tutorial dialogue. A fragment of annotated dialogue is shown in Figure 1.

Table 1 presents a correspondence between the language in the PACO model and our classifications, which are discussed in more detail below. As the table shows, the SP model, along with its implementation in PACO, provides ways of thinking about and representing discourse that can describe a wide range of student utterances in tutoring. However, the tutoring setting makes available to students a variety of utterances whose form, content or function differs from those more traditional collaborators might typically use.

Help requests. These are utterances that indicate that the student does not know what the next step is, or does not know how to perform a specific step (or perhaps is not confident enough to perform it). We subdivided help requests into generic and specific requests. Generic requests indicate that the student is stuck, for example, "I don't know what to do", but don't provide further clues as to what the problem is. Specific help requests identify explicitly what the students do not know, *e.g.*, "I don't know what the common factor is", or "$z = cos(x)$. I don't know about y".

Generic help requests correspond to asking what to do next in the CPS model. Specific help requests correspond either to asking about task and action parameters, or to asking about the appropriate recipe (how to perform a given step).

Verifying next steps and goals. Instead of doing the step directly, students often describe it first (without indicating the end result), in order to get confirmation from the tutor, for example, "Should I multiply the 3 and the 15", or "Should I simplify this further?" This utterance category corresponds to proposing steps in the CPS model. However, the language is quite different compared to that typically seen in CPS, where proposals are normally offered up for negotiation, *e.g.*, "Let's start engine two". This shows that students are not really negotiating how to select the next step from among a set of possibilities. Rather, students who produce this type of utterance may be uncertain how to proceed, and need the tutor's help.

Clarification requests. As in any dialogue, there are cases where students do not understand what the tutor has said (as opposed to not knowing what to do next, or how to perform a step). Students then attempt to clarify the meaning of the comment, either on the level of terminology, or a more conceptual level. For example, "What do you mean by $3x^2$" or "How can there be an inner layer when there is no power". Clarifications are part of conversational expertise, and should be accounted for by general dialogue models.

Dialogue progression. These are steps students usually take to acknowledge the tutor's instructions, to indicate that they understand the mate-

Student utterance class	PACO language type
Generic help request	asking what to do next
Specific help requests	asking about task and action parameters, or how tasks should be accomplished
Next step verification	Proposing steps, goals, action parameters or recipes
Dialogue progression	Agreeing and disagreeing
Progress evaluation	Consistent with the SP model, but not explicitly in PACO
Explanations	Indicating which task was accomplished, other kinds not modeled in detail
Task progression	Consistent with the SP model, but not explicitly in PACO
Stating their knowledge	Consistent with the SP model, but not explicitly in PACO
Editing the solution in language	Can be treated as just performing the step directly

Table 1: Correspondences between our coding scheme and language categories in (Rickel et al., 2001)

rial,[2] and in general to establish that the tutor has been understood and to advance the dialogue. We expect that these dialogue moves can be covered by a general model of collaborative dialogue, because they are not in any way specific to tutoring.

Progress evaluation. Students often either evaluate their own progress (*e.g.*, "I made a mistake"), or ask the tutor to evaluate their progress (*e.g.*, "Is this right?").[3] These utterances are consistent with the SP model, but in PACO only tutors' evaluations of students' progress are included, and they are treated as accepting or rejecting the student actions. This is not appropriate for students evaluating their own progress, or asking for evaluation.

One may think progress evaluation is an example of grounding (Clark and Brennan, 1991), rather than a specific CPS move. However many other CPS moves actually let interlocutors show or check that they have achieved mutual understanding. Moreover, modeling progress evaluation is not just a matter of allowing discussion at set points in plan-execution. Progress evaluation suggests that in tutoring, unlike domain-oriented problem solving, tutors allow students to make mistakes and expect that they may not necessarily recognize problems. The whole CPS process for tutoring therefore has to underconstrain actions to include errors and underconstrain context to allow for ignorance, and has to be more explicit about how progress is evaluated, including allowing students to ask for evaluation or give it themselves.

Explanations. Students may verbalize what they are doing as they are solving the problem ("forward-looking" explanations) or, usually after a mistake, reflect on what they did and why ("backward-looking" explanations). An example forward-looking explanation is the utterance "and put back in the original form $15x^2/3(5x^3-6)^4$". In this case, the formula in the utterance is the re-writing of the previous solution step, which is necessary to finish up the differentiation procedure.[4] An example of a backward-looking explanation is, in reply to the tutor's request "Why did you do that?", "because I thought you multiplied the powers when they were bracketed". Tutors occasionally asked the students how they arrived at their (incorrect) solutions, but sometimes students offered their own explanations spontaneously after the tutor corrected their mistake. Explanations of student actions have also been observed in tutoring algebra proofs (Wolska and Kruijff-Korbayová, 2004).

These explanations are not modeled in sufficient detail by existing CPS models. Sidner's negotiation language contains `provide-support` as a possible action, and the PACO implementation incorporates the tutor giving explanations within recipes specific to tutoring. A student's asking of "why" questions is modeled as a proposal to the tutor to provide support for what should be done. These models only describe utterances that motivate actions *before* they are agreed on, while our explanations often accompany actions as they are

---

[2] Saying "I don't understand" would usually be classified as a help request unless it is related to surface form of tutor's words

[3] We classified as progress evaluation only the items which were "content-free" and could be covered with generic "oops" and "am I right?" buttons. If the student described the problem in more detail, the utterance was classified as an edit or an explanation.

[4] If the language was not present, the formula by itself would be counted as performing the re-writing step.

done or afterward.

Explanations should be modeled more explicitly as part of the tutoring process. Forward-looking explanations can be seen as behaviors that disambiguate the place of an ongoing action in the plan. While typically collaborators agree on actions before they do them, in some cases one collaborator may decide to act independently to further the joint activity. In such cases, the actor may have to describe the action for their collaborators to recognize the step they are performing. We are not surprised to see this more prominently in tutoring than in other CPS because students are being coached to carry out the procedure on their own.

Backward-looking explanations occur after the relevant problem-solving step has been completed. We can see why such explanations might further students' and tutors' joint activity. These moves may allow students to provide evidence about their understanding of the rules and relationships involved in problem-solving, although a second possible motive is social (to allow students to save face). Such moves may therefore contribute to patterns of interaction between tutor and student that establish a correct mutual understanding of the subject-matter that students should learn. Again, it is no surprise that such backward-looking activity might be more frequent in learning dialogues than typical CPS dialogues, which simply aim at achieving real-world goals.

Task progress indications. Students often indicate whether they are continuing with the problem, or are finished, with progress markers like "first", or "the final answer is …". These again are consistent with the general SP model where they correspond to cue phrases starting a new discourse segment, under the assumption that each student action starts a new discourse segment which can end immediately when a tutor accepts it, or continue with remediation. Current implementations, however, do not reason about these cues specifically. Similarly to forward-looking explanations, domain reasoning should be sufficient to infer where the student is in executing the task without these markers. However, when the student provides them it may be important for tutoring, because they explicitly indicate where the student thinks she is in the process of solving the problem. Consider the case when the student is differentiating a function and writes $-3 * sin(x)^{-3}$. This expression should be simplified, and it is not always clear if the student thinks that he is done, in which case the tutor needs to remind them to simplify the formula, or if the student is still working on the problem and will simplify on the next step. But if the student says "the final answer is $-3 * sin(x)^{-3}$", then it is a clear indication that he thinks the problem is finished, and the tutor needs to intervene.

Stating knowledge of rules and principles. Students make statements (correct or incorrect) about the rules or principles they know, for example "The derivative of $\sin$ is $\cos$". These usually don't appear by themselves, but are used to support general meta-level tutoring talk, in particular help requests and explanations. We chose to tag these utterances as a separate class because they contain very explicit statements about what students know and believe, as opposed to more indirect indications when students state what they don't know in a help request. In the general SP framework (not explicitly covered in the PACO model) this corresponds to stating or proposing recipes. However, in CPS, proposing recipes is done at the negotiation stage, where different courses of actions are possible to achieve a goal. In a tutorial setting, the function of stating rules is to support tutoring rather than problem-solving per se — it is an attempt from the student to expose their knowledge to the tutor, with the goal of confirming it is complete and correct.

Using language to edit answers. Students sometimes describe a portion of the answer instead of providing the full formula, for example, "Ah, so the top part is $-15x^3$", or correct themselves immediately after supplying the answer (without tutor intervention), *e.g.,* "I meant to put a to the power 6 on the bracket". These utterances can be viewed as doing the step directly in the CPS model in most cases, equivalent to submitting a full formula. Some of these are specific to mathematical dialogue, where the mixture of informal and formal language can be used to describe math expressions (Wolska and Kruijff-Korbayová, 2004). However, self-corrections can be important in other domains, especially if student actions are "non-reversible" (*e.g.,* pressing a button in a simulator). These cases may then require a different strategy on the part of the tutor.

Input not related to tutoring or problem-solving. Task management (e.g, transitions between problems, greetings and closings), and jokes are obviously part of social interaction in any conversation. We expect these to be accounted for with a more general model of conversation, and tagged them as a class of "non-task-related".

4 Corpus Analysis and Discussion

We annotated student input in 19 dialogues with this scheme. There were a total of 656 student utterances. Out of those, 323 (49%) contained only a mathematical formula contributing to the solution and no language. Our annotation was done over the remaining 333 student utterances which contained at least one word. Out of those, 99 (30%) were judged as not relevant to tutoring or the task of differentiation (greetings and closings, transitions between tasks, jokes, etc.). The distribution of tags among the 234 remaining utterances in our corpus is shown in Table 2. To verify the inter-rater reliability, two annotators independently coded four dialogues (102 utterances) with the scheme, resulting in inter-rater agreement of 84% and $\kappa = 0.78$ (*i.e.* 'good' agreement).

The categories of student input which are not directly accounted for in the CPS model (evaluation, explanation and knowledge) together account for about 28% of all student language input. This underscores the importance of including those phenomena in a model of tutoring dialogue. Our categorization is a first step in identifying the phenomena which need to be accounted for in a formal model of tutorial dialogue in procedural domains, which is the next step in our work.

Other questions which arise in this line of research are the importance of the individual categories from the point of view of practical systems, as well as the importance of student language in general in tutoring. Our study contributes to answering these questions.

Many student utterances offer information for student modeling that goes beyond what can be derived from the sequence of steps the students execute. In particular, specific help requests, explanations and knowledge statements give indications of student knowledge and misconceptions, and task progression markers and evaluations may help evaluate student knowledge as well as their confidence level.

These categories cover 38% of student utterances, which can be interpreted as an indicator that dialogue participants considered it important in some way. This is not sufficient to make definitive conclusions about the importance of language in tutoring dialogue, because we do not know if the students who used more language learned more.[5] However, our data contain tutors' assessments of student aptitude, and we plan on investigating if they correlate with the use of language.

Tag	Count	Tag %	Mean	Stdev
Help requests	47	21%		
generic	27	12%	1.42	1.64
specific	20	9%	1.05	1.31
Step requests	19	8%	1.00	1.41
Clarifications	5	2%	0.26	0.56
Dialogue progression	54	23%	2.84	2.14
Edit	15	7%		
forward	9	4%	0.47	0.84
backward	6	3%	0.32	0.58
Evaluation	31	13%		
request	17	7%	0.89	1.20
state	14	6%	0.74	0.87
Explanation	17	8%		
forward	4	2%	0.21	0.42
backward	13	6%	0.68	1.25
Knowledge	15	6%		
global	10	4%	0.53	1.22
problem	5	2%	0.26	0.56
Task progression	5	2%	0.26	0.93
other	26	11%	1.37	2.19

Table 2: Tag distribution in our corpus. Tag % is the percentage of tag occurrences out of the overall tag count; mean and stdev refer to the average number of tag occurrences per dialogue.

Different students used different strategies in their language. For example, the percentage of backward explanations varied from 0 to 27%, and the percentage of specific help requests from 0 to 33%. Thus it is difficult to make predictions comparing the frequencies of individual tags in our corpus. One of the important tasks of further corpus analysis is to determine the cause of this variation, which may be due to individual differences, student aptitude and motivation,[6] or other features

---

[5] This study did not measure learning gains, which is necessary to assess the amount learned by each student.

[6] We observed that poor students generally talked more and were more specific in their requests, which needs to be confirmed with further analysis

of the interaction.

Most of the categories important to student modeling discussed above rely on non-trivial language which is dependent on the context and cannot be covered by a set of simple questions or buttons (specific help requests, step requests, clarification requests, explanations, and knowledge statements). This correspondence again suggests the importance of extending the models to cover these more complex interactions. The next step would be to determine which tutorial strategies would be appropriate in response to each of those utterance classes, and confirm the correlation with corpus analysis.

5 Related Work

Many tutoring systems for procedural tasks have been built around simulation environments (Rickel et al., 2001; Pon-Barry et al., 2004; Ong and Noneman, 2000). These systems use a task model, augmented with plan recognition, to recognize student actions and intentions and provide feedback and directions accordingly, with very limited student language input. For example the NASA RPOT tutor (Ong and Noneman, 2000) is based on a generic task tutor toolkit which contains a task model and can answer 3 questions: "What do I do", "Why do I do that", and "How do I do that". Rickel et al. (2001) accounts for a subset of student input consistent with the artificial CPS modeling language (Sidner, 1994). Our paper continues this line of work by investigating the student language not covered by existing models.

Shah et al. (2002) provide a model of student initiatives and tutor's responses in the CIRCSIM system (i.e. utterances which go beyond responding to tutor's questions), classifying them along four dimensions. Our classification is closest to their communicative goal dimension, which includes requests for confirmation and for information, challenging the tutor, refusal to answer and conversational repair. Our categorization covers all student utterances regardless of the initiative, with the goal of building model of tutorial dialogue covering the behavior of both dialogue participants, and we intend to study the classes of tutor utterances which are the appropriate follow-ups to student utterances in the future.

A large amount of work in the tutoring literature is dedicated to modeling tutoring strategies, that is, the actions the tutor takes during the interaction (McArthur et al., 1990; Zhou et al., 1999; Pilkington, 1999; Graesser et al., 1999; Pon-Barry et al., 2004). McArthur et al. (1990) propose a model of tutoring in solving algebraic equations in which problems are solved step-by-step according to the task model. Tutors execute "microplans" at each step, which consist of introducing a problem or a step, solving it (done by the student generally), evaluation, remediation if necessary, and an optional wrapup step where the tutor may summarize the step or the problem. This model is "tutor-centric" in the sense that it does not account for the student's actions. Our goal is to develop a model of both student and tutor behavior, which can be used to inform the implementation of a tutorial dialogue system.

Much research has been done in identifying what makes human-human tutoring effective. Self-explanation (Chi et al., 1989), interactivity (VanLehn et al., to appear), and student initiative and "student talk" (Core et al., 2003) have been studied as possible predictors of student learning. But there is currently no definitive study confirming that talking in natural language, and specifically what kind of language, improves learning compared to reading or limited forms of input such as multiple-choice answers.

We are currently conducting a study to evaluate the role of student language input in tutoring. We analyzed a corpus of human-human tutoring dialogues (in a conceptual domain) where the teaching material was designed to elicit a different amount of student language under different conditions. This will allow us to see if there is a correlation between the amount and type of language students use and learning gains. Additionally, we are considering a study in our current domain comparing tutoring with free language input to tutoring where students are only allowed to input formulas and have a small set of buttons to ask for help and confirm or disconfirm their understanding of what the tutor said. We are planning to use these studies to gain further understanding of the role of student natural language input in learning from tutoring.

6 Conclusions

We provided a description of student language in a corpus of procedural tutoring which can serve as an initial model for implementing a tutorial system. We identified student language categories which are not sufficiently modeled in the existing

CPS model, and showed that 28% of student utterance in our corpus fall under those categories. All of those classes fall under the categories of utterances which may be important to student modeling. We argue that the existing CPS models need to be extended to cover these classes of utterances in tutoring dialogue. Our scheme provides an initial categorization of phenomena which need to be included in formal models, as well as working dialogue systems to account for student language in addition to actions.

Acknowledgments

Helen Pain and Kaśka Poryska-Pomsta collected the data used in this analysis. This work is supported by a grant from The Office of Naval Research number N000149910165, European Union 6th framework programme grant EC-FP6-2002-IST-1-507826 (LeActiveMath), NSF HLC 0308121, the Leverhulme Trust and Rutgers University.

References

N. Blaylock and J. Allen. 2005. A collaborative problem-solving model of dialogue. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 200–211, Lisbon, September.

B. S. Bloom. 1984. The two sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:3–16.

J. S. Brown and R. R. Burton. 1978. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science: A Multidisciplinary Journal*, 2(2):155–192.

C. Callaway, M. Dzikovska, C. Matheson, J. Moore, and C. Zinn. 2006. Using dialogue to learn math in the LeActiveMath project. In *Proceedings of the ECAI Workshop on Language-Enhanced Educational Technology*, pages 1–8, August.

M. T. H. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.

H. H. Clark and S. E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, D.C.

M. G. Core, J. D. Moore, and C. Zinn. 2003. The role of initiative in tutorial dialogue. In *Proceedings of EACL-03*, pages 67–74.

A. C. Graesser, P. Wiemer-Hastings, P. Wiemer-Hastings, and R. Kreuz. 1999. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1:35–51.

B. Grosz and S. Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

B. J. Grosz and C. L. Sidner. 1986. Attentions, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

K. E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

D. McArthur, C. Stasz, and M. Zmuidzinas. 1990. Tutoring techniques in algebra. *Cognition and Instruction*, 7:197–244.

J. C. Ong and S. R. Noneman. 2000. Intelligent tutoring systems for procedural task training of remote payload operations at NASA. In *I/ITSEC 2000*.

R. M. Pilkington. 1999. Analysing educational discourse: The discount scheme. Technical Report 99/2, Computer Based Learning Unit, Univ. of Leeds.

H. Pon-Barry, B. Clark, K. Schultz, E. O. Bratt, and S. Peters. 2004. Advantages of spoken language interaction in dialogue-based intelligent tutoring systems. In *Proceedings of ITS-2004*, pages 390–400.

J. Rickel, N. Lesh, C. Rich, C. Sidner, and A. Gertner. 2001. Building a bridge between intelligent tutoring and collaborative dialogue systems. In *Proceedings of AIED-2001*, pages 592–594.

F. Shah, M. W. Evens, J. Michael, and A. Rovick. 2002. Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, 33(1).

C. L. Sidner. 1994. An artificial discourse language for collaborative negotiation. In *Proceedings of AAAI*, pages 814–819.

K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé. (to appear). When are tutorial dialogues more effective than reading? *Cognitive Science*.

M. Wolska and I. Kruijff-Korbayová. 2004. Analysis of mixed natural and symbolic language input in mathematical dialogs. In *Proceedings of ACL-04*, pages 25–32.

Y. Zhou, R. Freedman, M. Glass, J. A. Michael, A. A. Rovick, and M. W. Evens. 1999. What should the tutor do when the student cannot answer a question? In *12th FLAIRS Conference*, pages 187–191.

C. Zinn. 2006. Supporting tutorial feedback to student help requests and errors in symbolic differentiation. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, pages 349–359.

A multi-speaker dialogue system for computer-aided language learning

Peter Vlugter and Alistair Knott
Dept of Computer Science, University of Otago

Abstract

The main topic of this paper is how to configure a dialogue system to support computer-aided language learning. The paper also serves to introduce our new multi-speaker dialogue system, and highlight some of its novel features.

1 Introduction: language-learning dialogues

The dialogue system described in this paper is intended for use in a language-learning application. Computer-aided language learning (or CALL) is an interesting application for dialogue systems, for several reasons. Firstly, from a practical point of view, a language-learner's grammar and vocabulary are smaller than those of a native speaker, and therefore the problem of interpreting and generating unrestricted free text may be simpler in this domain than in other domains. On the other hand, the student's knowledge of dialogue management largely carries over from her native language to the language being learned. Elements such as clarification subdialogues, and checking questions are common in CALL dialogues—perhaps even more common than in ordinary dialogues. Thus a language-learning dialogue might be quite complex at the level of dialogue structure, even if its constituent utterances are syntactically restricted. This makes CALL dialogue an interesting test domain for a practical dialogue system.

Equally importantly, dialogue is a natural medium for language-learning. What the student is learning to do is to speak to other people in the new language, so in talking to the tutor, the student is actually performing the very task to be learned. The tutor essentially plays two roles: firstly, as a conversational partner he engages in a real dialogue, about some chosen topic or task; secondly, as tutor he gives feedback about the student's utterances, and is available to answer questions about the language.

Language-learning dialogues often take place between two participants: a student and a tutor. However, many of the most important elements of a new language being learned are hard to exercise in this envrironment. In any new language, one of the important early topics is the system of **personal pronouns**: referring expressions, or possessives, or relative pronouns. To provide a natural environment for these expressions, more than two participants are needed in the conversation. A similar consideration applies to verb inflections (which in many languages must agree with the person and number of the subject and/or object in the sentence), and addressee terms (which are only needed if there is more than one potential addressee).

In this paper, we will describe a multi-speaker dialogue system which we have configured for use in a CALL environment for learning Māori, the indigenous language of New Zealand. In Section 2, we introduce the theoretical background for the system and the basic dialogue management principles. In Section 3, we describe how the system supports multi-speaker dialogue. In Section 4, we show how the multi-speaker system provides a natural environment for practice with the Māori pronoun system (which is particularly fiendish). In Section 5, we describe how the system implements a set of educational goals for a given interaction, and monitors the progress of these goals by maintaining a student model. In Section 6, we describe how the system-played characters choose their contributions with an educational goal in mind, to create opportunities for the student to acquire targeted constructions. Throughout the paper, we emphasise the topic of personal pronouns, to motivate the need for a multi-speaker CALL environment.

2 Background to the dialogue system

2.1 Semantic framework

Our system, called Te Kaitito, is designed to teach the Māori language. Our treatment of utterance

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 18–25.

syntax and semantics has been described elsewhere (Vlugter *et al.* (2004); Bayard *et al.* (2002)) so we will provide a short summary of these topics here.

The system supports ‘deep’ semantic processing. The user enters her utterances as written text. Each utterance is first parsed, using an HPSG-style grammar, which delivers a semantic representation as output, in the formalism of Minimal Recursion Semantics (MRS; Copestake *et al* (1999)). The MRS representation is converted into a DRT-like format, which represents an utterance as an **context-update** operation. Each update is specified for a **dialogue act**, a **speaker**, an **addressee**, and a **message**. The message comprises a set of **presuppositions** which must be satisfied against a representation of the common ground, and a **nucleus**, which holds the propositional part of the utterance, whose force is different for different dialogue acts.

The common ground is represented as a DRS. At any time, the common ground contains a set of **referents**, and a set of propositions about these referents. To support the interpretation of plurals, these referents can be **group entities**, whose semantics are defined as in Kamp and Reyle (1993). For example, the referent associated with the coordinated NP *John and Mary* (in fact presupposed by this NP) would be as shown in Figure 1.

x1 x2 x3
x3 = x1 $\oplus$ x2

named(x1, “John”)
named(x2, “Mary”)

Figure 1: Group entity presupposed by *John and Mary*

2.2 Dialogue model for two-speaker dialogue

For a two-speaker dialogue, our model of dialogue structure includes a fairly standard set of dialogue moves. Questions can be QUERY (a genuine question), CLARIFY (for a clarification question) and CHECK (for a question which explicitly checks the student’s knowledge). Assertions can be ASSERT (for a new fact apropos of nothing) and ANSWER (a question). Yes-no questions and their answers are distinguished from *wh*-questions in each case. We also have various acts for greeting and farewell.

The dialogue model specifies legal sequences of dialogue acts. The simplest structures involve a forward-looking act paired with a backward-looking act; for instance QUERY followed by ANSWER, or ASSERT followed by ACCEPT. More complex examples can be formed by nested subdialogues. Several types of utterance can be followed by a CLARIFY question; as is now common, such a question is pushed onto a dialogue stack, so that a subsequent response is understood as addressing the most recent question. A resolved clarification subdialogue is removed from the stack altogether (and as a side-effect, the semantics of the utterance which has been clarified is updated).

2.3 Mixed-initiative dialogue

Our system supports a variety of mixed-initiative dialogue with the user. When there is nothing on the stack, the user is given the initiative, but she can choose to concede it simply by hitting ‘return’. If the user takes the initiative, the system will interpret the utterance, and then generate a response utterance. If the user concedes the initiative, the system will generate an initiative of its own, and then ask for a response from the user. If the system’s initiative was a question, the user must answer it (or ask a clarification); if the system’s initiative was an assertion, the user can give an explicit acknowledgement, or generate a new initiative (which will be treated as an implicit acknowledgement of the assertion). Thus there are two situations in which the system interprets the user’s utterance—first if it is an initiative, and second if it is a response to a system initiative.

2.4 The utterance interpretation pipeline

Interpretation of a user utterance involves the same sequence of processes, whether the utterance is an initiative or a response. The sentence is first parsed, producing a set of alternative parses. Each parse is associated with a semantic interpretation. Each interpretation is converted into an update, yielding a set of candidate updates for the sentence. To choose the most appropriate update, an attempt is made to attach each candidate to the common ground, by resolving its presuppositions and establishing a relationship with dialogue acts on the stack. The update which is easiest to attach is selected as the preferred interpretation. Any ties are resolved by statistical parsing techniques and various types of clarification question; for details see Lurcock *et al* (2004).

3 Our current multi-speaker system

To support a multi-speaker environment, our system is designed to 'play' several different characters, who have separate knowledge bases, and who can communicate both with the user and with each other. The system-played characters are relatively autonomous, keeping their own private version of the common ground, and updating it independently. The multi-speaker system is described and motivated in detail in Knott and Vlugter (in press); in this section, we summarise the most important aspects of the system.

3.1 The conversation management algorithm

The interaction between the system and the user is governed by a high-level **conversation manager**. Basically, the user and the system take turns to contribute to the conversation; but while the user contributes at most one utterance, the conversation manager implements a number of loops over all system-played characters. There are three basic loops, which operate at different points in the conversation.

3.1.1 The system initiative loop

At any point when the stack becomes empty, initiative is passed to the user, but the user can choose to concede the initiative by hitting 'return'. If the user concedes, the system must generate a new initiative.

Since the system plays several characters, generating an initiative is not simply a matter of choosing what to say: it must also be decided which character will talk next. These two decisions interact, because each character has a separate knowledge base, and therefore the options open to one character may not be open to another.

In our two-speaker system, initiative selection is done by generating a set of candidate utterances and giving each a score, with the highest-scoring utterance being delivered to the user (Slabbers, 2005). In the multi-speaker system, the highest-scoring utterance is computed for each character, and the character with the highest score is selected as the one to talk next. There is a certain amount of private communication between system characters in this process; basically, the best utterance is selected without regard for who will say it.

In the multi-speaker system, the character who is selected to speak next must also decide who to *address*. Again, since there is nothing on the stack when an initiative is taken, the addressee is unconstrained. In our system, we implement a simple rule, which is dictated by the educational domain: system characters always address the user, so that the user is maximally involved in the conversation.

3.1.2 The inter-character update loop

We assume that all system characters are 'active overhearers'—what Goffman (1976) calls **ratified side-participants**—who actively interpret each utterance in the conversation, even if it is not addressed to them. Therefore, when one system character generates an utterance, each other system character must update its common ground accordingly.

Naturally, the other characters do not need to interpret the utterance from scratch. However, since they all maintain their own version of the common ground, each must do a certain amount of work independently. The point in the interpretation pipeline where the private work begins is the point where the MRS of the selected parse (i.e. its semantic representation) is converted to an update. Therefore, after every utterance by a system character, the MRS associated with that utterance is passed to each other system character, and each processes the MRS privately. Effectively, system characters communicate with each other at the level of MRS representations, rather than sentences. This eliminates the possibility of syntactic ambiguity, but note that the potential for pragmatic ambiguity remains, if any discrepancies arise between different characters' versions of the common ground.

3.1.3 The system response loop

If the user chooses to take an initiative, it can be an assertion or a question. In either case, each system character needs to interpret it, and update their common ground. Then those system characters who were addressed need to give responses.

Interpreting the user's utterance There is no need for each character to parse the user's utterance, or disambiguate it. A single character is chosen at random to perform these processes, and to engage in clarification subdialogues with the user if necessary. The result of these processes is a single update; the MRS associated with this update is then communicated to all system characters. Each character then creates their own version of the update for this MRS, and incorporates this update into their private common ground.

Group responses In a multi-speaker conversation, an utterance can be addressed to the full group of participants, or to a specific participant, or to a subset of participants: i.e. for every utterance there is an **addressee group**. A basic principle in our system is that every participant in the addressee group of a forward-looking utterance has the right to respond to it—and that at least one member of this group *must* respond to it. To keep things simple, we specify that responses by different members of the addressee group should not overlap.[1] Thus group responses are implemented as strict sequences of utterances by members of the addressed group. There is some evidence that group responses are generated in this way in spoken conversation, at least for some dialogue acts; see e.g. Ginzburg and Fernández (2005).

The system's response to a user initiative is implemented as a loop over all system characters, chosen in an arbitrary order. Naturally, not all of these characters will be in the addressee group. Each character determines independently whether they are an addressee. (See Section 3.2 for a discussion of how this is done.) If they are not an addressee, they remain silent. If they are an addressee, and a response has not yet been given, they are obliged to give one.

Responses are not delivered directly to the user. Each new utterance is added to a buffer containing the full group response, which is delivered in one piece to the user when complete. Thus the user is not able to interrupt in between respondents. Also, each character who delivers a response broadcasts the MRS for their response to all other system characters, who process it and add the update to a private data structure called the **group response record**, which they consult when generating a subsequent response of their own.

If a character adds a response to a non-empty group response record, they can choose to preface the response with an **inter-respondent dialogue act**, signalling agreement or disagreement with the previous respondent before presenting its own response.[2] A character can also choose to give no response, which is interpreted as signalling agreement with all the responses given so far.

A group response generated by our system is given in Dialogue 1.

(1)
User: John, Bill, where do Mary and Sue live?
Bill: Well Mary lives in Dunedin.
John: That's right. And Sue lives in Auckland.

Here, the user addresses a question to two system characters, John and Bill, who respond in turn. Bill responds first, providing a partial answer (introduced by by *Well*, to signal that it is incomplete). John responds next; he begins by agreeing with Bill's answer, and then fleshes this out with an answer of his own. More details about how group responses are generated can be found in Knott and Vlugter (in press).

3.2 Addressee selection

When interpreting a user utterance, whether it is an initiative or a response, each system character must determine to whom it is addressed, to decide whether or not they should formulate a response. There are various types of information which contribute to this decision in real-life dialogue (see Sacks *et al.* (1974), and much subsequent work). Some of these involve eye gaze or gesture, and are beyond the scope of our current system. But there are nonetheless several different sources of information about the addressee which are available; in this section, they are discussed in turn.

Most obviously, the addressee of an utterance can be specified linguistically, as a modifer of the utterance, as in the first utterance of Dialogue1. Much has been written about the grammar and semantics of addressee terms, which are quite different from ordinary referring expressions (see e.g. Longobardi (1994)). Clearly, an explicit addressee term unambiguously identifies the intended addressee of an utterance.

If there is no explicit addressee, the structure of the preceding conversation provides some indication of the addressee group. Two types of context can be distinguished. If there is a forward-looking dialogue act on the stack, and the next utterance responds to this act, the addressee of the response utterance should obviously be the speaker of the act being responded to. Thus in Dialogue 2, the user's response should be understood as addressing Bill.

(2)
Bill [to User]: What is your name?
User: My name is Ursula.

We refer to the addressee of a backward-looking dialogue act as a **structurally-defined addressee**.

---

[1] In fact, given that our system uses written text as an interface, rather than spoken text, it is not even clear what overlapping responses would look like.

[2] In fact, only agreement is implemented so far.

The structural constraint seems quite strong; if an explicit addressee term is given which conflicts with it, the result is quite jarring. However, if there is a conflict, our system gives precedence to the explicitly specified addressee. This element of our addressee selection algorithm is very similar to that proposed by Traum (2004).

If there is no forward-looking dialogue act on the stack, the next utterance is a new initiative, which can in principle be addressed to any subgroup of participants. However, we suggest that if the user produces a new initiative without an explicit addressee term, (s)he is probably continuing an interaction with participants who have just been involved in the conversation, rather than making a switch to a new group of participants. We therefore define a **default addressee group** for an utterance making a new initiative, which consists of the group of participants involved in the most recent subdialogue (minus the speaker of the utterance). In Dialogue 3, for instance, it seems likely that Sue's second utterance is addressed to both Bill and Bob, rather than to one or the other separately, or (still worse) to someone else entirely.

(3)
Sue: What's the time?
Bill: Five o'clock.
Bob: Yup, five o'clock.
Sue: Do you want to go for tea?

This notion of default addressee is somewhat different from that proposed in Traum (2004); see Knott and Vlugter (in press) for a more detailed comparison. The default addressee is obviously a fairly weak notion, and an explicit addressee overrides it very easily. However, it makes for natural dialogues if the user fails to specify an addressee.

A final interesting issue in our system concerns the identity of the addressees of utterances occurring as part of a group response. Who is each member of a responding group addressing in their utterance? The most recent respondent? The speaker of the forward-looking act being responded to? The full group of respondents? Our solution is to sidestep these questions, as they do not seem to have clearcut answers. Since the notion of the addressee group is primarily used to decide who goes next, our solution is simply to specify a prearranged sequence of speakers during group responses, as described in Section 3.1.3.

3.3 Pronoun interpretation and generation

Recall that the purpose of our multi-speaker dialogue system is to create an environment in which a language learner can exercise various aspects of the language being learned. In our case, to provide a natural environment for exploring the system of personal pronouns, we needed a system which allows a conversation between multiple speakers. In this section, we describe how the multi-speaker system supports the generation and interpretation of a range of personal pronouns, focussing on our target language, M¯aori.

The pronoun system in M¯aori is more complex than that of English. Pronouns are specified for 'number' and 'person', but there are three values for the 'number' feature (singular, dual and plural); in addition, for first-person dual and plural pronouns, there is a distinction between **inclusive** pronouns (which include the addressee) and **exclusive** pronouns (which do not). Thus, for instance, the pronoun *māua* means 'us two (not including you)', while *tātou* means 'us three (or more), including you'. Clearly, these are expressions which a student will need to practice.

In our system, pronouns are presuppositional; different pronouns presuppose groups with different properties. For the sake of generality, all pronouns presuppose a group entity: a singular pronoun requires the group entity to have cardinality 1, a dual pronoun requires it to have cardinality 2, and so on. A first-person pronoun requires a group entity which includes the speaker as one of its members; a second person pronoun requires a group entity which includes the addressee group, and which does not include the speaker. A third-person pronoun requires a group entity which includes neither the speaker nor the addressee group. Finally, all pronouns presuppose groups which are linguistically **salient**. The speaker and addressee group are salient automatically, by virtue of their role in the current interaction. In our implementation, the only other way a group can become salient is through having been referred to in the recent conversation. (Our precise algorithm for determining salience is very simple, but could of course be elaborated in many different directions.)

Having specified the appropriate properties for pronouns, the interpretation of an utterance containing a pronoun is handled fairly automatically by the presupposition resolution system. If the referent of a pronoun is unclear (typically because there are several candidates with roughly equal salience), a referential clarification question is asked (e.g. *WHICH man is your friend?*). However, the generation of utterances containing pronouns is more complex. First and second-person singular pronouns are unambiguous, and thus relatively straightforward, but third-person pronouns and plural pronouns are often ambiguous; for instance, *you(pl)* could refer to several different groups including the addressee. If the alternative groups have similar salience, the system chooses to generate a full referring expression (which will involve coordination for plural referents). But if the intended referent is clearly more salient than its competitors, a pronoun is allowed.

4 Multi-speaker conversation as a CALL environment for learning pronouns

The previous section summarised how the important aspects of our multi-speaker environment are implemented. In this section, we provide an example of the kind of CALL dialogue which our system supports.

The interaction can be divided into two phases. In one phase, the student engages in an open-ended dialogue with the system characters, and information is exchanged between all participants. An example of this phase is given in Figure 2. Note that while the student is relatively passive in this interaction, she is exposed to a range of different personal pronouns. In a second phase, the system generates a sequence of checking or comprehension questions for the student. An example of this phase is given in Figure 3. Note that checking questions are generated by different characters; this is to ensure that a good mixture of referring expressions is generated, to test the student's knowledge of these concepts.

User: Hello.
Mary: Hello.
John: Hi.
Bill: Hello.
User: ⟨return⟩
Mary: Ursula, where are you going today?
User: I am going to the beach.
Mary: Okay.
User: ⟨return⟩
John: I am going to the beach too, Ursula.
User: ⟨return⟩
Mary: Bill and I are going to a restaurant, Ursula.
User: ⟨return⟩
Mary: We(dual,excl) always go to Comme Chez Soi.
User: ⟨return⟩
John: I always go to Comme Chez Soi too, Ursula.
User: ⟨return⟩

Figure 2: First phase of a CALL dialogue (translated from M¯aori)

John: Where are we(dual,incl) going today, Ursula?
User: We(dual,incl) are going to the beach.
John: Well done!
Mary: Where are you and John going today, Ursula?
User: We (dual,excl) are going to the beach.
Mary: Well done!
John: Where do Bill and Mary always go, Ursula?
User: They(dual) always go to Comme Chez Soi.
John: Well done!
Mary: Where do Bill, John and I always go, Ursula?
User: You(plur) always go to Comme Chez Soi.
Fred: Where do John, Bill and Mary always go, Ursula?
User: They(plur) always go to Comme Chez Soi.

Figure 3: Second phase of a CALL dialogue: checking questions (translated from M¯aori)

5 Implementing educational goals and student modelling

The dialogues in Figures 2 and 3 may appear serendipitous; the phase 1 dialogue happens to introduce a number of groups, which allow the phase 2 checking questions to target certain pronouns. In fact, the initiative-selection routine in our system is configured to encourage the right groups to be introduced, and the right checking questions to be asked. The initiatives taken by system characters are evaluated primarily on their educational usefulness, rather than on their content *per se*. In this section, we describe how this evaluation takes place.

5.1 Lesson authoring

To begin with, we need a method for telling the system what the educational agenda is for a given lesson. In our system, we create a sequence of lessons, each of which takes the form of a dialogue, with new topics about the language introduced in each lesson.

To create the content for a given lesson, characters need to be given an appropriate knowledge base. To build these knowledge bases, our system supports a special kind of dialogue called an **authoring dialogue**, in which a human author simply tells each character what they know (Slabbers, 2005). The authoring dialogue for each lesson is preprocessed to build the educational agenda for the lesson. The agenda in our system is simply a list of the lexical types which appear in the current authoring dialogue, but not in previous ones. The student must demonstrate understanding of each of these new types during the dialogue in order to move to the next lesson.

5.2 Authoring the pronoun agenda

To create an agenda for pronouns, it is somewhat inconvenient to work directly from the authoring dialogue, since this requires the author to set up the right environment for using each pronoun to be included in the agenda. In our system, we simply introduce a **command** which the author can type instead of an utterance: the keyword `pronoun:` followed by a specific pronoun. This adds the pronoun in question to the agenda for the current lesson. For instance, the command `pronoun: tāua` would add *tāua* to the pronoun agenda.

5.3 The student model

The student model is an overlay on the educational agenda, indicating the degree to which each construction in the agenda has been assimilated by the student. At a given point in a lesson, for instance, the agenda might contain the pronouns *tāua* (we/dual-inclusive) and *māua* (we-dual-exclusive), and the student might have shown some assimilation of the former, but none of the latter.

There are two ways in which the student model can be updated. Firstly, if the student uses a sentence which can be successfully attached to the common ground, the assimilation score for each construction in the sentence is incremented. Secondly, if the student correctly answers a question, the assimilation score for each construction in the question is incremented. These methods are quite simplistic—and at present we have no method of *decrementing* assimilation scores in response to student errors. These are topics for future research. Basically, the lesson continues until the student has shown sufficient assimilation of each construction on the agenda.

6 Educational content selection strategies

The goal of our dialogue system is that the student assimilates all the constructions on the agenda for the current lesson. This is very different from the goal of most dialogue systems. Typically the goal is content-based—for instance, communicating a particular set of facts to the user, or obtaining a particular set of facts from the user. In our system, the goal is *form-based*—we require that the user correctly employs, or shows understanding of, a certain range of grammatical constructions.

One way the system can facilitate achievement of the goal is through a targeted content selection strategy. There are two ways content selection can help achieve the goal. The most direct way is to give the student the opportunity to display knowledge about a particular construction. For instance, by asking a checking question using a particular pronoun, or one whose answer requires the use of a particular pronoun, the student is given an opportunity to show they can use it by giving the right answer.

However, this direct approach is not always possible for pronouns, because it is not always possible to use a given pronoun felicitously. A more indirect content selection strategy should therefore encourage the creation of an *environment* in which the targeted pronouns can be felicitously used. For instance, to allow the student to use *tāua* (we-dual-inclusive), there needs to be a salient group containing exactly two individuals, which includes the student and the addressee. So a useful character initiative would be one which creates this group, if it does not already exist. What is more, there has to be something predicated of the group, to serve as the content of an utterance about the group.

In our system, characters create appropriate groups dynamically, simply by making facts up. For instance, if the user has asserted a predicate about themselves, a system character might take an initiative asserting that this predicate is also true of himself, or of a group of participants including himself. This has the additional benefit of adding **cohesion** to the dialogue. For instance, in Dialogue 2, John's utterance *I am going to the beach today too, Ursula* is a very appropriate initiative, since it maintains the topic of the user's previous utterance. However, the main reason for John to generate this utterance is to allow him later to ask the user the checking question *Where are we (dual,inclusive) going today, Ursula?* during phase two of the lesson. Note that the utterance also allows other checking questions; for instance, Mary can ask a question (*Where are you and John going today, Ursula?* which sets up the user to respond with a suitable dual exclusive pronoun.

7 Summary and further work

In this paper, we have presented a multi-speaker human-machine dialogue system which is configured to function as a CALL tool. The multi-speaker system is of interest in its own right, as it uses a novel conversation management architecture, and some novel methods for addressee selection. In particular, its approach to the handling of group responses has some novel features. However, it is also of interest in that it provides a novel kind of CALL environment, where the student can exercise her knowlege of elements of the language being learned which depend on multiple speakers. In our case, the focus is on a system of personal pronouns and addressee terms, but in other languages, the same multi-speaker environment would be useful in giving a student practice in other topics, for instance in verb conjugation.

It is interesting to compare the current approach to CALL dialogues to the type of exercises which are supported by conventional CALL software. Typical CALL exercises involve canned materials: a hand-built text is provided, followed by hand-built comprehension questions tailored to test the targeted knowledge. In our system, the user's interaction with the system is a lot more flexible; the user simply engages in a wide-ranging mixed-initiative dialogue. However the system-played characters are acting behind the scenes in the same kind of way as the human author of a canned CALL exercise, to create opportunities to test the student's knowledge.

It is not clear whether the extra flexibility in our CALL interactions is of any value to the student. Obviously, when learning a language it is very beneficial to interact with native speakers in as natural an environment as possible. Our aim is basically to recreate this kind of interaction (or at least to approximate it), while engineering the flow of conversation so as to create useful educational opportunities. In future work, we will evaluate the system to see if this aim is met.

References

I Bayard, A Knott, and S de Jager. 2002. A unification-based grammar for a fragment of M¯aori and English. In *Proceedings of the 2nd Australasian Natural Language Processing Workshop (ANLP 2002)*.

A Copestake, D Flickinger, I Sag, and C Pollard. 1999. Minimal Recursion Semantics: An introduction. Manuscript, CSLI, Stanford University.

J Ginzburg and R Fernández. 2005. Action at a distance: the difference between dialogue and multilogue. In C Gardent and B Gaiffe, editors, *Proceedings of DIALOR'05: The ninth workshop on the semantics and pragmatics of dialogue*, pages 85–92, Nancy, France.

E Goffman. 1976. Replies and responses. *Language in Society*, 5:257–313.

H Kamp and U Reyle. 1993. *From discourse to logic*. Kluwer Academic Publishers, Dordrecht.

A Knott and P Vlugter. in press. Multi-agent human-machine dialogue: issues in dialogue management and referring expression semantics. *AI journal*.

G Longobardi. 1994. Reference and proper names: a theory of N-movement in syntax and logical form. *Linguistic Inquiry*, 25(4):609–665.

P Lurcock, P Vlugter, and A Knott. 2004. A framework for utterance disambiguation in dialogue. In *Proceedings of the 2004 Australasian Language Technology Workshop (ALTW)*, pages 101–108, Macquarie University.

H Sacks, E Schegloff, and G Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

N Slabbers. 2005. A system for generating teaching initiatives in a computer-aided language learning dialogue. Technical Report OUCS-2005-02, Department of Computer Science, University of Otago, Dunedin, New Zealand.

D Traum. 2004. Issues in multi-party dialogues. In F Dignum, editor, *Advances in agent communication*, pages 201–211. Springer Verlag Lecture Notes in AI 2922.

P Vlugter, A Knott, and V Weatherall. 2004. A human-machine dialogue system for CALL. In *Proceedings of InSTIL/ICALL 2004: NLP and speech technologies in Advanced Language Learning Systems*, pages 215–218, Venice.

Correction and acceptance by contrastive focus

Elena Karagjosova
University of Oslo
elena.karagjosova@ilos.uio.no

Abstract

An account is presented of the focus properties, common ground effect and dialogue behaviour of the accented German discourse marker *doch* and the accented sentence negation *nicht*. It is argued that *doch* and *nicht* evoke as a focus alternative the logical complement of the proposition expressed by the sentence in which they occur, and that an analysis in terms of contrastive focus accounts for their effect on the common ground and their function in dialogue.

1 Introduction

Recently, there has been an increased interest in the relation between information structure and discourse structure (cf. e.g. (Kruijff-Korbayová and Steedman, 2003; Umbach, 2004; Jasinskaja et al., 2004)). This paper aims at contributing to the ongoing discussion by extending its scope further to dialogue structure. I present an account of the dialogue function of utterances containing focussed constituents of a type that has been scarcely studied with respect to their focus properties, namely the accented German discourse marker *doch* and the accented sentence negation *nicht*. I suggest that the focus properties of these items in the type of utterances I discuss are best captured in terms of contrastive focus, which in a focus-semantic framework such as Rooth (1992) allows viewing them as having anaphoric properties. These anaphoric properties are furthermore seen as responsible for a pattern I observe with respect to the behaviour of focussed *doch* and *nicht* in dialogue: typically, utterances containing accented *doch* and *nicht* serve as corrections in dialogue. However, when the context does not license an utterance to be corrected, the use of *doch* and *nicht* is not infelicitous but the utterance is interpreted as acceptance. On my account, accented *doch* and *nicht* evoke as a focus alternative the logical complement of the proposition expressed by the sentence in which they occur. When the context contains a suitable antecedent to which the focus alternative can be linked, then the utterance is interpreted as a correction, when not, it is interpreted as acceptance.

I argue furthermore that an analysis of *doch* and *nicht* in terms of contrastive focus accounts for their invariant effect on the common ground irrespective of their use in corrections or acceptances. I suggest that a taxonomy of dialogue moves that takes into consideration how the common ground is established and changed, such as the one proposed in Traum (1994), is needed to capture the dialogue behaviour of the utterances containing these focussed expressions.

The paper is organised as follows. Section 2 introduces the notion of correction and discusses examples of corrections involving focussed *doch* and *nicht*. Section 3 deals with the focus properties of *doch* and *nicht* and Section 4 explores the notion of contrastive focus with respect to correction utterances containing these expressions. Section 5 discusses examples of acceptance with focussed *doch* and *nicht* and the question of the suitable dialogue move taxonomy. The results are summarised in Section 6.

2 Correction

One of the most comprehensive treatments of correction I am aware of is Steube (2001). One condition for interpreting an utterance as correction is according to her that the sentence that is corrected,

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 26–33.

called *corrigendum*, is explicitly given immediately before the correcting sentence, called *corrigens*. I will adopt this terminology in what follows. According to Steube, the semantics of corrections boils down to indicating that the corrigendum is not true or correct. In terms of pragmatics, corrections suggest that the corrigendum should be removed from the common ground (CG). The corrigens proposes furthermore a replacement to be added to the CG instead of the corrigendum.

Steube (2001) distinguishes two types of correction constructions. The first type involves a complex sentence as corrigens, where the first part contains the explicitly negated corrigendum and the second part, introduced by *but* in English and *sondern* in German, provides the replacement, cf. (1a).[1] The second type is called "backward-related corrections with contrastive focus" and is illustrated by (1b). It involves a corrigens which contains a contrastively focussed replacement of the constituent to be corrected.

(1) A: [Paul]$_F$ kommt.
'Paul is coming'
  a. B: Nicht [Paul]$_F$ kommt, sondern [Peter]$_{CF}$.
  'Not Paul is coming but Peter.'
  b. B: [Peter]$_{CF}$ kommt.
  'Peter is coming.'

The cases of correction I will discuss here involve the focussed discourse marker *doch* and the focussed sentence negation *nicht*. They are not considered by Steube but can nevertheless be assigned to her second type of correction contructions, namely the backward-related corrections with contrastive focus. What is particular about corrections with focussed *doch* and *nicht* is that they involve two sentences which differ only in their polarity, which suggests that the polarity is the only part of the sentence that is corrected.[2]

Let's look at some examples. In corrections, accented *doch* has two possible realisations: (i) as a sentence equivalent, where it is categorised as a response particle, and (ii) in the middle field of the German sentence, where it is categorised as an adverb.[3] The response particle (henceforth, RP) *doch* has the function of refuting an immediately preceding negated sentence, thus asserting that the positive counterpart of the sentence is true. Intuitively, (2B) rejects the preceding statement that Karl was not at the party and asserts that, on the contrary, he was at the party:

(2) A: Karl war nicht auf meiner Party.
'Karl was not at my party.'
B: DOCH.[4] (= Karl war auf deiner Party.)
'He was indeed.'

From a dialogue structural point of view, (2B) serves as a correction. As a sentence equivalent, RP asserts the positive counterpart p of the proposition $\neg p$ expressed by the preceding sentence, as indicated in the example.[5] (2B) suggests that the proposition $\neg p$ expressed by the preceding sentence is not true and that it is not accepted as an update of the CG. The proposition p asserted by (2B) is suggested as a replacement to be added to the CG (cf. also Zeevat (2005) who points out that the intended change of stressed *doch* to the CG is a combination of retraction of not-ϕ and the addition of ϕ as a replacement, where ϕ is the content of the utterance).

A similar effect can be observed in the case of the other accented use of *doch* we will consider, namely adverbial *doch* in the middle field. In (3),[6] the *doch* utterance serves as a (self)correction.[7] A$_{2}$ offers a replacement p for the proposition $\neg p$ expressed by A$_{1}$ to be added to the CG:[8]

(3) A$_{1}$: **es geht nicht**.
'it does not work'
B$_{1}$: du musst die Schraube drehen, [...]
'you must turn the screw'

---

[1] F denotes focus and CF contrastive focus.

[2] This is reminiscent of what is known as "verum focus" ("polarity focus"). I'll point at some differences in Sec. 3.

[3] Adverbial *doch* can also occupy the initial field of the German sentence, where it is also accented, but without serving correction purposes.

[4] Small capitals denote accent.

[5] The preceding utterance may be uttered by the same speaker (self-correction) or a different speaker (correction).

[6] This example is taken from the Baufix corpus, http://www.sfb360.uni-bielefeld.de/transkript/

[7] The same example could be modified into a proper correction:

(1) A$_{1}$: **es geht nicht**.
'it does not work'
B$_{1}$: du musst die Schraube drehen, [...]
'you must turn the screw'
A$_{2}$: [...] hast recht
'you are right'
B$_{2}$: Na siehst du? **es geht** DOCH
'What did I tell you? It works.'

[8] Self-corrections like (3) are at the same time cases of belief revision, which can be seen as a special case of revision of the CG, cf. Section 5.

A$_2$: [...] hast recht, **es geht** DOCH
'you are right, it works'

One difference here compared to the RP use of *doch* is that the corrigendum lies further back in the dialogue context, which however does not affect the correction interpretation of (3A$_2$).[9]

Finally, accented sentence negation *nicht* can also serve as a correction (cf. also Zeevat (2004)). Consider (4), where we have a similar situation as in (2), except for the fact that the polarity of the sentences is switched:

(4) A: Karl war auf meiner Party.
'Karl was at my party.'
B: Er war NICHT da.
'He wasn't there.'

(4B) negates (4A), asserting thus the logical complement $\neg p$ of the proposition p expressed by (4A). Just like (2B), (4B) suggests that the proposition expressed by the preceding utterance, here p, is not true and therefore not accepted as an update of the CG. (4B) expressing $\neg p$ is offered as an update instead.

According to Steube, the entities in the focus domain of the contrastive focus replace type-identical entities of the corresponding corrigenda. In the examples of *doch* and *nicht* above, however, the entities that are proposed as replacements are the entire negative or positive propositions expressed by the respective corrigens (2B), (3B) and (4B). Nevertheless, the domain of the contrastive focus of *doch* and *nicht* cannot be seen as stretching over the entire sentence, since only the polarity is corrected, the rest being known from the context (given). The entities that are replaced are the reversed polarity propositions expressed by the respective corrigenda (2A), (3A) and (4A), i.e., the entire corrigens replaces the entire corrigendum.

Steube points out further that the correcting speaker regards the entity that is replaced by the corresponding entity in the focus domain of the contrastive focus as an untrue alternative. In alternative semantics, however, focussed expressions give rise to alternatives of the entire underlying proposition rather than of just the focussed element (Rooth, 1992). Moreover, contrastive focus evokes an alternative proposition that should be anaphorically recoverable from the context.[10] Intuitively, in our examples this alternative coincides with the corrigendum.

In order to spell out these intuitions in a more precise way, we next turn to the question of the focus properties of *doch* and *nicht*.

3 The focus properties of *doch* and *nicht*

In alternative semantics (Rooth, 1992), a focussed expression is accounted for by assuming that it adds a focus semantic value $[\![.]\!]^f$ to the semantic interpretation of the sentence. The focus semantic value represents a set of alternatives - a set of propositions which contrast with the ordinary semantic value $[\![.]\!]^o$ and which are "obtainable from the ordinary semantic value by making a substitution in the position corresponding to the focussed phrase" (ibid., p. 76). The ordinary semantic value is always an element of the focus semantic value. The set of alternatives is salient but not necessarily explicitly mentioned and contains only alternatives which are type-identical with the focussed expression.

In what follows, I examine what the focus sets of alternatives evoked by *doch* and *nicht* look like. I start with the focussed sentence negation *nicht* since it, in contrast to *doch*, has been studied before from this angle, albeit only rudimentarily, to my knowledge.

3.1 Focussed *nicht*

Höhle (1992) suggests that the focussed sentence negation constitutes a set of alternatives together with expressions like *vielleicht* ('maybe'), *bestimmt* ('surely') and *wahrscheinlich* ('probably') by means of which the speaker "expresses his opinion with respect to the truth content of the contextually given thought".[11] Höhle's considerations regarding this issue are very brief and remain at the intuitive level. On closer examination, they do not prove right. Consider the examples on which Höhle bases his assumption:

(5) A: VIELLEICHT hört er ihr zu.
'Maybe he listens to her.'
A': er hört ihr BESTIMMT zu
'Sure he listens to her.'

---

[9] Exchanges like *A: Es geht nicht. B: Es geht* DOCH. are considered marginal by native speakers. The shorter follow-up utterance with the RP, *B:* DOCH, is preferred.

[10] Steube's treatment of contrastive focus seems to be consistent with this view but is set in the more cognitively oriented "two-level semantics" framework.

[11] Höhle suggests furthermore that focussed sentence negation and verum focus give rise to one and the same set of alternatives.

B: (nein) er hört ihr NICHT zu
'(no) he does not listen to her'

The focussed negation particle *nicht* in (5B) seems to be contrasted with neither *vielleicht* nor *bestimmt* but just with the positive counterpart p of the proposition $\neg p$ expressed by (5B).[12] This is suggested by the fact that (5B) can only be interpreted as 'it is not the case that he listens to her' and not as 'it is not the case that it is possible/certain that he listens to her'. The reason for that is the fact that epistemic modals such as *vielleicht* and *bestimmt* do not contribute to the propositional content of the sentence and thus cannot be challenged by another speaker (cf. also Romero (2005)).

This suggests that the set of alternatives evoked by focussed *nicht* contains only two elements: the negated proposition that is asserted and represents the ordinary semantic value of the sentence, $\llbracket [_S\text{er hört ihr nicht zu}] \rrbracket^o$, and its positive counterpart: $\llbracket [_S \text{ er hört ihr } [\text{NICHT}]_F \text{ zu}] \rrbracket^f = \{ \llbracket [_S\text{er hört ihr nicht zu}] \rrbracket^o, \llbracket [_S\text{er hört ihr zu}] \rrbracket^o \}$. This can be generalised as $\llbracket [_S[\text{nicht}]_F S \rrbracket^f = \{\neg p, p\}$, where p is the proposition expressed by the sentence S.

3.2 Focussed *doch*

As in the case of focussed *nicht*, *doch* cannot be understood as being contrasted with modal expressions: (6B) asserts that Karl has lied, rather than that he has *actually* not lied (as opposed to him *possibly* not lying):

(6) A: Karl hat vielleicht nicht gelogen.
'Maybe Karl did not lie.'
B: DOCH. (=Karl hat gelogen.)
'He has indeed.'

This suggests that just like focussed sentence negation, RP *doch* contrasts the asserted proposition with its negative counterpart. In other words, the focus semantic value of RP *doch* is the set containing its ordinary semantic value $\llbracket [_S\text{doch}] \rrbracket^o = p$ and the alternative that contrasts with it, namely $\neg p$. I.e., $\llbracket [_S[\text{doch}]_F] \rrbracket^f = \{p, \neg p\}$, where p is a proposition negated (or asserted, cf. (14)) by the immediately preceding sentence. The same applies also for adverbial *doch*: in (7) (a modified version of (3)), *doch* is contrasted with the negation in (7A$_1$) rather than with the modal expression *bestimmt*:

(7) A$_1$: es geht BESTIMMT nicht.
'it certainly does not work'
B$_1$: du musst die Schraube drehen, [...]
A$_2$: [...] hast recht, es geht DOCH

In other words, $\llbracket [_S\text{doch}]_F S \rrbracket^f = \{p, \neg p\}$. Note however that, in contrast to the RP *doch*, the proposition asserted by adverbial *doch* may also be negative (cf. *Es geht* DOCH *nicht*).

This means that the focus sets of alternatives evoked by focussed *doch* and *nicht* are identical ($\llbracket [doch]_F \rrbracket^f = \llbracket [nicht]_F \rrbracket^f$), the difference being the distinct ordinary semantic values that are elements of the focus set of alternatives: p for the RP *doch*, $\neg p$ for *nicht* and either of the two in the case of adverbial *doch*.

4 Contrastive focus

Following Rooth (1992), focus may have two main functions, depending on how the uttered sentence is understood against the salient set of alternatives: exhaustive focus and contrastive focus. In the case of exhaustive focus, the function of the accent is to signal that the focussed expression is the only one that is true out of the set of alternatives, e.g., in question-answer pairs. In the case of contrastive focus, accent signals that the focussed expression contrasts with a previously uttered member of the focus set of alternatives. Following Rooth (1992), a phrase α is contrasting with a phrase β, if $\llbracket \beta \rrbracket^o \in \llbracket \alpha \rrbracket^f$ and $\llbracket \beta \rrbracket^o \neq \llbracket \alpha \rrbracket^o$.

Focus on *nicht* may be exhaustive, such as in question-answer contexts like (8):

(8) A: War Karl auf deiner Party? $p \vee \neg p$
'Was Karl at your party?'
B: Nein, er war NICHT da. $\neg p$
'No, he wasn't there.'

Here, the answer given, $\neg p$, is suggested as the only one that is true out of the set of alternatives $\{p, \neg p\}$ determined by the question.

On the other hand, in a context like (9A), the focus on *nicht* is contrastive. Here, the ordinary semantic value p of the contrasting phrase S is a previously uttered member of the focus set of alternatives evoked by *nicht*:

---

[12] This is also one of the differences to verum focus, which serves not only to establish a contrast between true and false, but also between possibly true/false and actually true/false, cf.:

(1) A: VIELLEICHT hört er ihr zu.
'Maybe he listens to her.'
B: er HÖRT ihr zu
'He DOES listen to her.'

(9) A: Karl war auf deiner Party. $[\![S]\!]^o = p$
'Karl was at your party.'
B: Er war NICHT da. $[\![[_S[\text{nicht}]_F S]\!]^f = \{\neg p, p\}$
'He wasn't there.'

Focus on *doch* seems to be always contrastive: in (10), what answers the question is $\neg p$ and not $doch\neg p$. In fact, the latter provides additional information about the expectations of the speaker, presenting thus an overinformative answer: Focus on *doch* signals that the opposite, p, was expected to hold:

(10) A: War Karl auf deiner Party? $p \vee \neg p$
'Was Karl at your party?'
B: Nein, er war DOCH nicht da. $\neg p$
'No, he wasn't there after all.'

In corrections like (11), the expectation (here $\neg p$) that is a member of the focus set of alternatives evoked by *doch*, is previously mentioned:

(11) A_1: es geht nicht. $[\![S]\!]^o = \neg p$
B_1: [...]
A_2: [...] es geht DOCH
$[\![[_S[\text{doch}]_F S]\!]^f = \{p, \neg p\}$

Similarly, focus on RP *doch* seems to be always contrastive: although *doch* answers the question $\neg p$? in (12), the context does not license a set of alternatives from which one could be chosen and presented as the only true one.[13]

(12) A: War Karl nicht auf deiner Party? $\neg p$?
'Wasn't Karl at your party?'
B: DOCH. p
'He was indeed.'

An analysis of the RP *doch* in terms of contrastive focus correctly predicts that the ordinary semantic value $\neg p$ of the contrasting phrase S is a previously uttered member of the focus set of alternatives evoked by *doch*:

(13) A: K war nicht auf der Party. $[\![S]\!]^o = \neg p$
B: DOCH. $[\![[_S[doch]_F]]\!]^f = \{p, \neg p\}$

5 Contrastive focus, discourse relations and dialogue moves

The focus properties of *doch* and *nicht* and their function as contrastive focus presented in the previous two sections account for their use in corrections: the preceding context contains a corrigendum which is an element of the focus set of alternatives of *doch* and *nicht* respectively.

Correction is usually viewed as a discourse relation which also can manifest itself in dialogue (cf. Asher (1998)). Umbach (2004) relates the discourse relation of correction to the information structural notion of contrast between alternatives. She views correction as a special case of contrast where one element of the set of alternatives evoked by an accented expression is excluded by substitution: the asserted element is presented as a replacement for the alternative, suggesting that the former should be added to the CG and the latter removed from it. Umbach assigns the same interpretation to contrastive focus, i.e. a common property of correction and contrastive focus is the exclusion of an alternative by means of substitution. This view is consistent with Steube (2001) and the data I presented.[14]

It can be argued, however, that the effect of contrastive focus on the CG we witnessed in the case of corrections with focussed *doch* and *nicht*, is preserved also in contexts in which the same utterances indicate agreement. The following examples illustrate this point.

In some cases, the RP *doch* can indicate acceptance. There it is used as a response to a positively formulated statement, cf. (14). The use of the RP is however understood as the result of reinterpreting the preceding utterance as expressing a negative bias towards the truth of the proposition, as the reconstructed negated question suggests (cf. Helbig (1988)):

---

[13] That focus on the RP *doch* cannot be interpreted as exhaustive is also suggested by the fact that even though *doch* can be used in the context of (8A), it is not understood as an answer to a polar question but to the biased question $\neg p$?, cf. also (14) in Section 5.

[14] Note, however, that Rooth's notion of contrastive focus (CF) cannot be equated with correction: no substitution is involved in his example *An American$_F$ farmer was talking to a Canadian$_F$ farmer.* This sentence can be interpreted as correction only when produced with *contrastive accents* on the focussed expressions. Thus, we need to distinguish two different concepts of CF: a broader, Roothian one in terms of contextual boundedness (corresponding to Umbach's contrast) involving normal accent, and a narrower one involving contrastive accent and signalling additionally context revision (corresponding to Steube's and Umbach's CF). In the case of *doch* and *nicht* we are dealing with the narrower notion, although the nature of their accent type has not been extensively studied yet.

(14) A: Das war sehr freundlich von ihm. p ($\rightarrow$ War das nicht sehr freundlich von ihm? $\neg p$?)
'This was very nice of him. (Wasn't it?)'
B: DOCH, das muss man sagen.
'It certainly was.'

Although (14B) does not serve as a correction since the context does not provide an explicit corrigendum, accent on *doch* evokes the alternative proposition $\neg p$ and indicates that this proposition should be retracted from the CG and replaced by its positive counterpart p.

The same effect can be observed with respect to the adverbial *doch*. Consider (15B), which does not serve the purpose of correcting, but is rather interpreted as the speaker accepting the information p just received and expressing an earlier opposite expectation $\neg p$. The latter is a member of the focus set of alternatives that *doch* gives rise to. This alternative is moreover discarded, i.e. (15B) indicates that $\neg p$ should be retracted from the CG and replaced by p.[15]

(15) A: Karl hat gelogen. p
'Karl lied.'
B: Er hat (also) DOCH gelogen. p
'He lied after all.'

Depending on whether the opposite expectation $\neg p$ was verbalised or not, (15B) may be seen as either a correction (with the respective verbalised expectation as corrigendum) or belief revision. Belief revision can be seen as a special case of CG-revision, since the CG (the things on which A and B agree) does not change with respect to a proposition p before p is added to or deleted from the set of private beliefs of the interlocutors.

Finally, utterances with the accented sentence negation *nicht* can also indicate acceptance. Consider (16) where the negated statement (16A) is accepted rather than denied by the negated statement in (16B). (16B) suggests furthermore that the opposite p was expected and that this expectation was not met:[16]

(16) A: Karl hat nicht gelogen. $\neg p$
'Karl did not lie.'
B: Er hat (also) NICHT gelogen. $\neg p$
'(So) He did not lie after all.'

Like in the *doch* examples above, *nicht* in (16B) evokes an alternative proposition, here p, and indicates that it should be replaced in the CG by what is asserted, namely $\neg p$. The use of the modal particle *also* in (16B) is a further indication of this interpretation. The particle *also* refers to a consequence (explanation, constatation, confirmation, summary or result) from a preceding utterance or a deliberation of the speaker or the hearer (König et al., 1990). Without *also*, and with the proper intonation, the utterance may also be understood as a clarification question motivated by some conflicting expectation. That there is a conflicting expectation is indicated by the accent on *nicht*:

(17) A: Karl hat nicht gelogen. $\neg p$
'Karl did not lie.'
B: Er hat NICHT gelogen? $\neg p$
'He did not lie?'

All the same, questions like (17B) induce changes of the CG, as they, like corrections, occur in dialogue phases where the content of the CG is negotiated.[17]

I suggest to account for the acceptance cases by assuming that the focus sets of alternatives evoked by *doch* and *nicht* are accommodated. According to Rooth, the set of alternatives is salient but need not be explicitly mentioned. It could be argued that in the case of acceptance, accent on *doch* and *nicht* makes the set of alternatives salient.[18]

Accommodation seems to be an especially suitable way of accounting for cases like (14). Here, accommodation makes sure that the evoked alternative $\neg p$ is added to the context of interpreting the *doch*-utterance, which is equivalent to reinterpreting the preceding sentence (14A) as being negated.

In the case of adverbial *doch*, the evoked alternative may have been mentioned earlier in the dialogue, by either of the interlocutors, in which case it can be bound. The resulting utterance is a

---

[15] The polarity of the sentences does not matter, as soon as it is the same in both utterances, i.e. we get the same interpretation when both sentences are negated, i.e. *A: Karl hat nicht gelogen. B: Er hat (also)* DOCH *nicht gelogen.*

[16] With a rising intonation, (16B) can be interpreted as a confirmation question motivated by a conflicting expectation, cf. Zeevat (2004).

[17] Another possible interpretation of (17B) is as a clarification request motivated by uncertainty on the part of B of whether he heard right. Intuitively, such a case fits into our analysis, since the part of the preceding utterance that requires clarification is its polarity.

[18] According to Zeevat (2004), a reason for prosodic prominence is that an alternative is activated. Here, however, the alternative becomes activated as a consequence of focussing, i.e. another reason for the prominence may be activation of nonsalient alternatives. In either case, contrastive focus requires that the alternative is removed from the CG and replaced by the prominent item.

(self)correction with respect to some earlier commitment and at the same time acceptance with respect to the immediately preceding contribution of the other interlocutor. The evoked alternative may however also represent a non-verbalised assumption of the speaker, in which case the information that the speaker used to believe the opposite is added to the CG, i.e. is accommodated.

The case of *nicht* seems to be analogous to adverbial *doch*, i.e. it is possible to imagine a situation where A or B has commited himself to the opposite belief at an earlier stage of the dialogue. Another possibility is that the belief was not manifested in the dialogue, in which case it must be accommodated.

Thus, although context is what determines whether the utterances we discussed serve as corrections or acceptances in dialogue, contrastive focus has in either case the effect of inducing CG-revision. How can this be accounted for in terms of the dialogue move(s) that the respective utterance performs? A look at existing dialogue move classifications and coding schemes reveals that the majority of them do not provide for a level at which the CG-revision takes place (cf. e.g. (Alexandersson et al., 1998), (Allen and Core, 1996)). Corrections are missing from these schemes, the only possibility being to tag them as rejections at the level of their content. The only model that seems suitable to account for the data I presented is Traum's (1994) classification of conversation acts into four different types anchored at different levels of action "necessary to express the content and maintaining the coherence of conversation": turn-taking acts, grounding acts, core speech acts and argumentation acts. In this scheme, the CG-revision aspect of contrastive focus can be accounted for at the level of grounding. This is also consistent with Steube's view that a corrigens blocks the continuous development of a text or dialogue and complies with our characterisation of corrections as representing a process of negotiating the CG. More closely, the scheme provides a grounding act Repair defined as an utterance that changes the content of the discourse unit and that may be a correction of previously uttered material or addition of omitted material. Repairs are characterised furthermore as concerning merely the grounding of content. Thus, the cases of what I called "corrections" with focussed *doch* and *nicht* would be labelled Repair at the level of grounding acts and Reject at the level of core speech acts. Analogously, the acceptance cases would represent the same grounding act of Repair, but a core speech act Accept.

6 Summary and conclusions

I argued that utterances containing focussed *doch* and *nicht* may function as either corrections or acceptances, depending on whether the preceding context contains an element of the set of alternatives that *doch* and *nicht* evoke, or not. In both the case of correction and acceptance, the focus on *doch* and *nicht* is contrastive and the utterance has the effect of revising the common ground. I suggest that a suitable dialogue move taxonomy has to provide for a level that captures the process of grounding.

It is a subject of further research to work out the broader implications that phenomena like the ones described here have for the relation between information structure and dialogue structure, as well as to work out the details of the analysis in a dynamic semantics framework.

Acknowledgements This work is supported by the SPRIK project (*Språk i Kontrast* [Languages in Contrast]), NFR 158447/530. I would like to thank Torgrim Solstad and two anonymous reviewers for their comments.

References

Jan Alexandersson, Bianka Buschbeck-Wolf, and al. 1998. Dialogue acts in Verbmobil-2. Verbmobil-report nr.226, DFKI GmbH, Saarbrücken.

James Allen and Mark Core. 1996. Draft of DAMSL. http://www.cs.rochester.edu/research/speech/damsl.

Nicholas Asher. 1998. Varieties of discourse structure in dialogue. In *Proceedings of the 2nd International Workshop on Dialogue*, Twente.

Gerhard Helbig. 1988. *Lexikon deutscher Partikeln*. Verlag Enzyklopädie, Leipzig.

Tilman N. Höhle. 1992. Über Verum-Focus im Deutschen. In Joachim Jacobs, editor, *Informationsstruktur und Grammatik*. Westdeutscher Verlag, Opladen.

Ekaterina Jasinskaja, Jörg Mayer, and David Schlangen. 2004. Discourse structure and information structure. In S. Ishihara, M. Schmitz, and A. Schwarz, editors, *Interdisciplinary Studies on Information Structure 1*. Universitätsverlag, Potsdam.

Ekkehard König, Detlef Stark, and Susanne Requardt. 1990. *Adverbien und Partikeln. Ein deutsch-englisches Wörterbuch*. Julius Groos, Heidelberg.

Ivana Kruijff-Korbayová and Marc Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information*, (12):249–259.

Maribel Romero. 2005. Biased yes/no questions, really, and answers. Presented at the Workshop on Formal and Computational Approaches to Discourse and Other Particles, Barcelona.

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, (1):75–116.

Anita Steube. 2001. Correction by Contrastive Focus. ZAS Papers in Linguistics 23.

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

Carla Umbach. 2004. On the notion of contrast in information structure and discourse structure. *Journal of Semantics*, 21(2):155–175.

Henk Zeevat. 2004. Contrastors. *Journal of Semantics*, 21(2):95–112.

Henk Zeevat. 2005. A dynamic approach to discourse particles. In Kerstin Fischer, editor, *Approaches to Discourse Particles*. Elsevier, Amsterdam.

Modelling Correction Signalled by “But” in Dialogue

Kavita E. Thomas
School of Informatics
University of Edinburgh*
kavita.e.thomas@gmail.com

Abstract

Claiming that cross-speaker “but” can signal correction in dialogue, we start by describing the types of corrections “but” can communicate by focusing on the Speech Act (SA) communicated in the previous turn and address the ways in which “but” can correct what is communicated. We address whether “but” corrects the proposition, the direct SA or the discourse relation communicated in the previous turn. We will also briefly address other relations signalled by cross-turn “but”. After presenting a typology of the situations “but” can correct, we will address how these corrections can be modelled in the Information State model of dialogue, motivating this work by showing how it can be used to potentially avoid misunderstandings. We wrap up by showing how the model presented here updates beliefs in the Information State representation of the dialogue and can be used to facilitate response deliberation.

1 Introduction

This paper addresses how cross-turn “but”[1] can signal correction in dialogue. We define correction as involving disagreement, denial or rejection of something in the previous turn and either an explanation for why this is disagreed with or the presentation of a replacement.

Although “but” shares the logical meaning of conjunction with “and”, it carries a *conventional implicature* of contrast, as was argued by (Grice, 1975). (Carlson, 1985) argues that only in extreme circumstances does “but” introduce a flat denial, e.g., A says “he’s dead,” and B cries “but he’s not dead!”. He adds that it is relatively more common for “but” to contradict the preceding premise without restating the premise itself, as in Ex. 1 below.

(1) A: Nobody can do that.
B: But she did it.

He also claims that elliptic dialogue (see Ex. 2 below, B2) conveys an alternative rather than the direct contradiction which can be conveyed in B, and that B2 lacks emphatic force.

(2) A: He is extremely good.
B1: But he is slow.
B2: But slow.

(Kreutel and Matheson, 2001) show that “but” corrections can involve direct opposition by indicating contradictory evidence rather than directly negating a prior claim, and may be used to indicate disagreement:

(3) A: Helen didn’t come to the party.
B: But I’m sure I saw her there.

We will start by describing the types of corrections “but” can communicate by focusing on the Speech Act (SA) communicated in the previous turn and address the ways in which “but” can correct what is communicated. We adhere to the idea that SAs act as wrappers around propositions (1962) and aim for a central definition of correction to arise from the analysis. To this end we will focus on previous turns communicating assertions, questions, commands and answers to questions and we will address whether “but” corrects the proposition, the direct SA or the discourse relation communicated in the previous turn. We will briefly discuss related relations signalled by cross-turn “but”. After presenting a typology of the situations “but” can correct, we will address how these

---

*This research was conducted at the University of Edinburgh.

[1] We consider that “but” relates across turns if the constituent it modifies is contrasted with the previous turn. Turn-initial “but” almost always relates across turns.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 34–41.

corrections can be modelled in the IS model of dialogue. We model correction in the Information State (IS) framework, assuming the PTT (Poesio and Traum, 1998) model of dialogue. We motivate our work by showing how it updates beliefs and facilitates response deliberation.

2 Distinguishing Different Types of Correction

Assuming that "but" corrects material in the previous turn, we will consider previous turns which are assertions, questions, commands and those which communicate implicit information like discourse relations. What we will investigate here is how cross-speaker "but" can correct utterances communicating different types of SAs. We investigate whether it corrects the proposition itself or whether it corrects the relations stated or inferred from interpreting the role of the SA.

2.1 Correcting Assertions

Assertions can involve a large range of propositions including propositions in which the clauses are related via local (intra-sentential) discourse relations e.g., causal relations, temporal ones, exemplification, etc. For example, speakers can assert that event X happened before event Y, or that something caused something else to happen, etc. Assertions can also be related to other turns in the preceding dialogue, or to information inferred from the context (including preceding discourse). For example, speakers can assert something as an example of some prior rule expressed in the dialogue, or as a cause for some prior event, etc. The example below involves A expressing a reason why chairs have four legs, which is a local causal relation.

(4) A: Chairs have four legs for stability.
B: But three-legged chairs can be equally stable, they just need to be well-designed.

Often asserted propositions relate to the preceding discourse, e.g., assertions that serve as answers to questions, as illustrated in the next subsection. So correcting assertions involves interpreting some relation asserted in the preceding turn, or the relation inferred by the corrector between the assertion and the preceding discourse or simply within the assertion itself (as above). Simply asserted material can also be corrected, as in the following two examples, where the assertions themselves (not relations within the assertion, as in the example above) are being corrected:

(5) A: Dogs are trustworthy animals.
B: (Oh,) But they just seem that way because they're dumb.
B': No they're just too stupid to be devious.

(6) A: The train from Brisbane gets in at 11 o'clock.
B: (Oh,) But that's the Perth train! The Brisbane train only gets in at 1.
B': No it doesn't, you're thinking of the Perth train.

Notice how both these examples are more definitively negated given a "no" as in B'. Indeed the "but" does not easily correct assertions since it logically communicates conjunction with the implicature of contrast/opposition. These "but" corrections of assertions often seem more understandable with the "Oh" beginning the turn, signalling surprise at the previous turn and indicating understanding and possibly also partial acceptance of the assertion, and disagreeing with either its content or role in the discourse. Disagreeing with an assertion's argumentative stance, usage, or role in the discourse (given discourse history) is similar to corrections of discourse relations which are discussed in the next few sections.

2.1.1 Correcting Answers

Answers to questions are SAs responding to preceding SAs (relational SAs, following (Poesio and Traum, 1998), and "but" can correct the answerhood relation itself, as presented in the correction conversation adapted from (Asher and Lascarides, 1998), which is simplified below:

(7) A: Why did (John get sent to jail)? r
B: (He was caught embezzling funds from the pension plan). p
C: Yes, (BUT (he went to jail) r because (he was convicted of tax evasion) $q1$) $q2$

The *why*-question here enables interpretation of B's turn as indicating a *reason* (and therefore an answer) to the situation for which A asks an explanation. We will take r to be the situation queried about in A, so r is "John got sent to jail". Interpreting B's assertion as an answer to the *why*-question, we have B communicating $reason(p, r)$. O Here the question sets up an expectation of an answer, which constrains interpretation. We interpret p as (1) an answer expressing (2) the reason for r and also, more basically (3), as an assertion of the situation described in B. Then C's "yes" is interpreted as accepting B's assertion, and the "but" clause ("but he went to jail because he was convicted of tax evasion", $q2$ in the example) indicates an alternative answer w.r.t. B's turn, since the first part of $q2$, "he went to jail" is r restated, and is explained (cued by "because") by $q1$ ("he was convicted of tax evasion"), which is expressed as a

Table 1: Graesser's Question-Types

Question	Abstract Specification
Comparison	How is X similar/different to/from Y?
Definition	What (category/properties) does X have?
Interpretation	How is an event interpreted/summarised
Feature Specification	What value/attribute does feature X have?
Causal Antecedents	What caused event X to occur?
Causal Consequents	What are the consequences of an event/state?
Goal Orientation	What are an agent's goals/motives?
Instrumental	How (plan) does an agent accomplish a goal?
Expectational	Why did some expected event not occur?

reason for r, so we get $reason(r, q1)$. Since we know that for a why-r question, anything which involves $reason(r, X)$ is an answer to the question, we can interpret this as an alternative answer to A's question. The "but" indicates that C's utterance is somehow contradictory to B's utterance, so we interpret this alternative answer as a correction of B, i.e., both (1) a rejection of B's utterance as an answer to A, and (2) the assertion of an alternative answer to A, which we will use as a rough definition for corrections like these.

The presence of the cue ("but") in Ex. 7 indicates that inferring that B's answer is wrong is obligatory[2]. However in either the cued or non-cued case, the inferred relation is still defeasible.

We take Graesser's taxonomy of inquiries (1992) as a basic set of question types and omit the categories in Graesser's taxonomy which involve single-valued (e.g., slot-filling) answers and prefer those which tend to require answers which attribute some predicate to a subject (so we have sentential rather than phrasal answers); see Table 1. Since several categories in his taxonomy have questions beginning with "what", some of which can also have *why*-questions, and the remaining three categories have *how*-questions, we will assume that we have the appropriate machinery to resolve several question-types. The benefit of using the taxonomy is that it provides us with useful clues about the nature of the answer, supposing the answerer to be honest and helpful (following Gricean reasoning).

2.1.2 Correcting Implicit Information

Implicitly communicated information like discourse relations, denied expectations, inferences, and defeasible rules can all also be corrected following our definition of correction where something is (1) either denied or rejected and (2) an alternative/replacement or explanation is presented. We will briefly explore correction of implicit information here.

Recall Ex. 4 where A's assertion communicates a reason for the stability of chairs, which is then refuted by B. Here B refutes the *reason* relation communicated by A by directly refuting the inference that A communicates that $four\_legs(chair) > stable(chair)$. Similarly, Ex. 7 showed that C could agree with the assertion B made but disagree with the inferred answerhood function of the assertion. The idea here is that implicit information like discourse relations can be corrected while not necessarily refuting what is explicitly asserted.

2.2 Correcting Presuppositions in Questions

What does it mean to correct a question? While there is much work involving the semantics of questions (e.g., (Asher and Lascarides, 1998), (Ginzburg, 1995), (Ginzburg, 1996), etc.), Asher and Lascarides argue that much of it falls either into the realm of dialogue planning or formal semantics, and neither type of approach bridges the gap in order to explain examples like the one below:

(8) A: How do I install the modem drivers in Linux?
B: But you're getting a Mac, so you don't need to install anything.

In order to get at how questions can be corrected, we will first consider some conditions on how they can be answered. Asher and Lascarides' approach to question-answering follows from the multiple notions of answerhood put forward in (Ginzburg, 1996), namely that (1) the information fully resolves the question, defined in terms of the interpreter's goal and mental state, or (2), that the information potentially resolves the question. That is, Ginzburg's notions of answerhood rely on context sensitivity and interpreter-specific responses, and his analysis identifies a proposition at the centre of the question, e.g., in the question above, the proposition would be "I install the modem drivers in Linux". In this sense, correcting a question itself would be very similar to correcting an assertion[3] except that it would involve adjustments to the answerer's obligations.

Additionally, different question types all expect specific answers, e.g., consider *why, how, what, when,* and *where* questions; in the case of "why X", the answer is usually a reason for X, where X is some proposition describing a state of events

---

[2] Thanks to a reviewer for this point.

[3] Thanks to Colin Matheson for this point.

or situation. “How X” expects an answer that provides a manner in which X might be accomplished, or a way to perform/achieve (X), where X is again a proposition containing a state of events. “What X” questions are less specific and X is often a phrase which is usually the subject of a transitive verb, so that the answer provides the direct object.

Of course these sketches of question-answering are very rough, and there is far more extensive work on the subject. The point here is that if we adopt Graesser’s taxonomy of inquiries discussed earlier in the section on answers, we get much more specific requirements for answerhood, and also, a clearer set of question-types. In any case, our goal is to see how this information can inform a more specific characterisation of the types of corrections that are coherent given the preceding question’s context. Given such specific information about what a question addresses and what sort of answer it expects, it becomes less difficult to see how the questions in Table 1 can be corrected. One way that correcting questions is different from answers is that the former involves pointing out how the question itself is invalid/incorrect/irrelevant (i.e., by providing an explanation for this incorrectness), rather than by disagreeing with and providing an alternative answer. Here are some example corrections responding to the first few categories in Table 1:

(9) Comparison: But X is the same as Y!
Definition: But X is undefined!
Interpretation: But it (the event) never happened!
Feature Spec: But I already told you!
Causal Ante: But nothing caused X to occur!

We notice here that corrections can deny the basis of the question, for example, that a comparison is valid in Comparison (above). Correction can also prove the question invalid, as in the Interpretation case, where it is impossible to interpret an event that never happened. Likewise, corrections can address meta-level issues as in Feature Specification above; here the corrector indicates that the question itself has already been answered. Notice the strong role played by the question category above; in many cases the correction hinges on the validity or relevance of the question category itself.

Another sort of correction of questions involves incorrect assumptions of slot-values in the question. Correcting misassumptions communicated in the question signals a difference in speakers’ beliefs, as seen below:

(10) A: When did you want to fly back from Boston?
B: But I want to fly back from New York!
B′: But I want a rail ticket!
B″: But I don’t want to leave Boston at all!

In B a slot-value is corrected, namely place of departure, and then presumably the question becomes valid. B′ indicates that the verb slot-value is incorrect; it should be “travel by train” instead. B″ questions the validity of the question itself by correcting the proposition at the centre of the question, since A asks when B wants to travel and B does not want to travel at all. This highlights an important point: questions presuppose the truth of their central propositions.

Precondition failure, constraints or mutually exclusive situations are also reasons why a question may be corrected. In the example below, B wants to go for a walk now and Hilda is not with B and therefore cannot come:

(11) B: I’m going for a walk.
A: Will you take Hilda with you?
B: But she’s at school now and I can’t wait.

B answers A while correcting A’s assumptions of B’s goals. We consider this correction because it involves rejection of A’s inferred goal (namely that B takes Hilda with her) and also provides an explanation for why the question is deemed invalid. So the ways in which questions can be corrected are as follows:

- Correcting an incorrect slot-value (Ex. 10)
- Indicating that the question is not valid
  - Because a necessary criterion/precondition/ constraint has not been met (Ex. 11)
  - Because the question-type itself does not apply to what is being asked about (See 9, e.g.: Comparison, Definition, Causal Antecedent)
  - Because it was already asked and so is redundant (E.g., Feature Specification)

2.3 Correcting Commands

Corrections of commands also often involve correction of what is presupposed by the command. Interactions between speakers’ plans often provide the source of disagreement evident in corrected commands. In the example below, B signals precondition failure (in order to shut the door, it must be open) and corrects A by asserting this.

(12) A: Shut the door.
B: But it’s already shut.
B′: But then it’ll get too hot; why don’t we shut the window instead?
B″: But then it’ll get too hot.

B′ communicates an undesirable effect of performing A’s commanded action, and proposes an alternative. B″ just communicates the undesirable effect and does not propose an alternative solution, and we do not consider this correction. We will assume that correcting commands like questions and assertions also involves both (1) disagreeing, rejecting or denying something in the previous turn and (2) proposing an alternative/replacement or explanation. The question then arises as to why B above is considered a correction, since it does not provide an alternative. We argue that since B shows that A’s action is invalid or impossible, it provides an explanation, similar to how the various question categories in Ex. 10 were deemed irrelevant or invalid. So in both these situations, the correction involves asserting what or why the question or command is invalid or irrelevant. The difference between B$''$ and B above is that B$''$ only presents an undesirable effect and neither invalidates the command nor rejects and presents an alternative or explanation, so it is not deemed as correction. B on the other hand asserts that the commanded action is impossible (rejection/denial/disagreement) and presents an explanation, and so can be regarded as a correction.

Speakers can also propose better alternative actions, as seen below, where the rationale behind A's command is brought into question by the contradictory fact that the flies are getting in through the windows more than the door:

(13) A: Shut the door so the flies don't come in.
B: But more flies are coming through the windows; if we shut those instead we'll still have a breeze.

Speaker B can also object to the discourse relation inferred from A's assertion. In the example below, B infers that A's commanded action is intended to achieve the goal of keeping the flies out.

(14) B: These flies are really getting to me.
A: So shut the door.
B: But they are coming in through the windows.

In this case, what is actually being corrected is the inferred relation between the command and the problem mentioned in B(1). B(2) is correcting A's assumption that the flies are getting in through the door, and therefore corrects A's solution relation to B's problem, rather than the command A issues by showing that A's solution is irrelevant.

2.4 Related Relations

Correction differs from the other cross-turn "but" signalled relations concession and denial of expectation (DofE) in that it does not involve inferring relations between the turns themselves. Cross-turn DofE involves the "but" speaker denying an expectation triggered from the previous turn, while cross-turn concession involves the "but" speaker arguing in opposition to the other speaker w.r.t. a salient claim under discussion (Thomas, 2005).

Correction can also be distinguished from denial and rejection and seen as a relation which is composed of both denial or rejection and communication of an explanation or alternative. Many of the examples in the previous section involved rejection of offers, negotiation (by introducing alternative possible actions), misunderstandings, argumentation (by proposing other arguments), etc. We can distinguish denial as countering the perceived truth-value of an assertion, while rejection turns down an offer, and is thus only seen where commands or offers are issued, and both fail to introduce new explanatory or corrective information. For example, if C wants to deny B's answer in Ex. 7, he simply needs to say something along the lines of "No he wasn't (caught embezzling funds from the pension plan)". A rejection of a command like "shut the door" would simply involve an asserted "no!". Denial and rejection almost never involve "but", since they do not involve any acceptance beyond understanding of what is denied or rejected.

3 Modelling Correction

Given the discussion so far, we now show a generalised sketch of the update procedure for correction. We model correction in the Information State (IS) framework, assuming the PTT (Poesio and Traum, 1998) model of dialogue. In the procedure below, CDU, PDU and UDU refer to current, previous and ungrounded dialogue unit respectively. DH refers to the dialogue history field. CA_j etc. refer to conversational acts (CAs, see (Poesio and Traum, 1998)).

If [CDU.DH has CA_z of the form $assert(but[Z])$ AND a CA_y of $[reject(CA_j)$ OR $disagree(CA_j)]$] AND [PDU has CA_w of the form $SA(W)$]

1. If $alternatives(Z, W)$ then update CDU with CA_x: $correct\ (current\ speaker, W)$
2. Else if $explanation(Z, W)$ then update CDU with CA_x: $correct\ (current\ speaker, W)$

SA will be replaced by $command$, $question$, $assert$, or a discourse relation depending on what the SA of the prior turn (PDU) is. We assume a CA interpretation procedure (see (Thomas, 2005)) which takes CAs and compares their contents w.r.t. the speakers' beliefs, plans, etc. and updates the IS with $alternatives$ or $explanation$ depending on the case. Alternatives generally provide non-identical information relating to the same topic which is not necessarily conflicting or mutually exclusive, though in the former case, the corrector generally assumes that his alternative is more appropriate or relevant to the given situation. Explanations in the case of correction generally indicate why something is incorrect. In order to determine either alternation or explanation, the CA interpretator needs to call a theorem prover with the two related arguments. The above sketch of the update algorithm does not account for answers, which would need a question in UDU to be specified and where PDU and CDU pose as alternate answers.

Applying the update procedure to Ex. 7 produces the IS shown in Fig. 1, illustrating the results of the update procedure. We omit irrelevant fields and acts here for brevity, and just show the relevant part of the IS[4]. The condition (COND field) says that if B accepts C's correction, then she is socially committed (see (Matheson et al., 2000)) to accepting C's answer.

We will now show a more specific procedure to address corrections of answers to why-questions:

If [UDU.DH contains an $ask(why[X])$ SA in CA_i] AND [PDU.DH has a CA_j that is of the form $assert(Y)$] AND [CDU.DH has a CA_z of the form $assert(but[Z])$] AND $CA\_interpreter$ (IS, CA_j, R) returns $R = answer$ $(speaker(CA_j), CA_i)$ then

1. If CDU.DH contains a CA_k with the SA $assert(Z)$ and if the $CA\_interpreter$ $(IS, CA_k, R2)$ returns $R2 = answer$ $(speaker(CA_k), CA_i)$ and $CA\_interpreter$ $(CA_j, CA_k, R3)$ returns $R3 = alternatives$, then add to CDU.DH CA_m: $reject($ $speaker(CA_k), R)$ and CA_n: $correct$ $(speaker$ $(CA_k), R)$. Add to CDU.COND $accept$ $(speaker$ $(CA_j), CA_n) \rightarrow scp(speaker(CA_j), R2)$
2. Else if CDU.DH contains a CA_k with the SA $assert(Z)$ and if the $CA\_interpreter$ $(IS,$ $CA_k, R2)$ returns $R2 = explanation$ $(speaker$ $(CA_k), CA_i)$, then add to CDU.DH CA_m: $reject$ $(speaker(CA_k), R)$ and CA_n: $correct$ $(speaker$ $(CA_k), R)$. Add to CDU.COND $accept$ $(speaker$ $(CA_j), CA_n) \rightarrow scp(speaker$ (CA_j), $R2)$

Notice that we do not address here whether the corrector accepts the assertion in PDU or not. We assume that the CA interpreter will determine disagreement when PDU is processed; i.e., it will check the speaker of CDU's beliefs, intentions, etc. to determine if there is any conflicts before updating the IS with accept/reject CAs. This procedure can be easily adapted to address questions of other types, and Graesser's taxonomy can be incorporated so that, provided we can parse questions into his categories, we know what sort of answers to expect, which can be utilised by the CA interpreter and theorem prover when determining if something is an answer or not. Lastly, we present the procedure for interpreting correction of questions:

If CDU.DH contains a turn-initial "but" and PDU.DH contains CA_j: $question$ $(speaker[PDU], X, T)$ (where X is the proposition at the centre of the question and T is the question's category in Graesser's taxonomy)

1. If CDU.DH contains CA_k: $disagree$ $(speaker$ $[CDU], part\_of[X])$[5] and also CA_l: $assert$ $(speaker$ $[CDU], Y)$ and calling the theorem prover with $part\_of(X)$ and Y returns that they are alternate values for the same attribute or that both share a topic, then add to CDU.DH CA_m: $correct$ $(speaker$ $[CA_k], part\_of[X])$
2. Else if CDU.DH contains CA_k: $disagree$ $(speaker$ $[CDU], X)$ and also CA_l: $assert$ $(speaker[CDU],$ $Y)$ and calling the theorem prover with X and Y returns that they are alternate values for the same attribute or that both share a topic, then add to CDU.DH CA_m: $correct$ $(speaker[CA_k], X)$
3. Else if CDU.DH contains CA_k: $disagree$ $(speaker$ $[CDU], CA_j)$ and also CA_l: $assert$ $(speaker$ $[CDU], Y)$ and calling the TA interpreter with TA_j (the TA equivalent for CA_j), X and Y returns that Y is a precondition/ constraint to be overcome of TA_j, then add to CDU.DH CA_m: $correct$ $(speaker$ $[CA_k], CA_j)$
4. Else if CDU.DH contains CA_k: $disagree$ $(speaker$ $[CDU], CA_j)$ and also CA_l: $assert$ $(speaker$ $[CDU], Y)$ and calling the theorem prover with CA_j and Y returns that $reason$ $(incompatible$ $[X, T], Y)$, then add to CDU.DH CA_m: $correct$ $(speaker[CA_k], CA_j)$

Cases (a) and (b) above differ depending on whether X or part of X is disagreed with, as is illustrated in Ex. 10 turns B and B$'$ (part of X) and B$''$ (X) respectively, and the corrector provides alternate information (Y) for X with respect to the attribute they describe or the topic (T) that they relate to. Case (c) addresses situations in which the corrector introduces a precondition or constraint that must be satisfied before the question makes sense; for example, B could say "But I need to book when I'm leaving for Boston before I know when I want to leave" in response to A's question in Ex. 10. Case (d) addresses situations in which the question-type is somehow incompatible with what is being asked about (in the topic, T), and the corrector gives a reason about why this is the case, e.g., see Ex. 9 for an illustration of how different question types (following Graesser's categories given in Table 1) may be corrected. The point to using Graesser's categories here is to reduce the workload of the theorem-prover by checking for specific types of questions to isolate how they are corrected. On another point, it is important to note that none of these tests for correction will always be completely accurate, since contexts can always be found where such situations do not need correcting. However this is not something to worry about, because even if correction is erroneously predicted, the interpretation of such relations hinges on the notion that they are defeasible (i.e., cancellable), so that even if we have an incorrect interpretation, this can be cancelled explicitly by the correctee in a subsequent

---

[4] The abbreviated fields are: Previous and Current Dialogue Unit (PDU and CDU), Ground (GND) and Conditions (COND) following the IS structure given in (Matheson et al., 2000).

[5] Or $reject$; both are interchangeable for the purposes of this paper. The same holds for $alternate$ and $replacement$.

turn. In the next section we will see how the correctee can make use of the updates presented here to respond to the correction based on her own beliefs.

4 Deliberating Responses to Correction

The type of CA which is corrected in the $correct$ CA will give a lot of useful information about what the corrector believes so that the hearer can respond appropriately, depending on how her own beliefs relate to what she infers about the corrector. We will assume here that the most informative information available in the case of a correction will be found in the $correct$ CA itself, since this indicates which CA (in the previous turn) is being corrected, and further information as to why or providing an alternative will be found in an $assert$ CA along with the $correct$ CA. The focus on just three CAs greatly simplifies deliberation of a response to correction.

Now let us consider the various facets of response one can make to a $correct$ CA. Correctees can agree or disagree, they can do so partially or wholly, and they can agree or disagree with content explicitly or implicitly or with the relevance of the content. Furthermore, correction could involve new information to the correctee or highlight common knowledge and illustrate its relevance. In any case, the correctee must interpret what is being corrected and then respond, either by accepting that the corrector is right, or by disagreeing and explaining why.

Depending on what is corrected, the correctee needs to determine his own standpoint w.r.t. the correction by checking his beliefs, via the appropriate fields of the IS representing his beliefs, plans, what he has said, etc. Deciding to accept correction involves comparing the new (corrected) information with his original views in order to determine if the new information is compatible or not given his own beliefs and reasoning. While it is possible for a speaker to accept a correction and revise his beliefs, he should also be able to disagree if he knows of something wrong or incompatible in the corrected information. In practice this will involve theorem proving or reasoning about plans, so assuming such reasoning, we consider how one might deliberate a response based on the results of this reasoning. We will first discuss an example and then formulate a more general outline of what must be considered in responding to corrections. Recall Ex. 7. Here C accepted B's assertion p but disagreed with it as an answer to why John got sent to jail (r). An alternative reason for why he went to jail is also asserted by C (i.e., $q1$). If B accepts this reason, he can say so, e.g.: "Oh, I didn't know he also got caught for tax evasion." If B does not accept C's assertion ($q1$), he can refute it, e.g., by saying "No he wasn't. They dropped the charges." He can also accept C's reason and assertion but comment on the relevance of the correction and add new information, e.g., "Yes, but he also went to jail for embezzling funds from the pension plan and that's what A was asking about." We outline below some of the considerations a procedure modelling deliberation of responses to corrections should have:

If CDU.DH contains CA_j: $correct(speaker[CDU], CA_i)$, CA_k: $assert(speaker[CDU], X)$ and CA_l: $[disagree/reject(speaker[CDU], CA_i)$, and if PDU.DH contains CA_i:

1. Check PRIVATE BELIEFS, TASK BELIEFS (TB) and PDU.DH for speaker(PDU) for any CAs, beliefs, etc. held by speaker(PDU) which conflict with X. If there are conflicting beliefs update CDU with these conflicting beliefs. They should be added either to PRIVATE BELIEFS, TB or INT[6] (depending on where the conflict arose) as elements of the form $conflict(X, Z)$ where Z is the new information.
2. For all elements W of the form $conflict(X, Z)$ in CDU.PRIVATE_BELIEFS, CDU.TB or CDU.INT, push these onto INT in the form of intentions to $assert(speaker(PDU), W)$ (speaker(PDU) is the next speaker).
3. Push contents of PDU onto UDU and CDU onto PDU. CDU will contain the response to the correction, and will show the conflicting assertions in CDU.INT.
4. Deliberate over which intentions to assert should be expressed next (among other things) and generate a response.
5. After expressing $assert(speaker(CDU), W)$, remove $assert(speaker(CDU), W)$ from INT. Then move W into a CA of the form $raised[conflict(X, Z)]$ in CDU.DH.

This procedure first compares what is corrected and asserted by the corrector with the beliefs of the correctee in order to update the IS with a list of the resulting conflicting information found in DH (since the correctee might have previously uttered conflicting information), Private Beliefs in NTOD or Task Beliefs in TOD. The conflicts are then turned into intentions to assert them and pushed onto the correctee's INT field. After deliberating over which intentions to assert should actually be expressed next and these assertions of conflict are expressed, they are removed from INT

---

[6] If Intentions (INT field) are updated, they will trigger a response to be generated in the update rules.

C: Yes, but he went to jail because he was convicted of tax evasion.

$$\left[\begin{array}{ll}
\text{UDU} & \left[\ \text{DH [CA1: } \mathit{ask}(A,(B,C),\mathit{why}(\mathit{sent}(\mathit{John},\mathit{to}(\mathit{jail}))))]\ \right] \\
\text{PDU} & \left[\begin{array}{ll} \text{DH} & \left[\begin{array}{l} \text{CA2: } \mathit{assert}(B,\mathit{caught}(\mathit{John},\mathit{embezzling}(\mathit{funds},\mathit{from}(\mathit{pension\_plan})))) \\ \text{CA3: } \mathit{answer}(B,CA1) \end{array}\right] \end{array}\right] \\
\text{CDU} & \left[\begin{array}{ll} \text{DH} & \left[\begin{array}{l} \text{CA4: } \mathit{accept}(C,CA2) \\ \text{CA5: } \mathit{assert}(C,\mathit{reason}[\mathit{sent}(\mathit{John},\mathit{to}(\mathit{jail})),\mathit{convicted}(\mathit{John},\mathit{tax\_evasion})]) \\ \text{CA6: } \mathit{answer}(C,CA1) \\ \text{CA7: } \mathit{reject}(C,CA3) \\ \text{CA8: } \mathit{alternatives}(CA6,CA3)\ \text{CA9: } \mathit{correct}(C,CA3) \end{array}\right] \\ \text{COND} & [\mathit{accept}(B,CA9) \rightarrow \mathit{scp}(B,CA6)] \end{array}\right]
\end{array}\right]$$

Figure 1: IS for Example 7

and CDU.DH is updated with a CA indicating that the conflict was raised. A benefit of this approach is that conflicts (raised by corrections in this case) are always straightened out without delay, hopefully reducing the number of misassumptions which might otherwise occur.

5 Conclusions and Future Directions

In this paper we present a novel treatment of cross-speaker correction when signalled by “but”. We started out by presenting a range of corrections in assertions, answers to questions, implicit relations, questions and commands in the hopes that a central treatment for correction would emerge based on the notion that SAs are wrappers around a central proposition. We saw that both the SA and the central proposition can be corrected, and that in all cases, the defining characteristics of correction involve (1) disagreement with or rejection of something which is communicated in the previous turn and (2) the assertion of either an explanation or an alternative perceived to be more appropriate/relevant/correct by the corrector. We then presented procedures describing how correction can be interpreted given differences in beliefs etc. discernable by the theorem-prover and CA interpreter, and how this interpretation updates the IS representation of the dialogue with the correction relation. We finished by showing how these updates enable the correctee to respond to the correction depending on her own beliefs, etc. The motivation here is that by interpreting speakers’ goals and beliefs and explicitly accounting for them, this approach helps to avoid potential misunderstandings.

In future work we hope to extend this analysis to account for multimodal corrections, in order to determine how the general approach can be extended to account for information communicated in nonlinguistic modalities as well as by linguistic means.

References

J. Allen and M. Core. 1996. DAMSL: Dialog Act Markup in Several Layers. University of Rochester.

N. Asher and A. Lascarides. 1998. Questions in Dialogue. *Linguistics and Philosophy vol.23:2.*

J. Austin. 1962. How To Do Things With Words. *Clarendon Press, Oxford.*

L. Carlson. 1985. Dialogue Games. *D.Reidel Publishing Company.*

J. Ginzburg. 1995. Resolving Questions, parts 1 and 2. Linguistics and Philosophy, vol.8.

J. Ginzburg. 1996. Interrogatives: Questions, Facts and Dialogue. In The Handbook of Contemporary Semantic Theory, Ed. S. Lappin.

A. Graesser, N. Person, and J. Huber. 1992. Mechanisms That Generate Questions. In Questions and Information Systems, Eds. T. Lauer, E. Peacock, and A. Graesser.

H.P. Grice. 1975. Speech Acts. *In Logic and Conversation, Eds. P. Cole and J. Morgan.*

J. Kreutel and C. Matheson. 2001. Incremental Information State Updates in an Obligation-Driven Dialogue Model. *Language and Computation.*

L. Lagerwerf. 1998. Causal Connectives Have Presuppositions. *Catholic Univ. of Brabant, Holland Academic Graphics, The Hague, The Netherlands.*

N. Lesh, C. Rich, and C. Sidner. 1999. Using Plan Recognition in Human-Computer Collaboration. *International Conference on User Modelling.*

D. Litman and J. Allen. 1987. A Plan Recognition Model for Subdialogues in Conversation. *Cognitive Science*, volume 11.

C. Matheson, M. Poesio, and D. Traum. 2000. Modelling Grounding and Discourse Obligations Using Update Rules. *Proceedings of the North American Association for Computational Linguistics.*

M. Poesio. and D. Traum. 1998. Towards an Axiomatization of Dialogue Acts. *Proceedings of Twente Workshop.*

E. Sweetser. 1990. From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. *Cambridge University Press, Cambridge.*

K.E. Thomas. 2005. *But What Do They Mean?* Modelling Contrast Between Speakers in Dialogue Signalled by “But”. Doctoral thesis, *School of Informatics, University of Edinburgh, Scotland.*

From Complex to Simple Speech Acts: a Bidimensional Analysis of Illocutionary Forces

Claire BEYSSADE
Institut J. Nicod
CNRS-EHESS-ENS
1 bis, avenue Lowendal 75007 Paris
`claire.beyssade@ehess.fr`

Jean-Marie MARANDIN
CNRS-Paris 7
Case 7003 - 2, place Jussieu
75251 Paris cedex 05
`marandin@linguist.jussieu.fr`

Abstract

We present a new analysis of illocutionary forces in dialogue. We analyze them as complex conversational moves involving two dimensions: what Speaker commits herself to and what she calls on Addressee to perform. We start from the analysis of speech acts such as confirmation requests or whimperatives, and extend the analysis to seemingly simple speech acts, such as statements and queries. Then, we show how to integrate our proposal in the framework of the Grammar for Conversation (Ginzburg, to app.), which is adequate for modelling agents' information states and how they get updated.

1 Introduction

One usually assumes a one-to-one relationship between clause types and illocutionary forces, as summarized in (1):

(1) a. The declarative type is associated with asserting.
b. The interrogative type is associated with questioning.
c. The imperative type is associated with requesting.
d. The exclamative type is associated with exclaiming.

But, it has been observed (i) that the same utterance can be used to perform different speech acts, and (ii) that an utterance may simultaneously convey more that one speech act. It is the case in particular for confirmation requests, which have been analyzed as 'a superposition' (a.o. Fontaney (1991)), or a composition of assertion and question (a.o. Asher and Reese (2005)).

In this paper, we leave aside the thorny grammatical issues raised by (1), (references will be given in the talk); we concentrate instead on how to analyze illocutionary forces in a dialogical perspective. We claim that utterances impact dialogue in two ways that we describe in terms of update. On the one hand, Speaker commits herself to some content: uttering amounts to update Speaker's commitments. On the other hand, Speaker calls on Addressee for him to change his own commitments. As Strawson puts it, an act of communication goes through if it is taken up as intended. Speaker's intention cannot be reduced to showing her own private knowledge, belief or desire; rather, Speaker intends to change the context by adding or removing something (a fact, a question...) in the shared ground. Thus, a second update concerns what Speaker requests from Addressee (that Addressee consider a new proposition as true or consider an issue as relevant for current purposes in the dialogue...). Speaker's and Addressee's updates are usually and tacitly assumed to be identical. However, we do claim here that such an identity is not always the case, and that these updates should be explicitly distinguished when analyzing illocutionary forces and the speech acts they give rise to.

We articulate our proposal in the framework of the Grammar for Conversation (Ginzburg, to app.), which (i) assumes a rich ontology of semantic content and (ii) accounts for the asymmetry between Speaker and Addressee in dialogue. We model illocutionary forces as types of moves in conversational games. Conversational moves lead from one set of shared commitments to another set of shared commitments. Commitments may be added or removed (as when a question is answered or a command is carried out). Illocu-

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 42–49.

tionary forces can be viewed as commitment change potentials.

Our point of departure is the description of phenomena that have been described in terms of polyfunctionality or indirect speech acts (§2). We borrow our examples in English from the literature; data in French are partly elicited or taken from written or speech corpora. In §3, we present some recent analyses and a critique of them. In §4, we show how to revisit the notions of illocutionary force in general, and we sketch out the modelling of the proposal in the Grammar for Conversation.

2 Challenging Data

2.1 Utterance Polyfunctionality

Since Gazdar (1981), it has been taken for granted that clause types do not determine the illocutionary type of the utterances in a one-to-one manner (as in (1)). Gazdar's example is (2). As observed by Gazdar, (2) supports a great number of speech acts, and more importantly, speech acts of different types (assertions, questions or directives).

(2) A.: You will go home tomorrow.

The polyfunctionality of (2) out of context is reflected in the gamut of responses Addressee may perform his turn when taking up (2).

(3) B.: a. How do you know?
b. Yes.
c. That's what you think.
d. Okay.

Gazdar's line of reasoning echoes the literature about so-called indirect speech acts. (4a) in context may be taken up as an order (roughly (4b)); (5a) as an assertion (roughly (5b)); and (6a) as a directive (roughly (6b)).

(4) a. It is cold in here!
b. Close the windows!
(5) a. Will the sun rise tomorrow?
b. Of course, yes.
(6) a. Can you clean up your room?
b. Clean up your room!

2.2 Sorts of Indirect Speech Acts

Green (1975) draws a crucial distinction among so-called indirect speech acts when discussing the directive use of interrogatives: hints *vs* whimperative constructions. The closed interrogative (7a) may be used as a directive only in a context where the addressee is expected to take away the garbage at a certain time, whereas the open interrogative (7b) conveys a directive in all contexts. According to Green, a reply to (7b) with *because* or whose content could be interpreted as a reason would not only be non-felicitous, but it would show a poor competence of English.

(7) a. Have you taken away the garbage?
b. Why don't you be quiet!

The distinction pertains to two types of resource Addressee may use to come up with an illocutionary assignment: (i) either context knowledge bearing on Addressee, Speaker, their relations in the world and the current conversation or (ii) grammaticalized features of the utterance which indicate how it should be taken up. The directive import of (7a) (or (2), (6a) above) -- Green uses the labels *hint* to refer to them-- is arrived at through a chain of inferences.[1] The directive import of (7b) is directly brought about by a construction (which is a specification of an interrogative construction), a whimperative construction. The whimperative construction illustrated in (7b) features *why*, inverted *do* in the negative and a verb in the base form.

The point here is that the type of update Speaker calls on Addressee for him to perform may be marked by grammatical means that do not belong to the clause type. There are two kinds of markers that specify the type of update Speaker calls on Addressee to perform:

- constructional markers such as the whimperative constructions of English;[2]
- lexical markers such as what we call 'tags' here and that we illustrate below in French.

Tags such as *n'est-ce pas* ['isn't it?'] or *s'il te plaît* ['please'] in French fully specify the call on Addressee that is intended by Speaker. For example, (8a) or (8b), even out of context, are no

---

[1] The number of steps in a chain of inferences may vary. E. g., it is bigger in the case of directive declaratives such as (4a) than for the directive use of (2), since the content of the directive has to be inferred entirely in the case of (4a).

[2] Below, some examples of whimperative constructions in French:

(i) a. Veux-tu bien te taire !
Would you be quiet
Be quiet!
b. Pourquoi pas acheter une voiture ?
Why not buy a car
What about buying a car?

longer "polyfunctional": (8a) is a demand of confirmation, i.e. an utterance by which Speaker calls on Addressee to commit himself to the issue whether Addressee will go home tomorrow, whereas (8b) is a declarative directive by which Speaker calls on Addressee to bring about a state of affair in which he will go home tomorrow.

(8) a. Tu rentreras à la maison demain, n'est-ce pas ?
You will go home tomorrow, won't you?
b. Tu rentreras à la maison demain, s'il te plaît.
You will go home tomorrow, please.

There is a wealth of tags that vary crosslinguistically. Here, we give a very short list of them in French:
- *point barre* is compatible with declarative and imperative sentences, and specifies that the utterance is directive.

(9) Marie ne sortira pas, point barre.
Marie will not go out, POINT BARRE
Marie won't go out, period!
(10) Ferme ta gueule, point barre !
Shut-IMP your mouth, POINT BARRE
Shut up, period!

- *oui ou non* also marks a directive move, when associated with an interrogative sentence. (11a) clearly means (11b).

(11) a. Est-ce que tu viens, oui ou non.
Are you coming, yes or no
b. Hurry up!

- *sans indiscrétion* is grammatical in two clause types, viz. interrogative and declarative, but specifies only one type of call on Addressee, viz questioning. This is why it sounds odd to take up a turn tagged with *sans indiscrétion* with expressions used for statement uptake (12b).

(12) a. Sans indiscrétion, Marie est arrivée
Without indiscretion, Marie has arrived
Without indiscretion, has Marie arrived?
b. A.: Sans indiscrétion, Marie est arrivée
B.: # Ah bon / Je ne le savais pas
Oh really / I didn't know that

To sum up, we must mark off two distinct phenomena: speech act assignement (SAA) based on inferences and SAA based on grammatical (lexical or constructional) means. In the latter case, it is crucial to note that there are means to specify the call on Addressee intended by Speaker, i.e. how Speaker expects her utterance to be taken up by Addressee.

2.3 Dialogical Behaviour of Indirect Speech Acts

It has been observed that so-called indirect speech acts are not exactly equivalent to their direct counterparts; either their felicity conditions or their impact on dialogue making or the relations between discourse participants are different. This is correlated to the fact that they do not trigger the same array of responses as their direct counterparts do.[3] We present below some examples in French.

Directives conveyed by interrogative clauses (13a) do not have the same impact in context as directives conveyed by imperative clauses (13b). The use of interrogative clauses is reputed more polite than that of imperatives. The question then is what brings in the politeness effect.

(13) a. Pouvez-vous fermer la porte, s'il vous plaît ?
Can you close the door, please?
b. Fermez la porte, s'il vous plaît !
Close the door, please!

In the same way, directive declaratives (see (14b)) may only convey commands whereas the gamut of directives conveyed by imperatives ranges from commands to suggestions or permissions. This requires an explanation.

(14) a. Viens demain, s'il te plait !
Come tomorrow, please!
b. Tu viendras demain, s'il te plait !
You come$_{\text{FUTUR}}$ tomorrow, please

Questions conveyed by declaratives (15a) are not felicitous in the same contexts as those conveyed by interrogatives (15b).[4] Questioning declaratives are much more natural in situations where Speaker has good grounds to know the

---

[3] Green (ibid. : 138) reports that the response to whimperatives is different from the response to imperatives in English.

[4] See Gunlogson (2003) for English questioning declaratives.

answer. For example, in a situation where Speaker, upon entering the department office, sees Mary's personal belongings on her desk, (15b) would be odd whereas (15a) would be appropriate as a question to Mary's colleague already at work.

(15) a. Marie est arrivée, n'est-ce pas ?
Marie has arrived, hasn't she?
b. Marie est-elle arrivée ?
Marie is-SHE$_{\text{CLITIC}}$ arrived
Has Marie arrived?

To sum up, so-called complex speech acts are different from their simple counterparts. This should be accounted for.

3 Recent Analyses

Here, we only consider two recent proposals to account for the features characterizing speech acts: Asher and Reese (2005) and Ginzburg (to app.). Both analyses are limited to specific cases. Asher and Reese deal with biased questions and propose an analysis in terms of composition of illocutionary forces. Ginzburg deals with the interactive working of assertions and questions and proposes an analysis in terms of simple or double update of the dialogue participants' gameboard. Our own proposal will retain some of the insights generated by these two approaches.

3.1 Complex Speech Acts

Asher and Reese (2005) observe that certain questions convey an expectation by Speaker of a negative answer. Such questions are said to be biased. Asher and Reese's main idea is that biased polar questions convey both an assertion and a question; accordingly, they propose to analyze them as complex speech acts of the type *assertion * question.*

Sadock (1974) uses the distribution of discourse markers (DM) as a criterion to sort out assertions and questions. *After all*, for example, selects assertions: it is incompatible with neutral questions.

(16) a. It is fine if you don't finish the article today. After all, your adviser is out of the country.
b. # It is fine if you don't finish the article today. After all, is your adviser out of the country?

As to *by any chance* or *tell me,* they select questions, rather than assertions.

(17) a. # John, by any chance, owns a car.
b. Does John, by any chance, own a car?

Applying these tests to biased questions such as (18a) shows that they behave as assertions and questions. Asher and Reese have observed that (18a) is compatible not only with *after all* and *by any chance,* but even with both of them in a single utterance.

(18) a. Has John ever voted for a democrat?
b. After all, has John ever voted for a democrat?
c. Has John by any chance ever voted for a democrat?
d. After all, has John by any chance ever voted for a democrat?

They conclude from the felicity of (18d) that (18a) simultaneously conveys an assertion and a question.

Asher and Reese's proposal could be extended to confirmation requests (ConfR), such as (15a) above or (19) below. Confirmation requests could be seen as simultaneously conveying an assertion (by the way, syntactically, ConfRs are declarative clauses) and a question. Moreover, positive ConfRs are biased for the positive answer. Thus, (19) would be seen as conveying both the assertion that Marie has arrived and the issue whether Marie has arrived. Accordingly, ConfRs too would be associated with a complex speech act of the type *assertion*question*. At first blush, this could be corroborated by arguments such as those used by Asher and Reese in (18).

(19) a. Après tout, Marie est arrivée, n'est-ce pas ?
After all, Mary has arrived, hasn't she?
b. Dis-moi, Marie est arrivée, n'est-ce pas ?
Tell me, Mary has arrived, hasn't she?
c. Après tout, dis-moi, Marie est arrivée, n'est-ce pas ?
After all, tell me, Mary has arrived, hasn't she?

Unfortunately, such a corroboration is shaky since other tests using different lexical criteria (compatible either with declaratives or interrogatives) that show that (biased) polar questions and ConfRs cannot be identified: for example, compatibility with *n'est-ce pas* (*n'est-ce pas* is felicitous with declaratives only).

(20) a. # Marie est-elle (jamais) venue, n'est-ce pas ?
Did Mary (ever) come, N'EST-CE PAS
b. Marie est venue, n'est-ce pas ?
Mary came, N'EST-CE PAS

In the same manner, one could analyze whimperatives as *question*directive*. But, counterexamples analogous to (20) for ConfRs would plague the attempt.[5] Moreover, such an analysis would fail to account for the observations in section 2.3.

3.2 Speech Acts and DGB Update

Ginzburg's grammar for interaction is predicated on the idea that dialogue can be conceived of as a game. Each turn brings about a change in the ongoing dialogue: the type and content of each change are registered in a dialogue gameboard (DGB). Each dialogue participant keeps her own DGB; the dynamics of dialogue making is reflected in the updates of DGBs that DPs operate at each turn.

Ginzburg, who argues against dialogue game formulations that are exclusively stated as operations on the Common Ground, distinguishes between a set of propositions called FACTS, and a set of questions, called QUD. He proposes to capture the dialogical difference between assertions and questions in terms of updates of these two sets in the DGB.

The dialogical impact of questions is viewed as adding a question to QUD. Crucially, the dialogical impact of assertions is viewed as a double update: update of QUD and update of FACTS. Here, Ginzburg follows Stalnaker (1978): when a speaker makes an assertion, she is committed to a proposition p, hence the updating of FACTS. But, this does not exhaust the effect of asserting. An assertion comes through only if Addressee too becomes committed to the same proposition. This is where QUD comes in. Since asserting p requires Addressee accepting p, Ginzburg proposes that the call on Addressee for him to accept p be modeled as a polar question whether p derived from p. Hence, "in general, both asserter and her addressee do have the issue p? in QUD as a consequence of an assertion p" (Ginzburg, 1997).

In a nutshell, asserting involves the conversational move of questioning, and a conversational move that is specific to asserting (the updating of FACTS).

Indeed, Ginzburg's proposal implies that the dialogical behaviour of ConfRs is the same as that of regular assertions. Which is not true (reference to be given in the talk).[6] Morevoer, asserting is not the only type of speech act that crucially involves Addressee's uptake. In this respect, it is hard to see how to use QUD to analyze the impact of directives.

But, the general idea of analyzing speech acts (and types of speech acts) as updates of a dialogical gameboard is not threatened by the difficulties of Ginzburg's original analysis of asserting *vs* questioning.

4 Proposal

Our proposal draws on the insights embedded in both proposals presented in section 3. From Asher and Reese, we take up the idea that speech acts can be complex. From Ginzburg, we take up the idea that speech acts can be analyzed as update operations on DGBs.

More precisely, we propose that all types of speech acts (or illocutionary forces) are complex, since they involve a double update; the update of Speaker's commitments and the update of the call on Addressee.

Postulating that speech acts are bi-dimensional enables us to analyze in a unified framework both complex speech acts and their simple counterparts, while accounting for their differences in context.

4.1 Speaker's Commitment and Speaker's Call on Addressee

It has long been observed that speech acts have a 'double face': one pertaining to Speaker and another to Addressee. In pragmatic analyses, this is often expressed in terms of Speaker's attitude (belief, ignorance, desire, etc) on the one hand and the kinds of obligation[7] layed on Addressee on the other hand.

---

[5] For example:
(i) a. Veux-tu bien te taire, oui ou non !
b. # Tais-toi, oui ou non !

[6] For example, the contextual association with intonation contours is different for assertions, questions or ConfRs.

[7] Traum and Allen (1994) claim that when an agent is asked a question, this creates an obligation to respond. They propose to add the attitude of *obligation* to the more usual attitudes of belief, goal, and intention in modelling social interaction.

The attitudinal analyses of the Speaker-oriented aspects of speech acts face severe drawbacks (references to be given in the talk) and the links between grammatical forms and types of obligations imposed by Speaker on Addressee have not been established in a precise way.[8] This is why we will analyze speech acts as moves in conversational games. We take it that speech acts lead from one set of shared commitments to another set of shared commitments: commitments may be added or removed, as when questions are answered or commands carried out.

4.2 Modelling Speaker's Commitment

Following a suggestion made by Gazdar (1981), we extend Hamblin's notion of commitment, which is restricted to propositional content, in order to account for all general types of speech acts, i. e. questions, directives and exclamations.[9] Gazdar (1981) proposes that "an assertion that Φ is a function that changes a context in which the speaker is not committed to justifiable true belief in Φ into a context he is so committed. A promise that Φ is a function that changes a context in which the speaker is not committed to bringing Φ into one in which he is so committed. A permission to Φ is a function that changes a context in which Φ is prohibited into one in which Φ is permissible". We use Ginzburg and Sag's ontology and Ginzburg's modelling of dialogue to make explicit this extension of the notion of commitment.

The public part of Ginzburg's DGB comprises two 'slots' (along with LATEST-MOVE that we do not consider here): a set of propositions (FACTS), a set of questions (QUD). In order to account for outcomes conveyed by imperative utterances, we follow a proposal made by Portner (2005) and add in the DGB a specific slot TO-DO-LIST (TDL). TDL is partitionned into TDL(Speaker) and TDL(Addressee). TDL(Addressee) is an ordered list of descriptions of situations the actualization of which depends on Addressee and towards which Speaker is positively oriented. It is incremented with the outcomes that Speaker presents as actualizable by Addressee. TDL(Speaker) is incremented with the outcomes that Speaker presents as actualizable by herself: either the outcomes brought about by imperative utterances of interlocutors or those brought about by promissives.[10] Consequently, three dimensions are now distinguished in the DGB, each of them consisting in a homogeneous set (a set of propositions, a set of questions, or a set of outcomes).

When Speaker utters an assertion, she makes a move by which she becomes committed to a propositional content. By saying that Mary has arrived, Speaker presents herself as ready to stand for the truth of the proposition that Mary has arrived. This is a matter of public presentation which does not necessarily correspond to Speaker's private belief.

When Speaker utters a question, she makes a move by which she becomes committed to an issue. By asking whether Mary has arrived, Speaker presents herself as being interested for current purposes in the issue of whether Mary has arrived. Once again, this is a matter of public presentation and does not correspond to one specific knowledge state.

When Speaker utters a directive utterance, she makes a move by which she becomes committed to an outcome. Outcomes correspond to states of affair in the future, the actualization of which more or less directly depends on Addressee. Speaker's commitment consists in "the affirmative stance towards the actualization of this potential" (Stefanowitsch, 2003). By ordering Mary to arrive, Speaker presents herself as positively oriented to the realization of Mary's arrival.

To summarize, an assertion brings about the incrementation of SHARED GROUND, uttering a question the incrementation of QUD and, uttering a directive utterance the incrementation of TDL(Addressee).[11]

4.3 Modelling the Call on Addressee

Following Stalnaker, successful assertions are utterances that convey a call on Addressee for him to become committed to the propositional content Speaker commits herself to. Other types of speech acts carry a specific call on Addressee as well. By questioning, Speaker calls on Addressee to commit himself to the issue she is committed to. By uttering a directive, Speaker

---

[8] Truckenbrodt (2004) is an exception here, who presents another systematic way of capturing the call on Addressee dimension of speech acts.

[9] We will not account for exclamations here (references and reasons to be given in the talk).

[10] TO-DO-LIST(Speaker) is also involved in the analysis of wishes (such as *Que le meilleur gagne* [Let the best one win!] or *Que Dieu entende ma prière* [Let God hear my prayer!]): outcomes toward which Speaker is positively oriented but the realization of which does not depend on Speaker's interlocutors.

[11] SHARED GROUND is the analog of FACTS in Ginzburg's framework. This will be articulated in the talk.

calls on Addressee to commit himself to the outcome she is committed to, i.e. adopt an affirmative stance towards the actualization of the outcomes.

Let us return to the cases presented in section 2. What are ConfRs (see examples (8a), (15a) (19) above)? Precisely, utterances by which Speaker calls on Addressee to take up her utterance as a question. In the same way, whimperative constructions (see (7b) above) are constructions which specify the call on Addressee: whimperative utterances should be taken up by Addressee as directives. What is the dialogical impact of expressions such as *n'est-ce pas* or *sans indiscrétion* (see (8), (12) above)? Once again, such tags specify the call on Addressee. By adding *n'est-ce pas* ? to her utterance, Speaker marks that she expects Addressee to take it up as a question.

In order to capture this dimension we propose to add in Speaker's DGB a slot which registers the specific call on Addressee performed by Speaker.[12] Hence, the architecture of the public part of the DGB we propose is schematized in (21).

(21)
$$\left[\text{DGB}\begin{bmatrix}\text{SG}\\ \text{QUD}\\ \text{TDL}\\ \text{CALL - ON - ADDRESSEE}\end{bmatrix}\right]$$

To sum up, Speaker commits herself either to a proposition, or to a question/issue or to an outcome/order. Simultaneously, she calls on Addressee to commit himself to a proposition, to a question/issue or to an outcome/order.

This give us the key to analyze complex speech acts (such as interrogative directives or ConfRs) along with simple speech acts (such as statements, queries or commands). In complex speech acts, Speaker's commitment and Speaker's call on Addressee are distinct, whereas they are identical in simple speech acts.

4.4 Modelling Complex and Simple Speech Acts

Complex speech acts are moves in which Speaker's commitment and Speaker's call on Addressee are different. They correspond to conversational moves that associate two updates of two different slots in Speaker's DGB with distinct contents.

For example, when uttering a ConfR, Speaker updates her SHARED GROUND with a propositional content (p) and her CALL ON ADDRESSEE with a question (?p), which corresponds to the fact that she calls on Addressee to take her utterance as a question. Tags such as *n'est-ce pas* trigger the same effect. As for the tag *s'il te plaît* (see (8b)), it signals that the propositional content added in SHARED GROUND is different from the content added in CALL ON ADDRESSEE, which is an outcome.

In the absence of marking (by a construction, a tag or intonational cues), the content and the type of the call on addressee by default is identified with Speaker's commitment (be it an update of SHARED GROUND, of QUD, or of TDL). We call simple speech acts moves in which Speaker's commitment and Speaker's call on Addressee are the same. They correspond to conversational moves that associate two updates of two different slots in Speaker's DGB with identical contents.

For example, when uttering a statement, Speaker updates her SHARED GROUND with a propositional content (p) and her CALL ON ADDRESSEE with the same propositional content (p), which corresponds to the fact that she calls on Addressee to take her utterance as an assertion.

In table 1, we give the taxonomy of types of speech acts we arrive at as they are analyzed in terms of updates of Speaker's DGB.[13]

5 Conclusion

We have proposed a unified framework to analyze speech acts, be they direct or indirect. In this abstract, we have focused on speech acts in which the call on Addressee is grammatically

---

[12] CALL-ON-ADDRESSEE registers the type and content of Speaker's call on Addressee. Like LATEST MOVE - and contrarily to SG and QUD, which are structured sets - CALL-ON-ADDRESSEE contains one and only one element which is updated utterance by utterance. In the present proposal, it plays the interactive part that was carried out by QUD in Ginzburg's original proposal.

[13] Given a proposition p, we use the following convention: p? represents the polar question associated to p, and !p represents the outcome built from p, i.e. p will be true in the situation in which the outcome !p is fullfilled. For instance, if p corresponds to the sentence 'John is beautiful', then p? correspond to 'Is John beautiful?', and !p to 'Be beautiful, John!'. In this table, q' corresponds to the proposition which resolves q, and o' to the proposition which fullfills o.
CoA is used to abbreviate CALL ON ADDRESSEE.

specified and in which the core content (in fact, the SOA in Ginzburg and Sag’s terms) of the update of Speaker’s commitments is identical to the core content of the update of Speaker’s call on Addressee. The current proposal could be extended to indirect speech acts (or hints) such as those in (4)-(6) above in which the type of the call on Addressee should be inferred from private belief about the goals of the interlocutors and in which the core content of the update of Speaker’s commitments may be completely different from the core content of the update of Speaker’s call on Addressee.

Clause type	**Semantic content type**	**Conversation move types**	
		Speaker-oriented Impact	**Addressee-oriented impact**
Declarative	Proposition p	Update (S, SG, p)	*Default:* Update (S, CoA, p)
			Update (S, CoA, p?)
			Update (S, CoA, !p)
Interrogative	Propositional Abstract q	Update (S, QUD, q)	*Default:* Update (S, CoA, q)
			Update (S, CoA, q’)
			Update (S, CoA, !q')
Imperative	Outcome o	Update (S, TDL_{A},o)	*Default:* Update (S, CoA ,o)
			Update (S, CoA, o')
			Update (S, CoA, o’?)

Table 1

References

Nicholas Asher and Brian Reese. 2005. 'Negative bias in polar questions'. In E. Maier, C. Bary, and J. Huitink (eds), Proceedings of SuB9, 30–43.

Louise Fontaney. 1991. 'A la lumière de l'intonation'. In C. Kerbrat-Orecchioni (ed.), La question. Lyon : PUL, 113-161.

Gerald Gazdar. 1981. 'Speech act assignment', in Joshi, Webber and Sag (eds.), Elements of Discourse Understanding. Cambridge: Cambridge University Press. 64-83.

Jonathan Ginzburg. 1997. 'On some semantic consequences of turn taking'. In P. Dekker, M. Stokhof, and Y. Venema (eds.) Proceedings of the 11th Amsterdam Colloquium, 145-150, ILLC, Amsterdam.

Jonathan Ginzburg. To appear. A Semantics for Interaction in Dialogue, CSLI Publications and University of Chicago Press.

Jonathan Ginzburg and Ivan A. Sag. 2000. Interrogative investigations. Stanford: CSLI.

Georgia M. Green. 1975. 'How to get people to do things with words'. Syntax and Semantics 3, 107-141.

Christine Gunlogson. 2003. True to Form: Rising and Falling Declaratives as Questions in English. New York: Routledge.

Paul Portner. 2005. 'The Semantics of Imperatives within a Theory of Clause Types'. In K. Watanabe and R. B. Young (eds.), Proceedings of Salt 14. Ithaca, NY: CLC Publications.

Jerrold M. Sadock. 1974. Toward a Linguistic Theory of Speech Acts, New-York, Academic Press.

Robert C. Stalnaker. 1978. 'Assertion'. In P. Cole (ed.) Pragmatics, 315-332. Also in Portner & Partee, Formal Semantics, 2002, 147-161.

Anatol Stefanowitsch. 2003. 'The English Imperative: a Construction-based Approach', ms.

David R. Traum and James F. Allen, 1994. 'Discourse Obligations in Dialogue Processing'. Proceedings of ACL94.

Hubert Truckenbrodt. 2004. 'Sentence Type Meanings', ms available at http://www2.sfs.nphil.uni-tuebigen.de/~hubert/Home/papers/.

Dietmar Zaefferer. 2001. 'Deconstructing a classical classification: A typological look at Searle's concept of illocution type'. Revue Internationale de Philosophie 2/2001. 209-225.

Question Intonation and Lexicalized Bias Expression

Yurie Hara
Japan Society for the Promotion of Science
Kyoto University
University of Massachusetts, Amherst
`yhara@ling.bun.kyoto-u.ac.jp`

Abstract

This paper examines the interaction between different utterance types and the Japanese modal particle *darou,* and proposes that the decision-theoretic semantics accounts for the interaction among *darou,* sentence types and intonation.

1 Introduction

Many languages express question meanings morpho-syntactically and prosodically. For example, in English, Subject-Aux inversion mutates a statement into an interrogative, which is often accompanied by a rising contour as in (1–a). However, question meanings can be expressed by morpho-syntax alone as in (1–b). Moreover, a declarative sentence can be rendered into a question solely by a rising intonation as in (1–c).

(1) a. Is John coming↑ R.Interog
b. Is John coming↓ F.Interog
c. John is coming↑ R.Decl

Similarly, in Japanese, a question particle *ka* marks a sentence as interrogative, which is sometimes accompanied by a rising contour (2–a,b). As in (2–c), a question meaning can also be expressed by a declarative sentence with a rising intonation.

(2) a. John-ga kuru ka↑
John-Nom come Q
'Is John coming?' RI
b. John-ga kuru ka↓ FI
c. John-ga kuru↑ RD

Although all of these utterance types express some kind of question meanings, previous analyses agree upon that they are not completely interchangeable (see Bartels (1997) and Merin and Bartels (1997) for distinctions between rising interrogatives and falling interrogatives; Gunlogson (2003) and Nilsenova (2002) for distinctions between falling declaratives and rising declaratives).

This paper examines the interaction between these utterance types and the Japanese modal particle *darou,* which offers an interesting test case and sheds new light on the ongoing discussion of the interpretations of each utterance type. *Darou* cannot be used in an interrogative construction with a rising intonation as in (3–a). If *darou* occurs within a falling interrogative, it is interpreted as a self-addressing question (3–b). *Darou* with a rising declarative appears to function as a tag question (3–c).

(3) a. *John-ga kuru **darou** ka↑ RI
b. John-ga kuru **darou** ka↓ FI
'I wonder whether John is coming.'/'Let's think about the question whether John is coming.'
c. John-ga kuru **darou**↑ RD
'John is coming, right?'
d. John-ga kuru **darou**↓ FD
'John is coming (I bet)'

This paper is structured as follows. First, I show empirical data regarding the use of *darou* in different sentence types. In section 2.1, I go over the basic semantics of *darou* in falling declarative sentences. In section 2.2, the pitch tracks of different sentence types with *darou* are presented. Next, in section 3, I briefly review two previous studies, Merin and Bartels (1997) and Nilsenova (2002) that employ Merin's (1994) decision-theoretic approach to the description of a conversation, and analyze the meaning of intonation in English. In section 4, I extend the model formulated by the pre-

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 50–54.

vious studies and account for the pattern sketched in section 2.2.

2 Data

2.1 Japanese *Darou*

Darou is a sentence-final particle that has a modal-flavor. When *darou* is used in a plain declarative as in (4), it expresses the speaker's bias toward the content of the prejacent proposition.

(4) John-ga kuru **darou**.
Jonn-Nom come DAROU
'John is coming (I bet).'≈'Probably, John is coming.'

Sugimura (2004) observes that *darou* can co-occur with high-probability adverbs, *tabun* 'probably' and *kitto* 'certainly' but cannot co-occur with a low-probability adverb, *moshikasuruto* 'maybe'. *Darou* semantically indicates a high probability, namely a bias (more than 50 %) toward the event denoted by the proposition, hence *darou* is not compatible with a low-probability adverb.

(5) kare-wa tabun/kitto/*moshikasuruto
he-Top probably/certainly/maybe
kuru darou.
come DAROU
'Probably/Certainly/*Maybe, he will come.' (Sugimura, 2004)

The following data show that the agent of the bias expressed by *darou* needs to be the speaker.[1] (6-b) is infelicitous because the speaker's bias toward 'it will rain' does not cause John to bring an umbrella. In contrast, the modal meaning of the adverb *tabun* 'probably' does not have to be attributed to the speaker, hence (6-c) is felicitous.

(6) a. boku-wa ame-ga furu **darou**
I-Top rain-Nom fall DAROU
kara kasa-o mot-te
because umbrella-Acc have-and
it-ta
go-Past
'Because it will rain (I bet), I took an umbrella with me.'

b. ??John-wa ame-ga furu **darou**
John-Top rain-Nom fall DAROU
kara kasa-o mot-te
because umbrella-Acc have-and
it-ta
go-Past
'Because it will rain (I bet), John took an umbrella with him.'

c. John-wa ame-ga **tabun** furu
John-Top rain-Nom probably fall
kara kasa-o mot-te
because umbrella-Acc have-and
it-ta
go-Past
'Because it will probably rain, John took an umbrella with him.'

The properties of *darou* are summarized as follows:

- *Darou* indicates a bias (more than 50 %) toward the embedded proposition, i.e., p is more likely than ¬ p.
- The agent of bias is the agent of the local speech act.

2.2 Question Intonation and *Darou*

Darou exhibits interesting and subtly distinct interpretations when it is used with different sentence types and different intonations.

Let us first look at the pitch track of a rising interrogative without *darou*. The F0 Contour was measured by an autocorrelation analysis of the PRAAT program (Boersma and Weenink, 2006). A high boundary tone is found at the end of the interrogative sentence (7) as in Figure 1 (see also Venditti, 1995).

(7) Yurie-wa wain-o nomu-ka↑
Yurie-Top wine-Acc drink-Q
'Does Yurie drink wine?'

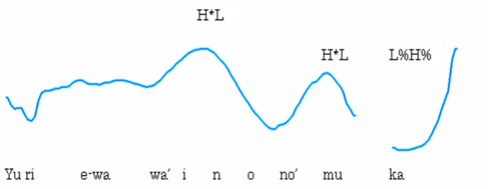


Figure 1: A rising interrogative with a high boundary tone. [speaker J.O.]

---

[1] Except for the case where *darou* is embedded under an attitude predicate. The notion of 'the speaker' can be shifted in embedded utterance contexts (see Schlenker, 2003). The discussion is omitted because it is out of the scope of this paper.

Darou cannot be used in an interrogative construction with a rising intonation. In Figure 2, the speaker is asked to pronounce (8) with a final high boundary tone, which is reported as ungrammatical by the speaker.

(8) *Yurie-wa wain-o nomu darou-ka↑
Yurie-Top wine-Acc drink DAROU-Q

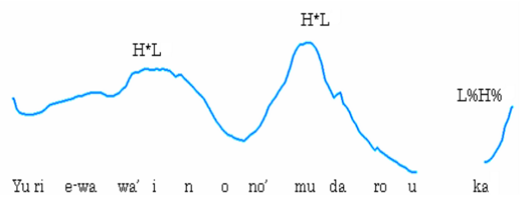


Figure 2: An interrogative construction with *darou*, which the speaker is asked to pronounce with a high boundary tone and she reports as ungrammatical. [speaker J.O.]

If *darou* occurs within a falling interrogative (Figure 3), it is interpreted as a self-addressing question.

(9) Yurie-wa wain-o nomu darou-ka↓
Yurie-Top wine-Acc drink DAROU-Q
‘I wonder if Yurie drinks wine.’

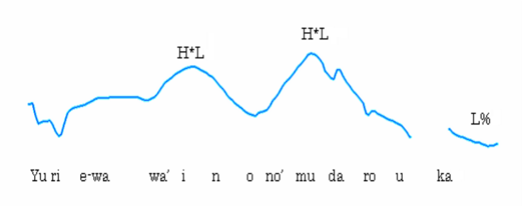


Figure 3: An interrogative construction with *darou* and a final low boundary tone. [speaker J.O.]

Darou with a rising declarative (Figure 4) appears to function as a tag question.

(10) Yurie-wa wain-o nomu darou↑
Yurie-Top wine-Acc drink DAROU
‘Yurie drinks wine, right?’

The influence of the boundary tone on interpretation is summarized in the following table.

This paper proposes that the decision-theoretic semantics accounts for the interaction among *darou*, sentence types and intonation.

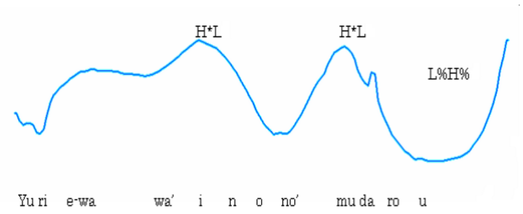


Figure 4: A declarative construction with *darou* and a final high boundary tone. [speaker J.O.]

Rising	
Interrogative	ungrammatical
Declarative	tag question (‘, right?’)
Falling	
Interrogative	self-addressing question (‘I wonder’/‘Let’s think’)
Declarative	statement (‘I have a bias’/‘I bet’)

3 Previous Studies

3.1 Merin 1994

Merin (1994) characterizes a conversation as a negotiation game among agents on what enters Common Ground. Each negotiation is executed by an Elementary Social Act, which is a transition from one negotiation state to another. Each negotiation state is instantiated by a 5-tuple, $< S, O, P, D, I >$. Values of Issue orientation [O] range over θ (a proposition) and $\neg\theta$. Values of Actor-role [S], Preference [P], Dominance [D] and Initiator-role [I] range over E (Ego, the speaker) and A (Alter, the addressee). Preference takes the value corresponding the agent who prefers adoption of the proposition θ, Dominance the social power of the agent, and Initiator-role the agent who has made a Claim pertaining to the current negotiation game. The parameter setting of each Elementary Social Act is summarized in the following table (simplified from Merin (1994)):

	S	O	P	D	I
Claim	E	θ	E	E	E
Concession	E	θ	A	A	A
Denial	E	θ	A	E	A
Retraction	E	θ	E	A	E

3.2 Merin and Bartels (1997) on Intonations

Employing Merin’s (1994) model, Merin and Bartels (1997) characterize intonation as “(re-

)allocation of [D]-parameter value":

(11) Final Rise (Merin and Bartels, 1997)
in asking, Ego is alienating choice among alternatives (sets of possible worlds) to Alter i.e. making a Concession.

(12) Final Fall (Merin and Bartels, 1997)
in requiring an answer from Alter, Ego is forcing Alter to commit himself to one mutually binding alternative (thus banning others from inclusion in the CG) and is thereby making a Claim.

Parameter settings for Rising interrogative and Falling interrogative are given in the following table:

	$< S, O, P, D, I >$
Rising Interrogative	$< E, \theta, A, A, A >$
Falling Interrogative	$< E, \theta, E, E, E >$

To illustrate, in (13), the Final Rise indicates shift of the authority of choice from Ego (the speaker) to Alter (the hearer).

(13) I didn't know John took a job all the way over in Redwood City.
Does he have a car now?
H* H-H%
(Merin and Bartels, 1997)

In Contrast, the Final Fall in (14) indicates Ego's demand to Alter for commitment:

(14) John did good work for us last year. But I doubt that we could still have him going round on his bicycle.
Does he have a car now?
H* L-L%
(Merin and Bartels, 1997)

3.3 Nilsenova (2002) on Rising Declaratives

Nilsenova (2002) examines distinctions between rising interrogatives and rising declaratives and extends Merin and Bartels's (1997) model by proposing that a rising declarative specifies Initiator-role as E (the speaker), and Dominance as A (the addressee):

	$< S, O, P, D, I >$
Rising Interrogative	$< E, \theta, A, A, A >$
Rising Declarative	$< E, \theta, A, A, E >$

To illustrate, in the following examples, Initiator-role is set to Ego, since the game of whether a proposition θ should enter the common ground is initiated by Ego. Furthermore, Dominance is set to Alter. This is either because Ego does not have necessary information to make a commitment or because Ego gives up the authority of choice to Alter in order to indicate his politeness.

(15) a. At Tim's graduation. Tim is standing next to a woman in her sixties.
Jack: You are Tim's mother?
(Nilsenova, 2002)
b. Waiter (to customer): My name is Carl? I'll be your waiter tonight?
(Gussenhoven and Chen 2000)

4 Lexical specification of *darou*

Integrating Merin and Bartels (1997), and Nilsenova (2002), I take the following table to be the classification of utterance types in terms of *default* parameter settings of Elementary Social Acts. (I assume with Merin and Bartels (1997) that Preference is either underspecified or defeasible within morphosyntactic and prosodic classification of utterance types (i.e., without *darou*).)

	$< S, O, P, D, I >$
Rising Interrogative	$< E, \theta, A, A, A >$
Rising Declarative	$< E, \theta, A, A, E >$
Falling Interrogative	$< E, \theta, E, E, E >$
Falling Declarative	$< E, \theta, E, E, E >$

Furthermore, I propose that the semantics of *darou*, i.e., the speaker's bias, lexically specifies the values of Preference and Initiator-role as Ego (the speaker).

(16) *darou*: $< E, \theta, E, (\cdot), E >$

Let us go back to the interaction between the semantics of *darou* and the typology of utterance types. As we have seen in (8) repeated here as (17), a rising interrogative is not compatible with *darou*. Now, remember that Merin and Bartels (1997) define a rising interrogative as a Concession: Ego (the speaker/the questioner) is ready to accept Alter's Claim to be in Common Ground. Therefore, the Initiator-role of a rising interrogative is Alter, which conflicts with the meaning of *darou*

(17) *Yurie-wa wain-o nomu darou-ka↑
Yurie-Top wine-Acc drink DAROU-Q

The falling interrogative (9) (repeated here as (18)) is compatible with *darou*, which results in the interpretation that Ego is demanding commitment from himself.

(18) Yurie-wa wain-o nomu darou-ka↓
Yurie-Top wine-Acc drink DAROU-Q
'I wonder if Yurie drinks wine.'

Finally, the rising declarative (10) (repeated here as (19)) is also compatible with *darou* and the combination yields the interpretation of a tag question, since *darou* indicates that Ego (the speaker) prefers the adoption of the proposition.

(19) Yurie-wa wain-o nomu darou↑
Yurie-Top wine-Acc drink DAROU
'Yurie drinks wine, right?'

The interaction between the lexical specification of *darou* and the meaning of utterance types is summarized below:

		$< S, O, P, D, I >$
RI	*darou-ka*↑ ungrammatical	N/A
RD	*darou*↑ tag question (', right?')	$< E, \theta, E, A, E >$
FI	*darou-ka*↓ self-addressing question ('I wonder'/'Let's think')	$< E, \theta, E, E, E >$
FD	*darou*↓ statement ('I have a bias'/'I bet')	$< E, \theta, E, E, E >$

5 Conclusion

To conclude, I have accounted for the influence of intonation and sentence types on interpretation of sentences with the Japanese modal particle *darou*. In particular, I integrated two previous studies on English intonation, Merin and Bartels (1997) and Nilsenova (2002), in order to provide parameter settings of four basic utterance types: rising interrogative, rising declarative, falling interrogative and falling declarative. Furthermore, I propose that *darou* lexically specifies the values of Preference and Initiator-role as the speaker. Together with the typology of social acts discussed by Merin and Bartels (1997) and Nilsenova (2002), my proposal correctly predicts the infelicity of the use of *darou* in rising interrogatives and distinct interpretations observed in rising declaratives, falling interrogatives and falling declaratives.

References

Bartels, C. (1997). *Towards a compositional interpretation of English statement and question intonation*. Ph. D. thesis, University of Massachusetts, Amherst.

Boersma, P. and D. Weenink (2006). Praat: Doint phonetics by computer. ver. 4.4.16.

Gunlogson, C. (2003). *True to Form: Rising and Falling Declaratives as Questions in English*. New York: Routledge.

Merin, A. (1994). Algebra of elementary social acts. In S. Tshotzidis (Ed.), *Foundations of Speech Act Theory*. London: Routledge.

Merin, A. and C. Bartels (1997). Decision-theoretic semantics for intonation.

Nilsenova, M. (2002). A game-theoretical approach to the meaning of intonation in rising declaratives and negative polar questions.

Schlenker, P. (2003). A plea for monsters. *Linguistics and Philosophy 26*(1), 29–120.

Sugimura, Y. (2004). Gaizensei o arawasu fukushi to bunmatsu no modality keishiki 'adverbs of probability and sentence-final modality expressions'. *Gengo Bunka Ronshuu 25*(2).

Venditti, J. J. (1995, October). *Japanese ToBI Labelling Guidelines*. Ohio State University.

Question-Answer Congruence and the Proper Representation of Focus (invited presentation)

Manfred Krifka
ZAS and Humboldt University Berlin, Germany
`krifka@rz.hu-berlin.de`
`krifka@zas.gwz-berlin.de`

One of the typical usages of focus is to mark the congruence of an answer to a question, such as (1), where the focused item *tomorrow* corresponds to the question constituent *when*.

(1) When will Fritz go to Potsdam?
Fritz will go to Potsdam TOMORROW.

In this talk I will review a number of arguments centering aroung congruent question-answer relations that have informed us about the nature and proper representation of focus. In particular, I will argue for a structured meaning approach for focus, and for a distinction between focus and focus phrases.

References

Krifka, M. 2001. For a Structured Account of Questions and Answers. In C. Féry and W. Sternefeld (eds.), *Audiatur vox sapientiae. A Festschrift for Achim von Stechow*, pages 287–319, Akademie-Verlag, Berlin.

Krifka, M. 2006. Association with Focus Phrases. In V. Molnar and S. Winkler (eds.), *The Architecture of Focus*, Mouton de Gruyter, Berlin.

Krifka, M. (to appear). The Semantics of Questions and the Focusation of Answers, in C. Lee, M. Gordon and D. Büring (eds.), *Topic and Focus: A cross-linguistic Perspective*, pages 139–151, Kluwer, Dordrecht.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, page 55.

Identifying reference to abstract objects in dialogue

Ron Artstein and Massimo Poesio
Department of Computer Science
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
`artstein|poesio [at] essex.ac.uk`

Abstract

In two experiments, many annotators marked antecedents for discourse deixis as unconstrained regions of text. The experiments show that annotators do converge on the identity of these text regions, though much of what they do can be captured by a simple model. Demonstrative pronouns are more likely than definite descriptions to be marked with discourse antecedents. We suggest that our methodology is suitable for the systematic study of discourse deixis.

1 Introduction

This paper describes two experiments that used corpus annotation to characterize discourse deixis (Webber, 1991)—an anaphoric relation in dialogue, where the reference of an anaphoric expression is present in the preceding text but not in the form of an explicit antecedent. An example of such a relation can be seen in the interpretation of the demonstrative pronoun *that* in the following snippet, taken from dialogue 2.2 of the TRAINS-91 corpus (Gross et al., 1993).[1]

(1) 7.3 : so we ship one
7.4 : boxcar
7.5 : of oranges to Elmira
7.6 : and that takes another 2 hours

The reference of *that* clearly depends on the preceding text, and in this sense the pronoun is an anaphor. The meaning of *that* in this context can perhaps be expressed with a nominalization such as *the shipping of one boxcar of oranges to Elmira*. Such a nominalization is not present in the text—but something very close to it is. This paper addresses the problem of how the appropriate antecedent can be identified through corpus annotation.

Previous work on annotating discourse-deictic relations has achieved reliability at the cost of severe restrictions on the annotation (Byron, 2002; Eckert and Strube, 2000; Navarretta, 2000). However, there is a need for empirical work to determine the degree of objectivity concerning the identification of specific references to abstract objects, even if only to conclude that such references are interpreted so subjectively that it wouldn't make sense for a system to resolve them. The experiments reported here were designed to assess the feasibility of identifying such anaphoric relations using a fairly unconstrained annotation format and a large number of linguistically naive annotators. We exchanged the highly knowledgeable opinions (and prejudices) of experts with the collective wisdom of many speakers, looking for interesting patterns that would emerge.

The references of the anaphors in question are often abstract, and do not necessarily correspond to any particular phrase or clause in the text. It is often possible to characterize an abstract referent with a textual description, as we did for the reference of the anaphor in example (1); however, we have no systematic way to compare characterizations by different annotators. In the absence of an explicit representation of all the potential referents, we chose to have our annotators point out the required antecedents by marking unconstrained regions in the text of the dialogue; this allowed comparing the annotations while retaining a high degree of precision.

2 The TRAINS dialogues

The dialogues annotated in the experiments come from the first edition of the TRAINS corpus collected at the University of Rochester in 1991 (Gross et al., 1993). This corpus consists of tran-

---

[1] The TRAINS-91 dialogue transcripts are available at `ftp://ftp.cs.rochester.edu/pub/papers/ai/92.tn1.trains_91_dialogues.txt`

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 56–63.

scripts of dialogues between two humans. One of the humans plays the 'manager' of a railway company, who needs to develop a plan to deliver specific goods at particular stations by a given deadline. The other participant in the dialogue plays a 'system', whose role is to provide the manager with required information such as journey times and equipment availability. The corpus consists of sixteen dialogues performed by eight different 'managers'—each manager has a short dialogue with a simple problem to become familiarized with the task, and a longer dialogue with a more complicated problem to solve. The 'system' in all sixteen dialogues is played by the same person.

The dialogues thus have a quite limited domain. The participants refer often to objects in the 'TRAINS world' such as engines, cars, stations, and commodities; they talk about routes, distances and times, and about different possibilities for moving the objects around. They formulate plans and identify conflicts between them. Because the goals of the dialogues are constrained, the range of abstract objects that are discussed in the dialogues is also quite limited. This is an advantage for the present study because it makes the (unconstrained) responses of the annotators fairly tractable and interpretable.

3 Annotation

The coding manual used in the experiments was based on the approach developed in the projects MATE (Poesio et al., 1999) and GNOME (Poesio, 2004). The task and instructions were simplified by eliminating the annotation of bridging references; on the other hand, we added instructions for marking multiple antecedents for ambiguous anaphoric expressions, and for marking text regions to represent abstract antecedents.

The dialogue transcripts were annotated on a computer, using the MMAX 2 annotation tool (Müller and Strube, 2003).[2] This tool uses an XML format which allows the definition of multiple levels of *markables* on top of a base text, for example phrase markables and utterance markables. We used the tool's project wizard to create the experimental texts from the plain-text transcripts and to automatically create utterance-level markables; we then manually defined the phrase-level markables, which included all the noun phrases in the text (except temporal ones). The phrase-level and utterance-level markables were the same for all the experiment participants, except for very few cases where in the course of annotation a participant inadvertently deleted or redefined a markable (this was due to a limitation of the tool, which does not afford the possibility of fixing the identity of markables while marking their attributes; the tool does make sure, however, that participants cannot alter the base text).

The participants entered their annotations using the graphical interface of MMAX 2. Their task was to determine, for all the predefined phrase-level markables in the text, whether they were anaphoric, and to identify antecedents for the anaphoric ones; antecedents were marked by creating *pointers* from an anaphoric markable to another markable representing the antecedent. If the antecedent was mentioned previously by an expression which was a phrase-level markable, a pointer was set from the anaphoric markable to the antecedent markable. If the antecedent was *not* mentioned previously by a phrase-level markable, then a text region was marked as the antecedent. The marking of text regions was done somewhat differently in the two experiments. In experiment 1, participants defined markables on a separate level, the *segment level* (hence the term "segment antecedent"); they were thus able to mark arbitrary regions of text to represent abstract antecedents (even discontinuous regions). This allowed the annotators to make very fine-grained distinctions. For example, a reasonable interpretation of the following part of dialogue 2.2 gives slightly different referents to the pronouns *that* in utterances 3.6 and 3.7: the pronoun *that* in 3.6 refers to getting the boxcar and engine to Corning, while the pronoun *that* in 3.7 refers to getting the boxcar and engine to Corning *from Elmira*.

(2) 3.1 M: so
3.2 : essentially we have to
3.3 : ... again get the boxcar
3.4 : and engine
3.5 : to Corning
3.6 : so the fastest way to do that is from Elmira
3.7 : so we'll do that

Indeed, one of our annotators captured this distinction by pointing the first pronoun to the text *get the boxcar and engine to Corning* while pointing the second pronoun to the text *get the boxcar and*

---

[2] `http://mmax.eml-research.de/`

engine to Corning from Elmira (note that the latter is a discontinuous portion of text, which also does not correspond to any syntactic constituent). However, most of the experiment participants did not make such fine-grained distinctions, and the need to define segment markables caused difficulties for some of the participants in the interaction with the software. Therefore in experiment 2 we chose a simpler design in which participants did not define new markables, but rather marked segment antecedents with multiple pointers to individual utterances: segment antecedents were not collections of words, but collections of utterances. This coarser marking of segment antecedents simplified the annotation procedure considerably.

The annotated dialogues (in XML format) were processed with custom-built perl scripts to extract the references to text regions and present them in a form suitable for analysis. Part of this processing involved propagation of these references down the coreference chains. This was needed because sometimes the same abstract object is referred to more than once in the dialogue. For example, in the following snippet from dialogue 2.2, the pronoun *that* in utterance 30.1 may refer to the same plan as the pronoun *it* in utterance 29.2.

(3) 29.1 M: mkay
29.2 : and how long would it take
30.1 S: that would take
30.2 : um
30.3 : ... six hours from .. Elmira

The annotation instructions specified that if the two markables (*it* and *that*) refer to the same object, then the first markable (*it*) should be marked as the antecedent of the second markable (*that*) regardless of whether the referent is concrete or abstract. For the purpose of this study we are interested in identifying all the references to the kind of objects represented by segment antecedents, and therefore for the purpose of analysis we propagated references to segment antecedents down the chains.

4 Experiment 1

This experiment tested the feasibility of marking text regions to represent abstract antecedents, using a large number of naive annotators; it was based on an earlier pilot which showed that inexperienced participants can be trained quickly to master enough of the MMAX 2 software to allow for reasonable annotation performance.

4.1 Experimental setup

Materials Dialogue 2.2 from the TRAINS-91 corpus; dialogue 2.1 was used for training.

Participants Twenty paid undergraduates, native speakers of English, without any previous training in corpus annotation (except one who had previously participated in a similar experiment; subsequent clustering to identify outliers failed to distinguish this participant from the others).

Procedure The participants performed the experiment together in one lab, each working on a separate computer. The experiment was run in two sessions, each consisting of two hour-long parts separated by a 30 minute break. The first part of the first session was devoted to training: participants were given the annotation manual and a map of the 'TRAINS world' and taught how to use the software, and then annotated the training text together. After the break, the participants started annotating the experimental dialogue. The second session took place five days later, and each participant continued from the point they had stopped on the previous day. Nineteen of the twenty participants completed the annotation, and continued on to annotate a newswire text as part of a separate experiment.

4.2 Results

Of the 181 phrase markables in the dialogue, 35 were annotated with a segment antecedent by three or more annotators. We chose to ignore the annotations on all markables which were given a segment antecedent by just one or two annotators, as it appears that with 20 annotators in total, such rare annotations are most likely to be errors: of the 26 markables which were identified by only one annotator, all but one appear to be in error, and of the 12 markables identified by just two annotators, at least 6 appear to be in error. The large number of singular annotations is partly due to antecedent propagation: for example, one participant linked the ten occurrences of *orange juice* and *the orange juice* in an anaphoric chain, and marked the top of the chain with a segment antecedent (annotator error); because of antecedent propagation, all ten markables appear to have a segment antecedent—but only by one annotator. A total of eight annotators contributed such singular annotations; thus, the errors do not appear to come from particular annotators who misunderstood the dialogue or instructions, but rather look like arbitrary mistakes.

Agreement By and large, annotators seemed to agree with one another on the identity of the segment antecedents they had marked. It is not clear what is the best way to measure the amount of such agreement. One simple measure is to check what percentage of annotators formed the most common choice for each markable. As an example we can look at the following bit of dialogue.

(4) 3.6 : so the fastest way to do that is from Elmira
3.7 : so we'll do that
⋮
7.3 : so we ship one
7.4 : boxcar
7.5 : of oranges to Elmira
7.6 : and that takes another 2 hours

Ten annotators marked segment antecedents for the pronoun *that* in utterance 7.6, and their chosen antecedents are shown in the following table.

Antecedent	*N*
(3.6) the … that (3.7)	1
(7.3) so … Elmira (7.5)	3
(7.3) we … Elmira (7.5)	2
(7.3) ship … Elmira (7.5)	3
(7.3) one … Elmira (7.5)	1

The most commonly chosen word for the beginning of the antecedent was either *so* or *ship*, each chosen by 3 annotators (30%); the most common choice for the end of the antecedent was *Elmira*, agreed upon by 9 annotators (90%). Averaging these percentages over the 16 most readily identifiable anaphors (those given segment antecedents by 8–12 annotators), we found that 42% of the time coders agreed with the most popular choice for the beginning of an antecedent, and 64% of the time they agreed with the most popular choice for the end. While simplistic, this measure seems appropriate for showing that agreement was higher on where the segments ended than on where they began.

One problem with the above measure is that it fails to take into account the fact that the words *so*, *we*, *ship*, and *one* in utterance 7.3 are very close, and that the antecedents that begin with these words overlap to a substantial extent. An anonymous reviewer suggested using measures from topic segmentation such as P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002); however, it is not clear to us how to adapt these measures to multiple coders, and to a situation where only small segments are selected, rather than a segmentation of the whole text. Another possibility is to use Krippendorff's α (Krippendorff, 1980; Krippendorff, 2004), a chance-corrected coefficient that allows various distance metrics between the coded categories. Alpha measures the *observed distance* D_o, which is the mean distance between all pairs of judgments that pertain to the same markable, and the *expected distance* D_e, which is the mean distance between all pairs of judgments without regard to markables; alpha is then defined as a coefficient which ranges from -1 to 1, with 1 signifying perfect agreement ($D_o = 0$), and 0 signifying chance agreement ($D_o = D_e$).

$$\alpha = 1 - \frac{D_o}{D_e}$$

Previous work has used α to calculate agreement on anaphoric chains, treating each anaphoric chain as a set of markables and using measures of set differences as distances between the chains (Passonneau, 2004; Poesio and Artstein, 2005a; Poesio and Artstein, 2005b). A similar approach treats segment antecedents as sets of words; we calculated alpha values for the 16 most readily identifiable anaphors using three distance metrics – Jaccard, Dice, and Passonneau.

	Jaccard	Dice	Passonneau
D_o	0.53	0.43	0.43
D_e	0.95	0.94	0.94
α	0.45	0.55	0.55

These measures show a fair amount of overlap between the chosen segment antecedents, though not close to perfect. It is interesting to note that the expected distance D_e is close to maximal (unity): the reason for this is that there is little overlap between the segment antecedents of different anaphors—we do not find many instances of multiple references to the same abstract object (represented by a text region). Therefore α pretty much reflects the observed agreement ($1 - D_o$), as there is little overlap expected by chance.

Treating antecedents as sets of words does not allow us to see easily where the differences between the annotators lie. We can treat beginnings and endings of words separately by using the interval version of Krippendorff's α, using individual word indices as a linear scale. For a particular markable, the observed distance is the sum of the squares of the distances between all the pairs of words chosen as antecedent beginnings or ends (this is equivalent to twice the variance about the mean, σ^2); the overall observed distance D_o is the sum of observed distances for all markables. Calculated this way for the 16 most readily identifiable anaphors, the observed distance of the beginnings of antecedents is about 2.5 times the observed distance of the ends of antecedents, confirming our previous observation that agreement on antecedent beginnings is lower than on antecedent ends. The expected distance D_e is the sum of the squares of the distances between all the pairs of words chosen as antecedents for any markable. This gives α values of 0.998 for the beginnings of antecedents and 0.999 for the ends of antecedents, which looks like very high agreement. The reason for this high value of α is an extremely high expected distance D_e, caused by the fact that the segment antecedents are spread over the entire dialogue (1421 words), whereas the segment antecedents of each particular markable tend to be in the same vicinity. The high value of α tells us that annotators are performing much better than choosing antecedent starting and ending points at random from all over the dialogue; this is to be expected, given that segment antecedents tend to be close to the anaphors (Passonneau, 1993).

Since we know that segment antecedents tend to be close to the anaphors, we can try an alternative model for chance agreement: assume that antecedents are always marked a fixed distance from their anaphors. This would associate each antecedent beginning or end with its distance from the beginning of the anaphor. The observed distance D_o remains as before, since for each anaphor all the antecedent beginnings and ends are changed by a constant. The expected distance D_e, however, is lowered considerably, since we have factored out the spreading of anaphors over the dialogue. Calculated this way, we get an α of 0.17 for the beginnings of antecedents and 0.12 for the ends of antecedents. This is extremely low: the annotators performed only 10–20% better than picking random points in relation to the anaphor! This low number is partly because interval α, like any measure of variance, takes squares of distances and is thus very sensitive to outliers. The 16 most readily identifiable anaphors comprise 155 individual annotations. In one of these annotations, the beginning and end of the antecedent lie more than 3 standard deviations away from the mean for the anaphor's antecedents; removing this single outlier brings α up to 0.21 for segment beginnings and 0.25 for segment ends. Removing six more data points where either the beginning or end of the antecedent lie 2.5–3 standard deviations away from the mean brings α up to 0.25 and 0.40, and removing an additional nine data points which lie 2–2.5 standard deviations away from the mean brings α up to 0.35 and 0.65. This shows that the extremely low value of α is the result of a small number of outliers, although even with those outliers removed agreement is far from perfect: a very primitive model of just picking an antecedent which is a fixed distance from the referring anaphor (with some random variation) accounts for much of what the annotators are doing. This could be either because the annotators or the annotation procedure are not very good, or because such a primitive model is fairly good at capturing segment antecedents.

The difference between α values for segment beginnings and ends appears to rise as we remove outliers. However, this is probably not meaningful, since this difference varies greatly depending on the cutoff point for outliers and on the minimum number of annotations a markable needs to receive in order to be considered in the comparison (we did not perform significance tests; see Krippendorff (2004) for the difficulties in calculating confidence intervals for α). The failure to show a difference in chance-corrected agreement for segment beginnings and endings means that the primitive model of a fixed distance from the anaphor is about equally good at describing the beginnings of segment antecedents and their ends; the higher agreement on segment endings is the result of lower variance around the fixed distance.

Demonstratives The annotations revealed an overwhelming preference to assign segment antecedents to demonstratives. With the exception of one instance of *that*, all the demonstrative pronouns were identified as referring to segment antecedents by at least three annotators, among them 20 instances of *that*, 4 instances of *this* and 2 instances of *those*. In contrast, only 2 of the 28 instances of the pronoun *it* were marked with a segment antecedent by three or more annotators.

These two *it* pronouns were marked by just four annotators each, and the segment interpretation of these pronouns is clearly not the only possible one. The first is the pronoun *it* in utterance 13.3: a few annotators marked it as referring to a text region containing utterance 13.1, presumably intending the action of moving the tanker; but clearly the pronoun can also refer to the tanker itself, as marked by the majority of coders.

(5) 13.1 M: so we have to move the tanker from Corning to Elmira
13.2 : ... uhm
13.3 : but we need an engine for it first

The other pronoun *it* marked with a segment antecedent by multiple coders was in utterance 29.2. It display an ambiguity which is very common in the TRAINS dialogues, between a route and a plan or action of moving trains along this route. More coders chose to mark it as coreferential with *the fastest route* than to give it a segment antecedent.

(6) 28.1 S: the fastest route is via Dansv / is
28.2 : yeah
28.3 : via Dansville
29.1 M: mkay
29.2 : and how long would it take

The observation that personal pronouns are much less likely than demonstratives to refer to abstract objects seems rather robust, in conformance with previous findings (Passonneau, 1993).

Demonstratives were also the easiest markables to identify as having segment antecedents. The eight markables which were given segment antecedents by the most annotators (between 10 and 12 annotators each) were all the pronoun *that*, occurring either as the object of *do/did* (4 instances) or as the subject of *takes/would take* (4 instances); they all referred to plans. The next eight markables, annotated with segment antecedents by 8 or 9 annotators, were also all demonstratives (six *that*, one *those* and one *that way*); they included five references to plans, one which displays the route/plan ambiguity, and two which denote activities that are not plans, for instance the activity of making orange juice.

(7) 21.1 M: um
21.2 : ‘bout how long does it take .. to make the oranges into orange juice
22.1 S: that takes an hour

Aside from demonstrative pronouns and the two instances of the pronoun *it* mentioned above, the only markables which reached the criterion of being assigned segment antecedents were definite descriptions with the head nouns *plan* or *way*. The non-demonstrative given segment antecedents by the most annotators was the NP *the plan*, identified by seven annotators. Interestingly enough, some definite descriptions whose form is highly suggestive of a segment antecedent, for example *the plan*, *the current plan* and *the banana problem*, failed to reach the criterion of identification by three annotators.

5 Experiment 2

This experiment tested whether using trained participants and a simplified coding scheme would provide improved results.

5.1 Experimental setup

Materials Dialogue 3.2 from the TRAINS-91 corpus; dialogue 3.1 was used for training.

Participants Four paid undergraduates, all of whom participated in experiment 1.

Procedure Similar to experiment 1, but slightly different marking of segment antecedents as explained above.

5.2 Results

Of the four participants, one didn’t mark even a single segment antecedent and was therefore excluded from the study. In order to have more data, we included one of the experimenters as an additional annotator (the experimenter’s annotations were produced at the same time as those of the experiment participants and without knowledge of their annotations).

In total, 35 of the 102 markables were identified with a segment antecedent by at least one annotator. Of these, 19 were identified by just one annotator; 15 of those appear to be in error—all and only those marked by one particular annotator, who apparently went for high recall at the expense of precision. The remaining four singular annotations (by three different annotators) appear to be plausible interpretations, so an acceptance criterion that requires agreement by two annotators seems too strong when there are just four annotators in total. We excluded the singular annotations of the overzealous annotator from the analysis.

As in the previous experiment, the annotators appeared to agree overall on the identity of segment antecedents, with a tendency to agree more on the ends of segments than on their beginnings. This is based on an impressionistic evaluation of the data—there are too few data points for a meaningful numerical analysis. This finding holds despite the fact that segment annotation was coarser (that is more constrained) in this experiment.

Also in line with the previous experiment, the most readily identifiable markables were demonstratives—the eight markables assigned segment antecedents by three or four coders were all instances of the pronoun *that*. The six markables which were given segment antecedents by all four annotators clearly referred to plans. The situation is less clear with regard to the two markables which were given segment antecedents by three annotators. The first of those was the word *that* in utterance 10.3, which displays a route/plan ambiguity.

(8) 10.1 S: okay the shortest route would be
10.2 : back through Dansville again
10.3 : that'll take 4 hours
10.4 : and get there
10.5 : get to Corning at 11

Indeed, the remaining annotator marked the word *that* in 10.3 as coreferential with the NP *the shortest route* in 10.1, as did one of the other annotators whose annotation received a discourse antecedent through propagation. A third annotator (the experimenter) marked the word as ambiguous between a segment antecedent and an object antecedent, intending to mark an ambiguity between a plan and a route. Only one annotator marked this unambiguously with a segment antecedent.

The second markable annotated with a segment antecedent by three participants was the word *that* in utterance 13.4. The matter here is more subtle: while the reference of *that* is related to the plan developed in the preceding utterances, it cannot actually denote the plan, but rather a fact about the identity of a plan.

(9) 13.1 M: and when our [2sec]
13.2 : engine and car .. arrives it at .. Corning
13.3 : I believe we're having it filled with oranges
13.4 : is that correct

Our method of marking antecedents as text regions is not sensitive enough to make such subtle distinctions. The three annotators who chose a segment antecedent for the pronoun *that* in 13.4 marked the preceding utterances (two chose 13.1–13.3 and one chose only 13.3); the fourth marked the pronoun as non-referring.

6 Discussion

If we impose a criterion which requires agreement by at least three of the 20 annotators in experiment 1, we find that 35 of the 181 markables in the dialogue (19.3%) have a plausible interpretation as an anaphor whose antecedent is discussed in preceding discourse but not mentioned by name. A similar figure obtains for experiment 2 after removing the annotations of the participant who appeared to have misunderstood the instructions (20 markables out of 102, or 19.6%). These percentages concur with the 22.6% figure reported by (Eckert and Strube, 2000) for their selection of dialogues from the Switchboard corpus, which is not task-oriented. The figures show that anaphora to entities not mentioned explicitly in the discourse is common enough to warrant treatment.

The fact that many of the segment antecedents in our study turned out to be plans is not surprising, and is due to the dialogues being collected as a planning task. The observation that demonstratives are more likely than other pronouns to have a segment antecedent confirms earlier findings. What is new is the finding that demonstratives are more readily identifiable as elements which require such antecedents—more so than definite descriptions with a highly suggestive head noun. This has implications for writing annotation guidelines, and possibly also for resolution.

The set-based measures show that there is substantial overlap between the annotators regarding the identity of segment antecedents; while far from perfect, this suggests that the task itself is a feasible one, and hopefully can be improved. As for the word-index measures, the fact that a simple model of picking antecedents at a fixed distance from the anaphor accounts for much of what the annotators are doing is in some ways encouraging, as it suggests that the correct vicinity (if not the exact antecedent) could, perhaps, be identified computationally. At the same time, this finding puts in question the added value of human annotation, since annotators have not shown much improvement over the base model. An anonymous reviewer points out that there may be a limit to what we can expect from the annotators because they are in a sense overhearers of the dialogue rather than participants in it, and therefore do not play a part in the grounding process that takes place between the participants (Schober and Clark, 1989). The attainable agreement among annotators may therefore be lower than a reflection of the understanding of dialogue participants.

The same reviewer also suggested out a possible explanation to the fact that annotators agree more on the endings of segment antecedents than on their beginnings, namely that candidate antecedents occur on the right frontier of the discourse structure (Webber, 1991), so their ends tend to coincide. However, the fact that we did not find a difference in chance-corrected agreement between beginnings and ends of antecedents suggests an alternative explanation—perhaps agreement is higher on the ends simply because the space for endings is more compressed. Of course, it could be that both explanations are right and the latter is the result of the former; we would need more experimentation to distinguish between these hypotheses.

It is encouraging that many annotators with little training can converge on roughly similar text regions as antecedents, as it shows that the judgments are not too subjective. Hopefully this should lead to a more systematic study of discourse deixis and discourse antecedents.

Acknowledgments

We wish to thank two anonymous reviewers and the participants of the Essex language and computation reading seminar. This work was in part supported by EPSRC project GR/S76434/01, ARRAU.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of 40th Annual Meeting of the ACL*, pages 80–87, Philadelphia, July.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Derek Gross, James F. Allen, and David R. Traum. 1993. The Trains 91 dialogues. TRAINS Technical Note 92-1, University of Rochester Computer Science Department, July.

Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, CA.

Klaus Krippendorff, 2004. *Content Analysis: An Introduction to Its Methodology*, chapter 11, pages 211–256. Sage, Thousand Oaks, CA, second edition.

Christoph Müller and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo, July.

Costanza Navarretta. 2000. Abstract anaphora resolution in Danish. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue*, Hong Kong, October.

Rebecca J. Passonneau. 1993. Getting and keeping the center of attention. In Madeleine Bates and Ralph M. Weischedel, editors, *Challenges in Natural Language Processing*, pages 179–227. Cambridge University Press, Cambridge.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*, volume 4, pages 1503–1506, Lisbon.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Massimo Poesio and Ron Artstein. 2005a. Annotating (anaphoric) ambiguity. In *Proceedings from the Corpus Linguistics Conference Series*, volume 1, Birmingham, England, July.

Massimo Poesio and Ron Artstein. 2005b. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In Marilyn Walker, editor, *Proceedings of the ACL Workshop Towards Standards and Tools for Discourse Tagging*, pages 65–74, College Park, Maryland, June. Association for Computational Linguistics.

Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, Massachusetts, April-May. Association for Computational Linguistics.

Michael F. Schober and Herbert H. Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Formal Semantics for Iconic Gesture

Alex Lascarides[1]
[1]Human Communication Research Centre
University of Edinburgh
Edinburgh EH8 9LW, UK
`alex@inf.ed.ac.uk`

Matthew Stone[1,2]
[2]Department of Computer Science
Rutgers University
Picataway, NJ 08845-8020
`Matthew.Stone@rutgers.edu`

Abstract

We present a formal analysis of iconic coverbal gesture. Our model describes the incomplete meaning of gesture that's derivable from its form, and the pragmatic reasoning that yields a more specific interpretation. Our formalism builds on established models of discourse interpretation to capture key insights from the descriptive literature on gesture: synchronous speech and gesture express a single thought, but while the form of iconic gesture is an important clue to its interpretation, the content of gesture can be resolved only by linking it to its context.

1 Introduction

Speakers use their whole bodies to present their ideas. Utterance (1), drawn from a lecture about speech errors,[1] shows how speakers can combine speech and gesture to flesh out their arguments in visible form.

(1) There are these very low level phonological errors that tend not to get reported.
The right hand is held in a fist and positioned below the mouth, where the previous gesture was performed; the hand iteratively moves in the sagittal plane (i.e., vertically outwards) in clockwise circles (as viewed from left).

In context, the gesture seems to visualise the continuous processes, operating below the level of awareness, that give rise to unreported errors.

Descriptive work on such gestures makes three key observations, which any theoretical account must respect. First, speech and gesture combine to express a single thought. Their contents fit together, forming the speaker's overall message (McNeill, 1992; Kendon, 2004). For example, in (1) the gesture visualises the subconscious nature of processes that cause low-level phonological errors, thereby explaining why they don't get reported.

Second, these gestures take a form that directly or metaphorically depicts what is described (McNeill, 1992; Kopp et al., 2004). For example, the iterative movement in (1) is a metaphorical depiction of a continuous process. However, not all aspects of a gesture have to be meaningful; e.g., the clockwise direction of motion in (1) doesn't contribute to interpretation.

Third, apart from conventionalised gestures (e.g., "thumbs up"), the form of a gesture on its own is insufficient for a coherent interpretation. For example, the gesture in (1) would be uninterpretable without simultaneous speech. A specific and coherent interpretation of gesture arises by linking it to simultaneous speech, and so it changes meaning in different speech contexts:

(2) The mouse ran on the wheel for a few minutes.
Gesture as in (1)

In (2), the gesture is still iconic: the physical movement of the hand depicts the path of the wheel's motion. But its interpretation is different from that in (1), and in particular the direction of motion now carries important information whereas it didn't in (1). A further kind of context-dependence arises through spatial distinctions maintained across multiple gestures (Emmorey et al., 2000). In (1), we recognise that the processes depicted are low-level in part by linking

---

[1] http://www.talkbank.org/media/Class/Lecture-unlinked/feb02/feb02-8.mov

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 64–71.

the gesture here to earlier gestures which have depicted the production of *noteworthy* errors through a trajectory leading from the mouth *upward*.

This paper describes a formal analysis of gesture that respects these three principles. In formalising these principles, we go beyond previous work—whether descriptive (McNeill, 1992; Kendon, 2004), psychological (Lozano and Tversky, 2004), or applied to embodied agents (Cassell, 2001; Kopp et al., 2004)—by drawing on formal models of semantics and pragmatics in discourse interpretation. Specifically, we argue in Section 2 that *rhetorical relations* provide a theoretical construct to explicate how speech and gesture cohere into a single thought. We explain in Section 3 how *underspecified representations of meaning* let us specify both how the form of gesture constrains its content and how the resulting representation needs to be augmented by contextual information to obtain a coherent logical form (LF). In Section 4 we represent LFs with *dynamic semantics* to capture the evolving structure of objects and spatial relationships that inform gesture interpretation. And in Section 5, this formal apparatus allows us to model how gesture is interpreted by drawing on its mappings from form to (underspecified) meaning, a context of salient objects and relationships, and rhetorical connections to synchronous speech.

While the resulting architecture captures descriptive insights into gesture, it in fact instantiates a general end-to-end model of pragmatic interpretation (Asher and Lascarides, 2003). We believe that these same principles apply to the interpretation of all communication—in whatever medium it takes place.

2 Relating gesture to speech

For Asher and Lascarides (2003), rhetorical relations are kinds of speech acts. That is, they offer an inventory of things that a speaker might be doing by providing content in discourse: he might be elaborating it, explaining it, continuing a narration, drawing a contrast, and so forth. When hearers infer rhetorical relations, they recognise the speaker's communicative intention and so discover why the discourse is coherent.

We propose that gesture is rhetorically related to simultaneous speech. For example, the gesture in (1) can be understood as providing an *explanation* in support of what is being said. The gesture in (2) can be understood as an *elaboration* that complements what is being said. On our view, the rhetorical connection is a tool which lets us formalise the intuition that the gesture is a communicative action which plays a part in the speaker's overall intention: rhetorical connections knit gesture and speech into a single thought.

Rhetorical relations are a vehicle for predicting implicatures, because their semantic consequences go beyond the compositional semantics of the utterances (and gestures) they connect, and inferring rhetorical relations during discourse interpretation involves commonsense reasoning with compositional semantics and contextual information such as world knowledge. Rhetorical relations also create a hierarchical structure to the discourse, where some communicative actions are completed and others remain open. This structure thus constrains the alternative ways coherent discourse can progress. The theory of rhetorical relations therefore serves to operationalise Grice's (1975) theory of communication as rational behaviour, articulating a precise interface between compositional semantics and pragmatics.

In essence, inferring rhetorical connections and inferring a gesture's specific meaning are logically co-dependent tasks. For example, interpreting the gesture in (1) as a continuous subconscious process causing speech errors supports an inference that the gesture and speech are related with *explanation*. This inference is justified partly by the semantics of *explanation* and partly by world knowledge: errors won't get reported if they aren't perceived; and the effects of continuous subconscious processes are normally hard to perceive.

Note that this specific content is compatible with the gesture's underspecified meaning as revealed by its form: as we shall see in Section 3, the fist can be interpreted as depicting the phonological errors being caused by something; the iterative, continuous motion of the hand can be interpreted as conveying that this cause is iterative and continuous; and the relatively low position of the hand can be interpreted as conveying that it is 'low down' or subconscious. However, the *explanation* relation predicts that the clockwise motion does not depict anything in this context.

There may be alternative specific interpretations of the gesture in (1), which in turn support inferences to alternative rhetorical connections, but as Asher and Lascarides (2003) argue, discourse interpretation is governed by a general principle of maximising coherence: one interprets discourse so that the highest possible quality of rhetorical connections is achieved (see Section 5 for further details). Of course, calculating a preferred interpretation using this principle does require formalising all the commonsense background involved.

Rhetorical relations thus help to model how context yields a more specific interpretation of the gesture from its underspecified meaning as revealed by its form. The remainder of this paper puts the case in formal terms. Now, generalising from Asher and Lascarides (2003), we would also expect that rhetorical relations can help to characterise the interpretation of speech and gesture in other ways—such as predicting when the interpretation of a gesture is coherent and when it is not in a way that other pragmatic knowledge sources, such as world knowledge, cannot do on their own; or modelling how a gesture can resolve ambiguities in synchronous speech. We leave these suggestions to future work.

3 Underspecifying iconic meaning

Underspecification is a common representational approach to interface an abstract linguistic meaning to its specific, contextualised interpretation e.g., (Alshawi and Crouch, 1992; Reyle, 1993). The contextualised interpretation is represented as a logical formula in a standard formal language; this plays the role of an LF in the model. (We will combine rhetorical relations with dynamic semantics to represent LFs; see Section 4.) The grammar, however, does not explicitly construct the LF. Instead, it builds a *partial description* of it, leaving open multiple alternatives. In this sense, the exact interpretation is left underspecified by compositional semantics. Accordingly, the underspecified elements in the description must be *resolved* pragmatically in interpretation.

We adopt *Robust minimal recursion semantics* (RMRS) as a formalism for underspecified semantic representation (Copestake, 2003). Like many formalisms, RMRS can underspecify semantic scope. In addition, it can represent partial information about *which predicates* appear in LF, *what arity* they have, and *what sorts of arguments* they take, a flexibility that isn't fully supported by other formalisms (e.g., Asher and Lascarides (2003) do not underspecify arity). We show that the form of iconic gesture constrains, but does not determine, all these aspects of interpretation.

Following earlier work (McNeill, 1992; Kopp et al., 2004), we characterise the link between the form and iconic meaning of gesture by representing gesture form in a multidimensional matrix. The rows in this matrix describe aspects of a gesture's form which potentially reveal things about its meaning—the hand shape, the orientations of the palm and finger, the position of the hands relative to the speaker's torso, the paths of the hands and the direction in which the hands move along those paths. For example, we represent the gesture form of (1) as the feature structure in (3).

$$(3)\quad \begin{bmatrix} \text{hand-shape} : \textit{asl-s} \\ \text{finger-direction} : \textit{down} \\ \text{palm-direction} : \textit{left} \\ \text{trajectory} : \textit{sagittal-circle} \\ \text{movement-direction} : \{\textit{iterative},\textit{clockwise}\} \\ \text{location} : \textit{central-right} \end{bmatrix}$$

Here each of the six attributes takes a particular value which characterises the physical realisation of the gesture. The matrix formalism highlights that the gesture morphology does not yield a hierarchical structure; rather, elements of the description combine via unification or 'conjunction'.

The gesture's iconicity consists in the fact that each of these attribute-value elements may convey a specific, analogous piece of content. With RMRS, we can formalise this in two straightforward steps. First, to each attribute-value element, we associate an *underspecified abstract predication* that must be resolved to a particular formula in the logical form of gesture. We introduce a convention that reads this underspecified predication directly off the feature structure, as in (4):

(4) $h_1{:}\mathit{hand\_shape\_asl\text{-}s}(i_1)$

Here h_1 is a uniquely indexed label that underspecifies the scope of the predication; i_1 is a uniquely indexed *metavariable* that underspecifies the main argument of the predication (an object, eventuality, etc); and *hand_shape_asl-s* underspecifies the property of i_1 that's depicted through the gesture's fist-shape. The compositional meaning of a gesture is just the conjunction of the underspecified predications associated with each of its form features. These predications must be resolved to give the gesture a specific interpretation.

Second, we constrain the possible resolutions of the underspecified predicates to a restricted inventory that states what alternative qualities we can depict with aspects of the gesture's form.

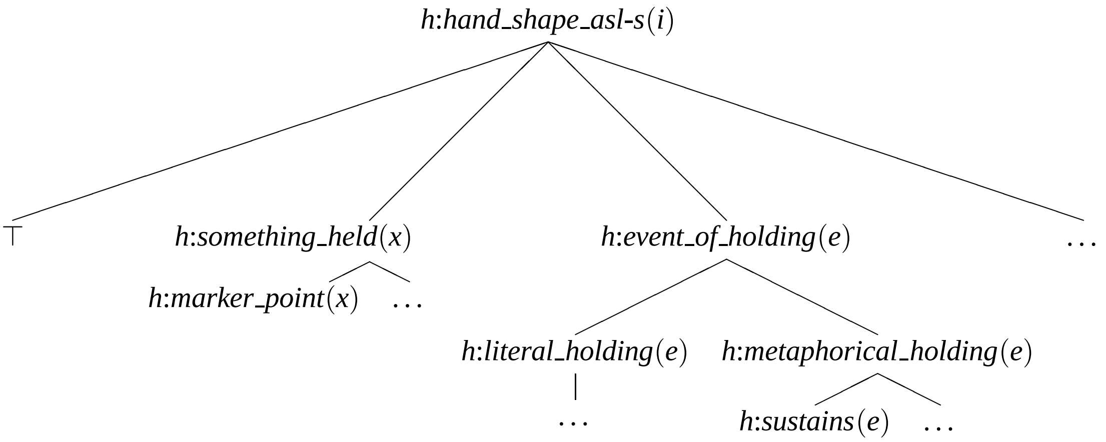


Figure 1: Part of the hierarchy of underspecified and fully-specified predications for *hand_shape_asl-s*.

We expect that each underspecified predicate admits a hierarchy of increasingly specified resolutions, as in Figure 1. While some of the leaves in this hierarchy correspond to fully specific interpretations, the creative use of metaphor makes interpretation open-ended. Therefore, some of the hierarchy's leaves correspond to more vague interpretations, and we envisage that either the speaker and hearer sometimes settle on a coherent but vague interpretation, or additional logical axioms will resolve a vague interpretation to a more specific one in the particular discourse context. To capture the (metaphorical) contribution of the fist in (1), we resolve *hand_shape_asl-s* as depicting a holding event, metaphorically interpreted as the event e of a process x *sustaining* speech errors y ("bearing them with it", as it were). At the same time, we can capture the contribution of the fist to the depiction of (2) by resolving *hand_shape_asl-s* as depicting something held, in particular a *marker-point* x indicating a designated location on the mouse's spinning wheel. Finally, all underspecified predications are resolvable to $\top$—the valid formula—since they might not contribute meaning in context. Underspecified predicates may also share specific resolutions: e.g., *marker-point* is also one way of resolving the underspecified predicate corresponding to a flat hand, and thus the gesture in (2) could have been performed with a flat hand instead of a fist.

Crucially, Figure 1 reflects the fact that, like all dimensions of iconic gesture, the fist shape doesn't determine how many entities are involved in the specific semantic relation it resolves to. The specific predications in Figure 1 vary in the number of arguments they take, and the factorised notation of RMRS lets us express this. In RMRS, additional arguments to predicates, over and above the 'primary' one, are expressed as separate binary relations: e.g., *sustains* is a 3-place predicate and h:*sustains*(e,x,y) is a notational variant of h:*sustains*(e), *ARG1*(h,x), *ARG2*(h,y), while *marker-point* is a 1-place predicate, and therefore h:*marker-point*(x), *ARG1*(h,y) is unsatisfiable. One can also underspecify the position of a variable in a predication: the binary relation *ARGn*(h,x) means that x is an argument to the predicate labelled by h, but its argument position is unknown (so *ARG1* $\sqsubset$ *ARGn*).

The divergent resolutions of the same gesture in different contexts highlight how we capture insights from previous work: we represent gesture meaning compositionally and iconically, yet in an underspecified form that requires context to resolve. You can compare predications like *hand_shape_asl-s* to Kopp et al.'s (2004) *image description features*, an abstract representation, distinct from form and content, that captures gesture meaning. By using RMRS, we can reinterpret these representations as analogous, both formally and substantively, to existing underspecified semantic representations for linguistic items. In particular, we show in Section 5 that we can therefore build reasoning mechanisms that combine information from speech and gesture to derive a single, overall coherent resolution of the logical form of discourse.

4 Representing meaning in context

In portraying objects and relationships, gesture exploits not just the iconic meaning of physical actions, but also the evolving discourse context.

For example, gesture, like speech, has access to the salient objects that have been evoked by noun phrases in the previous discourse. However, one striking difference between gesture and speech is that gesture is profoundly limited in its ability to introduce new entities into the context. We adapt the formalism of *segmented discourse representation structures* (SDRS) (Asher and Lascarides, 2003) to precisely model these similarities and differences between gesture and speech. An SDRS specifies a collection of update expressions which partially describe the evolution of context during the discourse. The SDRS also links these updates together using rhetorical relations to further constrain the interpretation and structure of discourse. We focus here on the updates themselves.

Individuals which are introduced in gesture seem to be subject to similar constraints on acceptability as *definite descriptions* in language: in both cases, the entities so-introduced must be related to an available antecedent through one of a constrained set of semantic relationships—including equality, in which case the entity is coreferent with its antecedent. We call these *bridging* relations, after Clark (1977). For instance, we infer in (2) that the *marker-point* represented by the fist indicates a *part of* the wheel. Thus there is a bridging relationship *part-of* between the gestural depiction and the noun phrase *the wheel* in the utterance.

The form of the gesture doesn't determine the bridging relation nor the antecedent, just as the form of definite descriptions doesn't. And so the form of gesture (and of definite descriptions) must impose the constraint that there is such a bridging relation, but underspecify its value. We follow Asher and Lascarides' (2003) realisation of this. For the sake of simplicity, we simply mention the notation and gloss its interpretation in words: $R(x,y) \wedge R = ? \wedge x = ?$ means that y is related to an (available) individual x with a relation R, but the *values* of x and R are underspecified. Following Chierchia's (1995) compositional semantics of definite descriptions, we include bridging constraints in the LF of gesture. These can be added to the RMRS produced by the grammar. We assume this addition occurs *outside* the grammar because bridging relations don't affect semantic composition from syntax. Rather, they impose constraints on the process of constructing the LF of discourse, stipulating that a particular relation to a particular available antecedent must be found for each individual variable.

At the same time, use of a gesture changes the referents available to subsequent discourse. This is the bread-and-butter of dynamic semantics—see e.g. Groenendijk and Stokhof (1991)—and we handle it in the usual way. We interpret formulae as transitions that update an input context to yield an output context. Among other things, these changing contexts make explicit what referents are available. However, an object introduced in gesture, like the point on the wheel in (2), can appear in subsequent gestural figurations, but cannot be picked up by a pronoun in subsequent speech. So we follow Asher and McCready (2006) in structuring our contexts to distinguish kinds of reference: we have one set of referents f available to speech and another set g (a superset in fact) available to gesture—see the dynamic semantic definition of indefinite quantification in (5a). Correspondingly, we annotate LFs for speech and gesture to indicate which kind of reference they participate in. That is, we introduce a 'gesture' modality $[\mathcal{G}]$, and the dynamic semantics of $[\mathcal{G}]\phi$ ensures that ϕ updates only the set g of referents available to *gesture*; see (5b):

(5) a. $\langle f,g\rangle [\![\exists x]\!]^M \langle f',g'\rangle$ iff
$dom(f') = dom(f) \cup \{x\}$ and
$\forall y \in dom(f),\ f'(y) = f(y)$
(i.e., $f \subseteq_x f'$), $g \subseteq_x g'$, and $f'(x) = g'(x)$.

b. $\langle f,g\rangle [\![[\mathcal{G}]\phi]\!]^M \langle f',g'\rangle$ iff $f = f'$ and $\langle g,g\rangle [\![\phi]\!]^M \langle g',g'\rangle$

One of the most interesting kinds of context dependence is the way successive gestures can establish a common frame of reference for spatial depiction (Emmorey et al., 2000; Kopp et al., 2004). We believe that dynamic semantics will provide an attractive formal setting in which to capture such connections precisely, since dynamic semantics has already proved an effective tool for modelling the evolving perspective in discourse—in time, space and information (Bittner, 2006). However, a model of spatial context in gesture will need substantial formal development, requiring a suitable formal ontology of space, a corresponding characterisation of spatial context, and rules for interpreting gesture meaning in terms of this spatial context. We leave this for the future, and here limit ourselves to the formalism sketched so far, which we can more immediately carry over from Asher and Lascarides (2003) and which in fact suffices to account for examples (1) and (2).

5 Interpreting gesture

We now address the problem of how the underspecified semantics revealed by form gets resolved to fully specific meanings in context. In Asher and Lascarides' (2003) SDRT model, this occurs as a byproduct of *discourse update*: the process of constructing the logical form of discourse.

Discourse update in SDRT starts from the compositional semantics derived from the grammar. To handle situated language, we work with the semantics for gesture derived from its form by iconicity. The compositional semantics of the gesture in (1) and (2) is shown in (6).

(6) $h_g{:}[\mathcal{G}](h)$,
$h \geq h_j$, for $1 \leq j \leq 6$,
$h_1{:}\mathit{hand\_shape\_asl\text{-}s}(i_1)$,
$h_2{:}\mathit{finger\_dir\_down}(i_2)$,
$h_3{:}\mathit{palm\_dir\_left}(i_3)$,
$h_4{:}\mathit{traj\_sagittal\_circle}(i_4)$,
$h_5{:}\mathit{move\_dir\_iterative}(i_5)$,
$h_5{:}\mathit{move\_dir\_clockwise}(i_5)$,
$h_6{:}\mathit{loc\_central\text{-}right}(i_6)$

In outline, this formula says that the final meaning will contain an expression h_g giving information specified through gesture, and that this information will resolve how the hand shape, finger direction, path, trajectory, direction of motion and location of the gesture (as labelled by $h_1 \ldots h_6$) work to describe salient generalised individuals (as labelled by $i_1 \ldots i_6$) from the context. Observe that the modality $[\mathcal{G}]$ outscopes the predications labelled h_1 to h_6, as required by the dynamic semantics in (5) of any of its resolved forms.

We assume, following Kopp et al. (2004), that gesture combines with its synchronous speech *within the grammar*, producing a single derivation tree. This assumption is necessary both to predict the fine-grained temporal synchrony between speech and gesture, and to capture the distinctive constraints on coreference and other semantic relations that apply to units of speech and gesture in coordination (e.g., a gesture and its synchronous speech cannot be combined with disjunction). Here the grammar yields the predication $h{:}\mathit{iconic\_rel}(h_s, h_g)$, where h_s labels the content of the speech. This predication underspecifies the rhetorical connection between the gesture and speech and must resolve to a value that's licensed by iconic gesture: e.g., *Explanation* or *Elaboration*, but not *Contrast* or *Disjunction*.

Discourse update derives an LF through commonsense reasoning, drawing on non-linguistic information, such as world knowledge, as well as compositional semantics. This reasoning is formalised using nonmonotonic inference rules that predict possible rhetorical connections from (shallow) representations of linguistic meaning and non-linguistic information. We refer collectively to this system as the *glue logic*. Its rules have the following form, where $A > B$ can be read as *If A then normally B*, and the symbols α and β are metavariables ranging over the labels of discourse segments in the SDRS representation:

$(\lambda{:}?(\alpha,\beta) \wedge \varphi) > \lambda{:}R(\alpha,\beta)$
(`Glue Logic Schema`)

In words: if the segment labelled β is to be connected to the segment labelled α with a rhetorical relation, and the result is to appear as part of the logical scope labelled λ, but we don't know the value of this relation yet, and moreover φ holds of the content labelled by λ, α and β, then normally the rhetorical relation is R. The conjunct φ is cashed out in terms of the (underspecified) LFs that α and β label, and the rules are justified either on the basis of underlying linguistic knowledge, world knowledge, or knowledge of the cognitive states of the conversational participants. Thus glue logic axioms encapsulate *prima facie* default inferences about which type of speech act was performed, on the basis of the content and context of the utterances.

In SDRT the inferences can flow in one of several directions. For example, if the premises of a glue logic axiom is satisfied by the information already available (e.g., by the underspecified semantics derived from the grammar), then one can infer a particular rhetorical relation and from its semantics infer how the underspecified conditions of the utterance or gesture are resolved. Alternatively, there are cases where the premises for inferring rhetorical relations are not satisfied by the underspecified compositional semantics. In this case, one can resolve the underspecified content so as to support an inference to a rhetorical relation. If one adopts this strategy, and moreover there is a choice of which way to resolve the underspecified content so as to infer a rhetorical relation from it, then one chooses an interpretation which maximises the *quality* of the rhetorical relations one can infer from it (see Asher and Lascarides (2003) for details).

Here, we indicate how discourse update can resolve the underspecified meaning of gesture with speech. Let's start with the analysis of the situated utterance (2). We introduce a glue logic axiom which captures the following intuition: if two propositions are rhetorically related somehow, and they both describe a movement event with the same participant and which can occur simultaneously, then there is evidence in the discourse that these events are in a subtype relation (following Asher and Lascarides (2003), we assume a notation where e_α and e_β are respectively the semantic indices of α and β):

(7) $(\lambda{:}?(\alpha,\beta) \wedge h{:}movement(e_\alpha) \wedge ARGn(h,x) \wedge$
$h'{:}movement(e_\beta) \wedge ARGn(h',x) \wedge$
$temporally\text{-}compatible(e_\alpha,e_\beta)) \rightarrow$
$Subtype_D(\beta,\alpha)$

The predication $Subtype_D(\beta,\alpha)$ does not entail that β and α are *actually* in a subtype relation; only that there is evidence in the discourse that they are. Note that the rule is monotonic, because either the evidence is present in the discourse, or it's not. This predicate is used to infer *Elaboration*:

(8) $(\lambda{:}?(\alpha,\beta) \wedge Subtype_D(\beta,\alpha)) >$
$\lambda{:}Elaboration(\alpha,\beta)$

If $Elaboration(\alpha,\beta)$ is inferred, then an *actual* subtype relation among their events follows.

Now returning to the situated utterance (2), the grammar imposes a constraint that the contents of speech and gesture are rhetorically connected by one of the relations that's licensed for gesture (as encapsulated in *iconic_rel*). So for (2) to be coherent, one must infer a particular rhetorical relation between them and also infer specific interpretations that support this relation.

In (2), the underspecified content on its own is insufficient for inferring a rhetorical relation, for although the gesture depicts movement, some of its possible specific interpretations do not entail physical movement (e.g., the movement could be metaphorical, or indeed the movement could resolve to $\top$ as explained in Section 3). Nor does the gesture's form specify the movement's participants. However, one of the possible resolved meanings of the gesture is one which satisfies the axiom (7). This is because one can resolve e_β (i.e., the semantic index of the gesture) to be the movement of the wheel in a circular, iterative clockwise direction, where the wheel is also the location of the running described in the sentence. This possible interpretation of the gesture is supported by world knowledge, which stipulates that when a marker point on a rigid object moves then so does that object. Moreover, world knowledge suggests that the moving object that's depicted cannot be the mouse, since the mouse runs on the spot. Thus with this specific interpretation of the gesture, the antecedent to (7) is satisfied by the content of the utterance and the gesture, with x in this axiom instantiated by the wheel. If the gesture is interpreted this way, then the axioms (7) and (8) lead to a (nonmonotonic) inference that the utterance and gesture are related with *Elaboration*. Suppose that this is the *only* possible resolved interpretation of the gesture that leads to an inference about which rhetorical relation connects the utterance and the gesture. Then discourse update in SDRT forces this specific interpretation (see (Asher and Lascarides, 2003) for formal details). Thus discourse update resolves the hand shape to $marker\text{-}point(y)$ and the accompanying bridging relation $part\text{-}of(y,x) \wedge wheel(x)$, where x is co-referent with the wheel denoted in (2); it resolves the underspecified predicate $traj\_sagittal\_circle(i)$ to $move(e_\beta,x) \wedge path(e_\beta,z) \wedge sagittal\_circle(z)$, and it resolves the underspecified predications $move\text{-}dir\_iterative(j)$ and $move\text{-}dir\_clockwise(j)$ to $direction(e_\beta,w) \wedge iterative(w) \wedge clockwise(w)$. Thus the gesture provides more information about the movement described in the utterance: the wheel is in a vertical plane (and fixed at a central point), and moves in a clockwise direction several times.

The analysis of (1) is similar to that of (2).

(1) There are these very low level phonological errors that tend not to get reported.

However, the specific interpretation of the gesture in (1) cannot satisfy the axiom (7) this time, because the sentence is not about physical movement. So another specific interpretation is needed to support a particular rhetorical connection between the speech and gesture. As we explained in Sections 2 and 3, the underspecified content of the gesture can resolve to denote a continuous, subconscious process which causes the phonological errors mentioned in (1). This particular interpretation satisfies the antecedent of an axiom whose consequent is $Cause_D(\beta,\alpha)$—i.e., there is evidence in the discourse of a causal relation. This in turn supports a default inference to *Explanation*:

(9) $(\lambda{:}?(\alpha,\beta) \wedge \mathit{Cause}_D(\beta,\alpha)) >$
$\lambda{:}\mathit{Explanation}(\alpha,\beta)$

If this is the specific interpretation which maximises the quality of the connection between the constituents, then discourse update dictates that the logical form of the discourse resolves the interpretations this way.

6 Conclusion and future work

We have provided a formal semantic analysis of iconic gesture which captures several compelling features that are described in the literature. First, it predicts that iconic gesture on its own doesn't receive a coherent interpretation: this is achieved by assigning a very underspecified content to iconic gesture as revealed by its form. Second, it predicts that speech and gesture together form a 'single thought'. This is achieved by integrating the content of gesture and synchronous speech in the grammar, and ensuring that their denotations are semantically related. The model then demands that one must compute the value of this rhetorical relation, using compositional semantics and contextual information as clues. Reasoning about this rhetorical connection leads to the gesture's underspecified content being resolved to a specific interpretation. Finally, we exploited discourse structure and the dynamics in dynamic semantics to account for dependencies on co-reference across speech and gesture and among different gestures in the discourse.

One virtue in our analysis is to demonstrate that existing mechanisms for representing the content of language can be exploited to model gesture as well. However, much future work needs to be done. For example, we need to specify in more detail the construction rules in the grammar which combine speech and gesture, and the meaning postulates which convey the range of possible meanings that the various dimensions of iconic gesture can depict. Concretely, that requires us to specify a hierarchy as in Figure 1 more fully, and to link the hierarchy to a family of interpretive instances of the Glue Logic Schema so as to predict a wide range of natural interpretations. In the dynamic semantic component, we need to integrate the interpretation of gesture with a commonsense view of space. We would also like to explore in more detail how a gesture's interpretation is constrained by prior gestures, as well as speech, and extend the analysis to other types of gesture, such as deixis and beats.

Acknowledgments

Thanks to Paul Tepper and anonymous reviewers for comments. Supported in part by the Leverhulme Trust, Rutgers University and NSF HLC0308121.

References

H. Alshawi and R. Crouch. 1992. Monotonic semantic interpretation. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 32–39, Delaware.

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

N. Asher and E. McCready. 2006. Were, would, might and a compositional account of counterfactuals. technical report; material will be presented at ESSLLI 2006.

M. Bittner. 2006. Online update: temporal, modal and *de se* anaphora in polysynthetic discourse. In C. Barker and P. Jacobson, editors, *Direct Compositionality*. Oxford.

J. Cassell. 2001. Embodied conversational agents: Representation and intelligence in user interface. *AI Magazine*, 22(3):67–83.

G. Chierchia. 1995. *Dynamics of Meaning: Anaphora, Presupposition and the Theory of Grammar*. University of Chicago Press, Chicago.

H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press.

A. Copestake. 2003. Report on the design of rmrs. Technical Report EU Deliverable for Project number IST-2001-37836, WP1a, Computer Laboratory, University of Cambridge.

K. B. Emmorey, B. Tversky, and H. Taylor. 2000. Using space to describe space: Perspective in speech, sign and gesture. *Spatial Cognition and Computation*, 2(3):157–180.

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

J. Groenendijk and M. Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100.

A. Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge.

S. Kopp, P. Tepper, and J. Cassell. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of ICMI*.

S. C. Lozano and B. Tversky. 2004. Communicative gestures benefit communicators. In *Proceedings of Cognitive Science*.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago.

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

SDRT and Multi-modal Situated Communication

Andy Lücking **Hannes Rieser** **Marc Staudacher**
Bielefeld University, CRC 360 “Situated Artifical Communicators”, B3
{andy.luecking,hannes.rieser,marc.staudacher}@uni-bielefeld.de

Abstract

Classical SDRT (Asher and Lascarides, 2003) discussed essential features of dialogue like adjacency pairs or corrections and up-dating. Recent work in SDRT (Asher, 2002, 2005) aims at the description of natural dialogue. We use this work to model situated communication, *i.e.* dialogue, in which sub-sentential utterances and gestures (pointing and grasping) are used as conventional modes of communication. We show that in addition to cognitive modelling in SDRT, capturing mental states and speech-act related goals, special postulates are needed to extract meaning out of contexts. Gestural meaning anchors Discourse Referents in contextually given domains. Both sorts of meaning are fused with the meaning of fragments to get at fully developed dialogue moves. This task accomplished, the standard SDRT machinery, tagged SDRSs, rhetorical relations, the up-date mechanism, and the Maximize Discourse Coherence constraint generate coherent structures. In sum, meanings from different verbal and non-verbal sources are assembled using extended SDRT to form coherent wholes.

1 Introduction

Recently, the interest in retrieval and representation of *non-sentential speech* has been growing, as the collection (Elugardo and Stainton, 2005) shows. The debate on how to account properly for the phenomena is still ongoing. However, it emerges that it puts further constraints on how mainstream linguistics *should* be done. Non-sentential speech is an essential part of language.[1] Thus, notions such as grammaticality and coherence have to be applicable to it. In this paper, we are concerned more specifically with issues of the semantics/pragmatics-interface of non-sentential speech. We understand this kind of language use as being part of *situated communication* and propose a theory for it. Thus, we start by characterising situated communication. Consider the two examples (1) and (2).

(1) World economic growth slowed noticeably in 2005 from the strong expansion in 2004.

(2) In a two-person dialogue between I and C in a room with some bolts on a table:
  a. I: This bolt in the rear there (while I is pointing)
  b. C: This one? (while C is grasping some bolt)
  c. I: Yes

In opposition to (1), the kind of language use as in (2) is what we call *situated communication*.[2] Language use of this kind can be recognized by a couple of characteristics. First, utterances are typically sub-sentences and not “full-fledged sentences” in a grammatical sense. On a standard account, only sentences (and not parts) express propositions. Still, sub-sentences can be used to express propositions. For example, (2-a) says of a particular bolt on the table that it is the one to be grasped. So, after all, utterances of sub-sentences can express propositional content.

Secondly, such utterances are typically accompanied by linguistically relevant non-verbal behaviour such as pointing gestures or graspings. Deixis is typical for this kind of language use. In (2-b), for example, it is asked of a certain bolt on the table whether it is the one I meant in (2-a). To establish the reference to that bolt, C’s grasping seems to be essential.

Thirdly, such utterances as in (2) can be used to perform speech acts. It can be meaningfully asked what the illocutionary role of such an utterance is (*e.g.* (2-b) is a `Check-back`) and which proposition is thereby expressed. However, it cannot be a property of the expression’s content that makes it express a certain speech act or proposition. For example consider an utterance of ‘scissors’ in a sewing shop, in the rock-paper-scissors-game, or

---

[1] For data on its frequent use, see (Schlangen and Lascarides, 2003). In our corpus (see (Lücking and Stegmann, 2005, p. 15)) 50 instructor’s requests were realized as definite NPs out of a total of 92 dialogue moves including acceptances and repairs.

[2] In contrast, the use of some fragments such as question-answer- or request-answer-pairs is determined by rules of grammar. We are interested in cases which are extremely context dependent and need inference for their resolution. These are cases calling for “resolution-via-inference” in the terms of the Schlangen and Lascarides (2003) approach.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 72–79.

on a shopping list. In each context, the utterance is used to express something different. While the first two can be taken to express a proposition, the inscription on the list cannot. It might merely be some mnemonic device to perform the shopping. Moreover, the different uses of ‘scissors’ seem to be governed by *conventions*. So, a special stock of conventions seems to regulate its interpretation. Being conventions each of them is mutually believed (in some dispositional sense). Together they allow agents the use of sub-sentential utterances and gestures to successfully communicate as (2) shows.

From these three properties of situated communication we derive the minimal requirements for a theory of situated communication. As a framework we are going to use SDRT. Given the use of non-sententials and nonverbal behaviour, the theory has to explain which sentential content a non-sentential utterance expresses and which dialogue move is performed. The explanation has to make use of a special stock of conventions. Moreover, discourse coherence should be explained.

For purposes of illustration we use discourse (3) as our main example:

(3) a. *I: Die rote* $\searrow_a$ *Holzscheibe*
I: The red $\searrow_a$ wooden disc
b. *C:* $\downdownarrows_a$ *Diese?*
C: $\downdownarrows_a$ This one?
c. *I: Ja*
I: Yes

Some comment about (3) is in order. Dialogues like (3) are called *Object Identification Games* and have been examined in project B3 of the Collaborative Research Centre “Situated Artificial Communicators” (SFB 360)[3].

In (3-a) the symbol ‘$\searrow_a$’ indicates, when the stroke of a pointing gesture occurs. The symbol is written after the word whose occurrence is immediately preceding in time. The index indicates the object a the pointing refers to. Likewise in (3-b), the symbol ‘$\downdownarrows_a$’ indicates the grasping of the object a. Two video-stills showing the pointing and the grasping in (3) are provided in Fig. 1(a) and 1(b), respectively.

(3) is a gloss for a corpus entry which has been built from the experimental data. Each corpus entry is a description of a dialogue which occurred in the experimental setting. The corpus annotation format features both verbal and non-verbal elements in such a way that the role of pointing gestures can be studied theoretically. Fig. 1(c) shows a graphical representation of a corpus entry.

In Object Identification Games, two persons, the *instructor* (I) and the *constructor* (C), are involved in a coordination task. It is a two-player-game of spotting an object in a given situation. The instructor has the role of the “description-giver”. The constructor has the role of the “object-identifier”. The players interact by performing moves in the game. The game starts with the instructor’s choosing a certain object out of the parts of a toy air plane spread on a table. She instructs the constructor to identify the object she has chosen by referring to it. The constructor then has to resolve the instructor’s reference act and to give feedback. Thereby, reference has to be negotiated and established using a special kind of dialogue game. The game ends, if the constructor has identified the correct object on the table and the instructor has accepted it.

This paper is organised as follows. We first introduce Standard SDRT. Next, we discuss recent SDRT developments and introduce the concepts doing the explanatory work. We then apply the theory to our main example. In the remaining sections, related research is reviewed and ideas for linking SDRT to Logical Description Grammars are presented.

2 Exposition of Standard SDRT

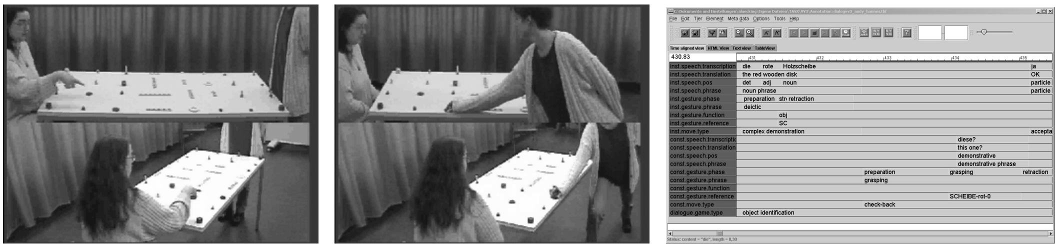
As a dynamic discourse representation theory modelling the semantics/pragmatics-interface, SDRT is an apt framework for modelling situated communication. For our purposes it is important to note that “standard” SDRT as presented in the 2003 book (Asher and Lascarides, 2003) requires its input to be of a type corresponding to sentences in the grammar. To understand this point we illustrate SDRT’s general architecture (Fig. 2) and its implicit notion of discourse construction using the sample dialogue (3).

Since SDRT provides no grammar, the *NL-input* is assumed to be available as *underspecified logical forms (ULFs)* constructed by a *parser*. The underspecification reflects the fact that, in general, the grammar does not determine a unique logical form but a set of possible forms corresponding to the interpretation licensed by the grammar alone. ULFs describe base logical forms, *i.e. SDRSs*.

So, SDRT’s processing begins by assuming

---

[3] `http://www.sfb360.uni-bielefeld.de/`


C:\Dokumente und Einstellungen\aluecking\Eigene Dateien\TASX\VV3\Annotation\dialogvv3_andy_hannes.tbf
File Edit Tier Element Meta data Options Tools Help
Time aligned view
HTML View
Text view
TableView
430.83
431
432
433
434
435

inst.speech.transcription	die rote Holzscheibe ja
inst.speech.translation	the red wooden disk OK
inst.speech.pos	det adj noun particle
inst.speech.phrase	noun phrase particle
inst.gesture.phase	preparation str retraction
inst.gesture.phrase	deictic
inst.gesture.function	obj
inst.gesture.reference	SC
inst.move.type	complex demonstration accepta
const.speech.transcriptic	diese?
const.speech.translation	this one?
const.speech.pos	demonstrative
const.speech.phrase	demonstrative phrase
const.gesture.phase	preparation grasping retraction
const.gesture.phrase	grasping
const.gesture.function	
const.gesture.reference	SCHEIBE-rot-0
const.move.type	check-back
dialogue.game.type	object identification

Status: content = "die", length = 0,30

(a) The pointing in (2a). (b) The grasping in (2b). (c) The annotation in TASX.

Figure 1: Annotation of a natural dialogue in project B3 of SFB 360.

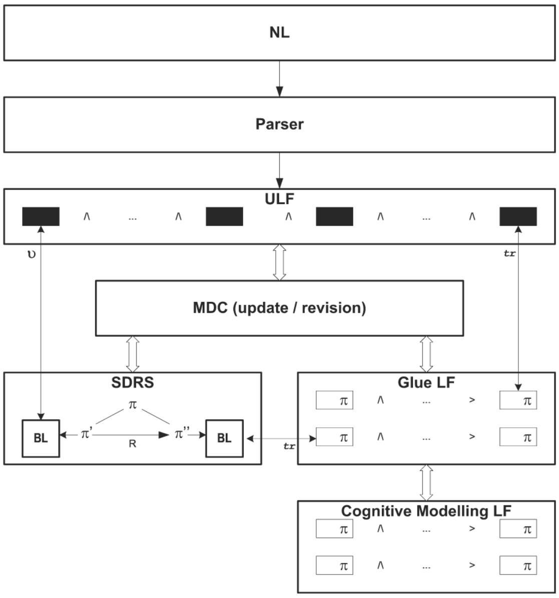


Figure 2: The general SDRT architecture.

some context C (a potentially empty description of SDRSs) and by assuming the ULF of the dialogue's first utterance (3-a) being part of it. In the next step, C is updated with the ULF of (3-b) yielding a new context C'. SDRT's *update mechanism* assumes that the new utterance is related to an available attachment point by means of an underspecified *discourse relation* $R(a,b)$. In this case, the most coherent one is (3-a). However, discourse relations relate only content having sentential satisfaction conditions.

At this point SDRT fails with sub-sentential utterances, if they just have their compositional meaning. The interpretations of (3-b) licensed by the grammar alone are not contents having sentential satisfaction conditions. Intuitively, the new context C' describes SDRSs in which a and b are related through some admissible resolution of R, *e.g.* $Q\!-\!Elab$. So, what $Q\!-\!Elab$ should relate is of the wrong semantic type. Something having satisfaction conditions is required, however in the case of (3-b) an NP-denotation is present.

To make the illustration of the general architecture complete, let us assume that (3-b) had a sentential content. Then its ULF (*inter alia*) would be translated to the *Glue logic* and to the *Cognitive Modelling Logic* in order to resolve underspecification by pragmatic reasoning. The resolutions are translated back to the logic of ULF and added to the description. The update mechanism restricts the resolutions to those that are consistent, *i.e.* describe well-formed SDRSs.

3 Recent SDRT Developments

Recently progress has been made on how difficult NL dialogue data can be handled using SDRT's full theoretical power plus some additional assumptions. For our purposes three papers are of relevance: (Asher, 2002) on Deixis, Binding and Presupposition, (Asher, 2005), Bielefeld Lectures on SDRT, and (Schlangen and Lascarides, 2003)'s work on the Interpretation of Non-Sentential Utterances in Dialogue.

Asher (2002) deals with the following issues, relevant for our example: treatment of presuppositions, analysis of definite descriptions, especially their deictic uses, anchoring of definites in the non-linguistic context, the notion of internal and external anchors, the relation between anchoring and speech act related goals (SARGs), the cognitive effects of anchoring, the generation of mutual belief with regard to an object anchored. These concepts are briefly and somewhat fragmentarily introduced below.

As to presuppositions, Asher argues that the Heim-van der Sandt-Geurts account is incomplete and yields wrong predictions, the reason being that presupposition accommodation in the case of deictically used NPs is not always adequate. Definite descriptions introduce an underspecified relation, called *bridging relation*, between the referent and some other contextually given object, set to identity by default.[4] Deictically used definites have to be anchored to some object in the non-linguistic context. As a consequence, anchoring involves a *de re* attitude towards the object, some sort of *knowing how* needed to solve the conversational goals (SARGs) of the speaker. SDRT uses, in opposition to specifying anchoring contextually as undertaken in Kaplan's Context Theory or Situation Semantics, DRT's external and internal anchors (Kamp, 1990). Anchoring requires linking an agent A's epistemic attitude to conversational goals. If an anchoring relation between the presupposition of a definite ψ and some element in the discourse context exists for the agent A, he is supposed to have a computable means of getting to the referent of ψ from the present non-linguistic context of utterance under some given purpose ϕ; to capture this, a notion of *path* is defined. If the anchoring function of a deictically used definite is accepted by the participants in dialogue, they are assumed to mutually believe that the definite picks out the same object for them. Hence, anchoring amounts to coordination or alignment.

Of similar importance as the discussion of definites, presupposition, binding and anchoring is the handling of fragments in dialogue, since, normally, natural dialogue does not come with utterances which can be mapped onto well-formed sentences in the theory of grammar sense. The idea in (Asher, 2005) is that fragments can be resolved iff the context in which the communication is situated provides us with two things: First, it must be mutual knowledge that a fragment with some meaning has been produced by an agent and secondly, it must be mutually believed that the fragment as produced expresses some more comprehensive content ϕ wrapped around the information reconstructed as a presupposition. In our example, the more comprehensive content ϕ is given by 'Grasp the red wooden disc!' and 'This one?', respectively. The status of these assumptions in the theoretical set up of SDRT is not yet clear, presumably, they belong to Cognitive Modelling, since mental states are involved.

Another approach to fragments is elaborated in (Schlangen and Lascarides, 2003). The idea is to assimilate sub-sentential utterances to sentences since such utterances express sentential content. Thereby, the problem discussed in the last section can be circumvented. From their point of view such utterances have "holes" which need to be filled in in order to express the intended content. Schlangen and Lascarides understand hole-fillers as the resolution of semantically underspecified content (and as such these are not syntactic ellipses). *I.e.* the linguistic form of such utterances is of the category "sentence fragment" which in turn consists of the usual linguistic items such as an NP. The logical form is assumed to have a semantically underspecified relation linking its variables such that each resolution expresses sentential contents, among them the intended one. Schlangen and Lascarides' main thesis is that the resolution of such utterances can be modelled as a by-product of establishing coherence in discourse.

Schlangen and Lascarides found that their approach is more problematic with regard to sub-sentential utterances which need a "resolution-via-inference", *i.e.* a resolution that cannot use the immediate linguistic context containing a "copy" of the material needed (as in the case of short answers to wh-questions). The reason is simply that domain-specific knowledge is necessary.

We, following Asher (2002, 2005), propose a new direction for accounting for this class of utterances. We don't treat such utterances as sentences. Our thesis is that competent speakers have linguistic knowledge in form of situated conventions allowing the speakers to properly use and understand such utterances. Moreover, our original data shows that the role of gestures and graspings is central to correctly resolve newly introduced definites. Without a notion of external anchoring resolution cannot be explained correctly. As a by-product of the introduction of the notion of external anchoring resolution-via-inference becomes more tractable.

4 Coherence from the 2005 SDRT Perspective: A Giant Step for SDRT

SDRT's notion of coherence up to (Asher, 2005) rested on several mechanisms, the use of rhetorical relations and their semantics, especially the division into coordinating and subordinating relations, the use of SDRSs as part of context change potentials in the Kamp-Heim-tradition, the extended definition of up-date capturing revision in dialogue and, finally, the filter mechanism "Maximize Discourse Coherence" (MDC). All these notions were ultimately founded upon the notion of *complete* meaning, of whatever type and however expli-

---

[4] See (Asher and Lascarides, 1998) for more on bridging.

cated. These meanings in turn were conceived of as coming solely from verbal expressions using a construction algorithm in DRT fashion.

This picture fundamentally changes with 2005's SDRT: First of all, the information provided by the fragments of the description giver 'the red wooden disc' and the object identifier 'this one?', respectively, are not complete. The intuition is that the fragments combine with meanings from the context to give us complete meanings. Roughly, we want 'Grasp the red wooden disc!' on the description giver's and 'Do you mean this one?' on the object identifier's side. Once we arrive at complete meanings, the normal SDRT machinery can be put to work again. However, in order to get there, we have to use special postulates, which under specific conditions let agents in cooperative dialogue use these fragments as directives and clarification questions, respectively. (Of course, our account is not restricted to directives and clarification questions. Other postulates would allow other uses.) The missing information for the directive comes from the context at the beginning of the object identification game, in which the director of the experiment assigns the roles of description giver and object identifier, saying for example, 'you, A, tell the other one to grasp one of the objects in the domain' and 'you, B, identify the object described, pointed at etc. and indicate whether you have identified it'. These roles are preserved throughout the contexts developed, at least as a fall back option. In terms of SDRT: The director of the experiment fixes the type of the speech-act-related goals (SARGs) of the participating agents. Secondly, the dialogue is multi-modal as the example shows, the object introduced by the description is anchored to the context by the demonstration. Similarly, the pure demonstrative used in the clarification question is anchored to the context by the object identifier's grasping. Definiteness information is treated as presuppositional, entertaining the idea that presuppositions are locally bound.

On the whole, detailed context information plays a much greater role in the 2005 SDRT version as compared to the standard one, due to the fact that the meaning of the fragments has to be filled up using context information.

5 Tying Things Together

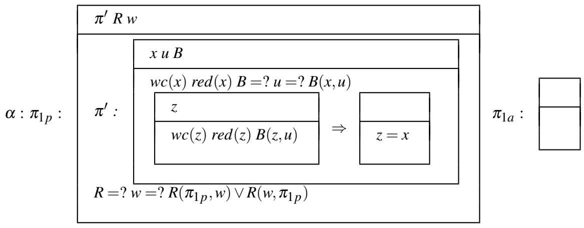
We now apply the theory to our main example (3) using a DRT-style notation. The application of the theory shows how demonstrations, discourse relations, a special stock of conventions and MDC interact in order to arrive at the intended interpretations. We assume that in the context of Object Identification Games a special stock of conventions holds which are represented as axioms of the following form:

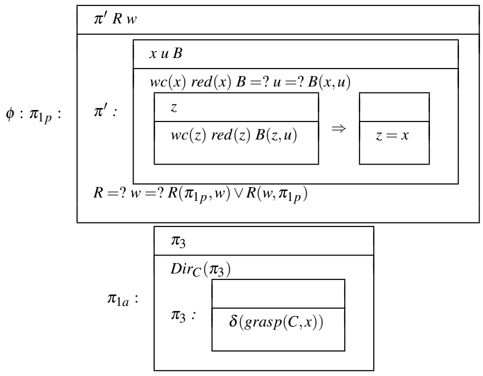
$$K_{I,C}(\alpha(\pi_1)\wedge Ag(\pi_1)=I\wedge(MB_{I,C}(\alpha(\pi_1)\wedge Ag(\pi_1)=I)\rightarrow Say_I(p_\phi)))\rightarrow\alpha(\pi_1)\text{ resolves to }\phi$$

Such conventions express linguistic knowledge which competent communicators in Object Identification Games are assumed to have. The formula can be read as follows: If both communicators I and C know ('$K_{I,C}$') that if I utters α and if it is mutually believed ('$MB_{I,C}$') that if I utters α she says that ϕ, then α resolves to ϕ.

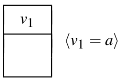
In our example dialogue, we assume that the following convention holds: If I utters 'The red wooden disc' in (3-a), it resolves to the directive addressed to C that she should grasp the object referred to. Of course, not all NP-utterances are directives. So, the relevant convention has to be restricted to situations of a certain type.

Recall that SDRT distinguishes with regard to definites between *presupposed and asserted information*. Consequently, the utterance of (3-a) gives us the presupposed information π_{1p} in α and the asserted information π_{1a} in α. π_{1a} in α should be read as 'There is an SDRS but I don't know which one'. ϕ, in turn, expresses what the utterance α resolves to if the antecedent of the axiom holds.

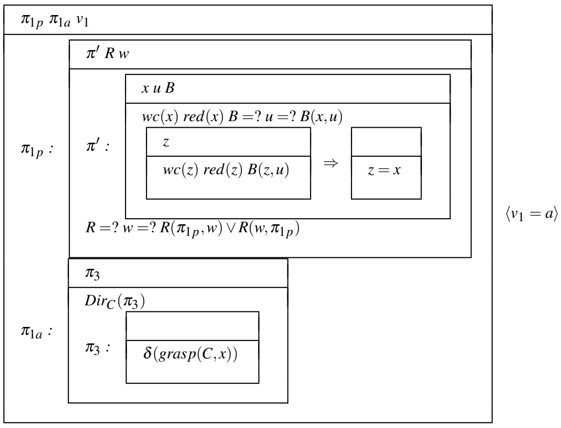




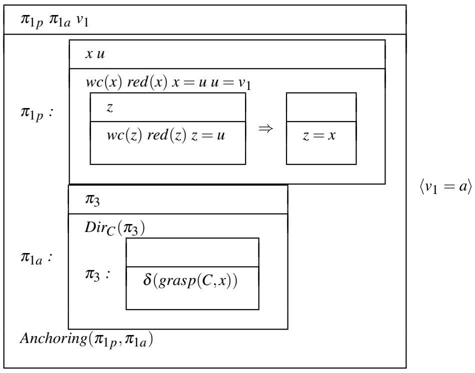
So, ϕ is what we get from the application of the linguistic information to the special convention. We assume a speech act theory style imperative semantics. Consequently, Dir_C is to be read as ‘C is commanded that ...’ and $\delta(grasp(C,x))$ in π_3 is the action commanded, namely that agent C grasp x. For the next step, we have to say how we represent gestural information. The pointing in (3-a) provides very little content. It merely relates some discourse referent to some external object:



Combining the linguistic and gestural information, the result of an apt multi-modal integration strategy is:

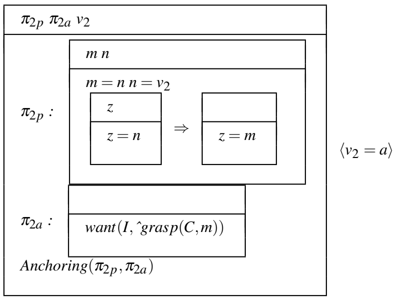


Now, underspecification can be resolved by using a tacit best-update-strategy. Thereby, we resolve the B-relation to identity $(\lambda x.\lambda y.x = y)$, u to the externally anchored v_1, w to π_{1a} and R to *Anchoring*. Thus we get:

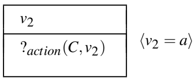


So, in the first turn I introduces a discourse referent v_1 which is externally anchored to the wooden disc a. The directive in π_{1a} presupposes that there is some object which can be grasped by C. The presupposition is satisfied through best-update’s resolution of R to *Anchoring* in such a way that π_{1p} anchors π_{1a}.

The next turn is analysed similarly. There is, likewise, a special convention regulating the interpretation of (3-b) which says that when C utters the deictic ‘Diese?’ she thereby says that she wants to satisfy the directive. Combining the presupposed and the asserted content as before we get:


$\pi_{2p}\ \pi_{2a}\ v_2$
π_{2p} :
$m\ n$
$m = n\ n = v_2$
z
$z = n$
$\Rightarrow$
$z = m$
π_{2a} :
$want(I, \hat{}grasp(C,m))$
$Anchoring(\pi_{2p},\pi_{2a})$
$\langle v_2 = a \rangle$

So, part of what the grasping does is that it externally anchors v_2. However, it seems that graspings have richer but underspecified content since they can be used to perform many things. We reflect this by assigning a highly underspecified content of type “action” to it:



In our dialogue, the grasping presumably carries out the action demanded by I. This suggests that the grasping in (3-b) is used to satisfy I’s request in π_{1a} and part of its SARG. Using best update, this amounts to saying that $?_{action}(C,v_2)$ resolves to $grasp(C,v_2)$ and that $Sat{-}Request(\pi_2 g, \pi_1 a)$ holds. Thus the grasping elaborates on π_{2a} yielding $Q{-}Elab(\pi_{2g},\pi_{2a})$. Usual reasoning additionally gives us $Q{-}Elab(\pi_{1a},\pi_{2a})$ and explains why ‘This one?’ in (3-b) is uttered. While the grasping satisfies the directive (see *Sat−Request*), it might not be *mutually believed* that it is satisfied. So, if $Q{-}Elab(\pi_{1a},\pi_{2a})$ holds, it also mutually believed that it does (using SDRT’s axiom schemata `Sincerity`, `Competence` and `Mutual Belief`). Moreover, by SARG-transitivity, the SARG of π_{1a} is (part of) the SARG of π_{2a}. Thus by satisfying π_{2a}’s SARG the SARG of π_{1a} is satisfied. So, finally, we get the resulting SDRS in Fig. 3.

6 Related Research

Dealing with natural multi-modal dialogue in our paper, we touch on several research areas. Leaving out special SDRT literature here, the focus

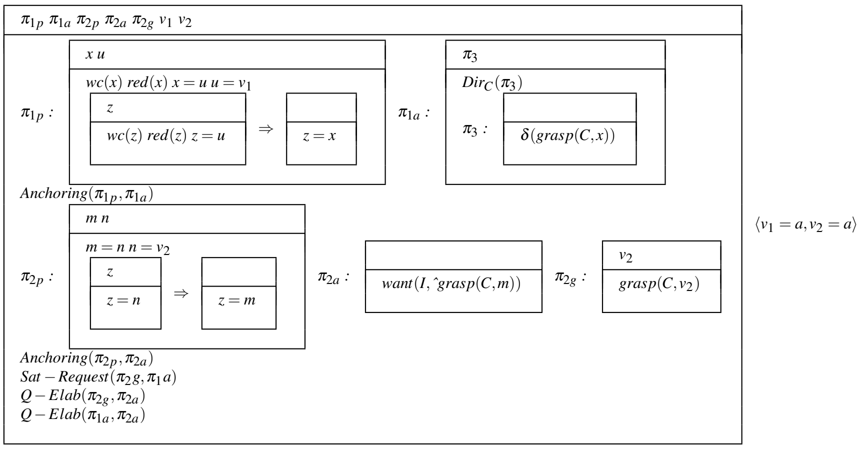

$\pi_{1p}\ \pi_{1a}\ \pi_{2p}\ \pi_{2a}\ \pi_{2g}\ v_1\ v_2$
π_{1p} :
$x\ u$
$wc(x)\ red(x)\ x = u\ u = v_1$
z
$wc(z)\ red(z)\ z = u$
⇒
$z = x$
π_{1a} :
π_3
$Dir_C(\pi_3)$
π_3 :
$\delta(grasp(C,x))$
$Anchoring(\pi_{1p}, \pi_{1a})$
π_{2p} :
$m\ n$
$m = n\ n = v_2$
z
$z = n$
⇒
$z = m$
π_{2a} :
$want(I, \hat{}grasp(C,m))$
π_{2g} :
v_2
$grasp(C, v_2)$
$\langle v_1 = a, v_2 = a \rangle$
$Anchoring(\pi_{2p}, \pi_{2a})$
$Sat - Request(\pi_2 g, \pi_1 a)$
$Q - Elab(\pi_{2g}, \pi_{2a})$
$Q - Elab(\pi_{1a}, \pi_{2a})$

Figure 3: The resulting SDRS.

is on grammar-in-dialogue, description of fragments, and problems of integrating information from other channels.

The issue of syntax-in-dialogue was treated by Schegloff (1979) from the perspective of discourse analysis, mainly focussing on hesitations, restarts, turn construction, and repairs. Clark and Wilkes-Gibbs (1990) generalised the ethnomethodological approach and studied cooperation in syntax production, formulating principles of cooperative contributions for NPs-in-dialogue. A corpus investigation from the perspective of syntax cooperation is provided in (Skuplik, 1999). Fine tuned coordination on all grammatical levels, named 'alignment', forms the backbone of Pickering and Garrod (2004)'s theory, completions and fragments being their favourite examples for establishing implicit common ground. Based on (Skuplik, 1999) and hooking up to SDRT, change of speaker roles, completions and inference in task-oriented dialogue were studied in (Poncin and Rieser, 2000) using Von Wright's Practical Syllogism and Asher and Morreau's Default Inference. A reconstruction of completions and similar phenomena within PTT is undertaken in (Poesio and Rieser, 2006). Recently, even if restricted to sentences/propositions, the interest in retrieval and representation of fragmentary information has been growing, as the collection of articles in (Elugardo and Stainton, 2005) and their introduction to the volume shows. Above all, representation of ellipsis and fragmentary information has been investigated in the paradigm of Dynamic Syntax (Cann et al., 2005; Purver et al., 2005; Purver and Kempson, 2004) for some time, using advanced theory of grammar.

Since SDRT does not come with a worked out construction algorithm, it does not have a multimodal interface. Its contribution to multi-modality issues lies therefore in applying the separation of presuppositional versus assertional information and especially in the notion of anchoring. Principles of interface construction and compositionality matters concerning speech and gesture integration are discussed in (Lücking et al., 2006), see also (Rieser, 2004, 2005), where one can see which problems have to be overcome. Once the mapping from verbal expressions to SDRSs is organised, SDRT could, in principle, be part of an MM interface.

7 Ideas for Linking SDRT Logical Description Grammars (LDGs)

Having sketched how non-sentential utterances can be accounted for from within SDRT, we now address the question how dialogue maps to ULF. In (Asher and Lascarides, 2003, p. 122) it is assumed that some syntax-semantics-interface maps verbal input into ULF, which, judging from the set-up of SDRT (p. 431), forms its bottom layer. ULFs have models in the logic of information content, represented as SDRSs of some sort. In the simple case, where we have no underspecification, we get only one model. In order to get the mapping from language to ULF going, we can start from Muskens' concept of Logical Description Grammars (Muskens, 2001). LDGs use a version of Lexicalized Tree Adjoining Grammar (LTAG) which can capture underspecification in a similar way as the Constraint Language for Lambda Structures (Egg et al., 2001) does, for example concerning PP-attachment, quantifier ambiguity and polysemy. The semantic structures which we can use to tag LTAG-trees can be either type-logical formulae, as in (Muskens, 2001) or DRSs in the style of compositional DRT as in (Muskens, 1996). These we can take as substitutes for single SDRSs. Underspecification could arise due to syntactic structure or semantic ambiguity, *i.e.* we could get several SDRSs for one LTAG-formula. Once we reach this level, we seem to be done, since ULFs can be translated into Glue Logic, the place where the axioms substantiating admissible rhetorical relations are introduced.

We haven't yet tested this assumption in detail, we hope to report about it in the workshop. Observe that with respect to our example we have to face additional problems due to the fragments encountered. As a consequence, we would have to use additional axioms in our mapping process.

8 Conclusion and Further Work

We have given a first sketch of a theory of situated communication by means of SDRT plus special conventions used to determine the communicative meaning of non-sentential utterances. The theory relates gestures to NPs by way of presupposition representation, Anchoring and MDC. It remains to be seen whether alternative grasping representations are better suited to the project. Further work relates to statistical investigations concerning fragments of dialogue moves, coverage results and generalisability.

Acknowledgements

We are grateful to Nicholas Asher for having taught us SDRT in the years 2003-2005 and for letting us work with unpublished SDRT material, especially (Asher, 2005). Our work on SDRT was supported by the CRC "Situated Artificial Communicators," project "Deixis in Construction Dialogue" (DEIKON) at Bielefeld University, funded by the German Research Foundation (DFG). Thanks to the anonymous reviewers whose remarks were helpful for improving our paper.

References

Nicholas Asher. Deixis, Binding and Presupposition. forthcoming in: Festschrift for Hans Kamp, 2002.

Nicholas Asher. Bielefeld Lectures on SDRT, 2005.

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

Nicholas Asher and Alex Lascarides. Bridging. *Journal of Semantics*, 15(1):83–113, 1998.

Ronnie Cann, Ruth Kempson, and Lutz Martin. *The dynamics of language: an introduction*. Syntax and Semantics; 35. Elsevier, Amsterdam [a.o.], 2005.

Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intention in communication*, pages 463–493. MIT Press, Cambridge, MA, 1990.

Markus Egg, Alexander Koller, and Joachim Niehren. The constraint language for lambda structures. *Journal of Logic, Language, and Information*, 10(4):457–485, 2001.

Reinaldo Elugardo and Robert J. Stainton, editors. *Ellipsis and nonsentential speech*, volume 81 of *Studies in Linguistics and Philosophy*. Springer, Dordrecht [a.o.], 2005.

Hans Kamp. Prolegomena to a structural theory of belief and other attitudes. In C. Anthony Anderson and Joseph Owens, editors, *Propositional attitudes: the role of content in logic, language, and mind*. CSLI, Stanford, 1990.

Andy Lücking, Hannes Rieser, and Marc Staudacher. Multimodal integration. Brandial'06, to appear, 2006.

Andy Lücking and Jens Stegmann. Assessing reliability on annotations (2): Statistical results for the DEIKON scheme. Technical Report 3, SFB 360, Bielefeld University, 2005.

Reinhard Muskens. Montague semantics and discourse representation. *Linguistics and Philosophy*, 19:143–186, 1996.

Reinhard Muskens. Talking about trees and truth-conditions. *Journal of Logic, Language and Information*, 10(4):417–455, 2001.

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

Massimo Poesio and Hannes Rieser. Prolegomena to a theory of completions, continuations, and coordination in dialogue. in prep., 2006.

Kristina Poncin and Hannes Rieser. Multi-speaker utterances and coordination in task-oriented dialogue. Report 2000/02, SFB 360, Bielefeld University, to appear in JoP 751. Technical Report 2000/06, SFB 360, Bielefeld University, 2000.

Matthew Purver and Ruth Kempson. Incremental parsing, or incremental grammar? In *Proceedings of the ACL Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pages 74–81. Barcelona, 2004.

Matthew Purver, Ronnie Cann, and Ruth Kempson. Grammars as parsers: meeting the dialogue challenge. To appear in: Research on Language and Computation, 2005.

Hannes Rieser. Pointing in dialogue. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04. Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue*, pages 93–101. Barcelona, 2004.

Hannes Rieser. Pointing and grasping in concert. In Manfred Stede, Christian Chiarcos, Michael Grabski, and Luuk Lagerwerf, editors, *Salience in discourse: multidisciplinary approaches to discourse*, pages 129–139. Nodus Publikationen, Münster, 2005.

Emmanuel A. Schegloff. The relevance of repair to syntax-for-conversation. In Talmy Givón, editor, *Syntax and semantics: Discourse and syntax*, volume 12, pages 261–286. Academic Press, New York, 1979.

David Schlangen and Alex Lascarides. The interpretation of non-sentential utterances in dialogue. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan, 2003.

Kristina Skuplik. Satzkooperationen. Definition und empirische Untersuchung. Technical Report 1999/03, SFB 360, Bielefeld University, 1999.

Engaging in a Multimodal Dialogue with Embodied Conversational Agents along the Virtuality Continuum
(invited presentation)

Elisabeth André
Multimedia Concepts and Applications
University of Augsburg, Germany
`andre@informatik.uni-augsburg.de`

In my talk, I will present a new generation of synthetic characters that are no longer bound to a flat screen, but able to enter a physical world and to engage in a conversation with a human user. Users and characters do no longer inhabit separated spaces, but share an informational and physical reality that is augmented by digital objects. As a consequence, communication has to take into account both the physical and the digital context. The talk will analyse new forms of deixis that are enabled by so-called directing-to and placing-for behaviours and may involve conversational locomotion. Directing-to behaviours rely on a number of verbal and non-verbal means including demonstrative pronouns, eye gaze or pointing gestures to signal the addressee what they are talking about. In my talk, I will show how to emulate such behaviours using an embodied conversational agent. Placing-for behaviours differ from directing-to behaviours by moving objects in the addressee's focus of attention. In my talk, I will present various tangible interfaces developed by our lab to enable placing-for behaviours in the physical space.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, page 80.

Handling Spatial Reference in Visually-Situated Dialogs
(invited presentation)

Geert-Jan M. Kruijff
DFKI GmbH
Saarbrücken, Germany
gj@dfki.de

John D. Kelleher
School of Computing
Dublin Institute of Techinology
john.kelleher@comp.dit.ie

This talk describes the application of computational models of spatial prepositions to visually situated dialog systems. An inherent aspect of these dialogs is reference to objects in the environment in which the agents are situated. The talk will present computational models of topological and projective spatial prepositions that are designed to handle spatial reference resolution and generation in complex visual environments containing multiple objects. These models have been implemented in a human-robot dialog system and the talk will conclude by describing how these spatial models were integrated into the robot architecture.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, page 81.

Measuring and Reconstructing Pointing in Visual Contexts

Alfred Kranstedt B3, Technology
Andy Lücking B3, Linguistics
Thies Pfeiffer C3, Technology
Hannes Rieser B3, Linguistics
Marc Staudacher B3, Linguistics
CRC 360 "Situated Artificial Communicators", Bielefeld University
`alfred.kranstedt@googlemail.com`
{`andy.luecking|thies.pfeiffer|hannes.rieser|marc.staudacher`}`@uni-bielefeld.de`

Abstract

We describe an experiment to gather original data on geometrical aspects of pointing. In particular, we are focusing upon the concept of the pointing cone, a geometrical model of a pointing's extension. In our setting we employed methodological and technical procedures of a new type to integrate data from annotations as well as from tracker recordings. We combined exact information on position and orientation with rater's classifications. Our first results seem to challenge classical linguistic and philosophical theories of demonstration in that they advise to separate pointings from reference.

1 Background

Dealing with pointing as a linguistic device implies dealing with two poles: On the one hand, pointing is bound up with reference. On the other hand, pointing is not precise.[1] Sources for the first pole can be found in philosophical literature, the second pole is supported by psychological research. Wittgenstein (1958, Blue Book, p. 50) gives away the philosophers' private detail that he "may know where a thing is and then point to it by virtue of that knowledge." Butterworth (2003, p. 25) sums up psycholinguistic investigation in stating that pointing "did not allow precise target localization."[2] Obviously, both positions do not fit together. The commonsense view that we can demonstrate objects seems to conflict with the fuzziness of vector extrapolation between index finger and target. Some years ago we started to hypothesize that the "blur" of pointings can be systematically couched in the geometrical concept of the pointing cone (Kranstedt et al., 2006a), and thereby deliver a model of a pointing's extension. This promises to be useful in both linguistics and artificial intelligence – see (Kranstedt et al., 2006b) for an overview. However, camera-based studies that aimed at delimiting the cone's apex angle suffered from the drawback that two-dimensional video data were too poor to derive exact three-dimensional topologies from. To overcome such limitations we pursue an original methodological approach employing audio, video, and body movement recordings simultaneously in a restricted, task-oriented object identification game setting and augmenting them with human annotation. We present some results gained by the empirical study (Section 2) in Section 3. The results play a prominent role in shaping the subsequent outlay of theorizing in Section 4.

2 Empirical Study

The empirical study involves two participants engaged in a restricted object identification game. This task was derived from earlier studies on the use of pointing gestures in referring (Lücking et al., 2004). Each participant gets a certain role, one is called *Description Giver* (henceforth DG) and the other *Object Identifier* (OI). DG and OI are placed in a CAVE-like environment which incorporates a marker-based optical tracking system with nine cameras (6DOF tracker). The information delivered by the cameras is integrated *via* special software and provides points and orientations in an absolute coordinate system, which origin lies in the center of the CAVE-like environment. We tracked the DG only. He sits on a stool and is equipped with carefully positioned

---

[1] We restrict ourselves to concrete pointings here. See (McNeill, 1992) for abstract pointings.

[2] See also (Butterworth and Itakura, 2000).

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 82–89.

markers for the tracking system measuring arm, index finger, hand, and head movements. It is clocked by a frame (1/25 sec.) so that longer movements deliver more tracking data. In addition, the whole scene is recorded from two different perspectives with digital cameras. Speech is captured with the DG's headset. The whole set-up with the prepared DG can be seen in Figure 1, a screenshot from our video recordings. The special gloves used to track the stretched index finger are displayed in Figure 2. Both OI and DG are located around a real table (77.5 × 155.5 cm) with 32 parts of a Lorentz Baufix toy airplane, the experimental domain. The objects' centers were lined up randomly on an underlying grid ensuring that they are laid out equidistantly, see Figure 3. This layout is used for all trials of the study. The outer objects' centers frame an area of 70 × 140 cm. That is, the distance between objects' centers of the same column is 20 cm in neighbouring rows. To exemplify the mapping from rows to distance measures: The distance of the third row from the left, DG's, side of the table is 47.75 cm (2 × 20 cm + 7.75 cm for the outer margin).

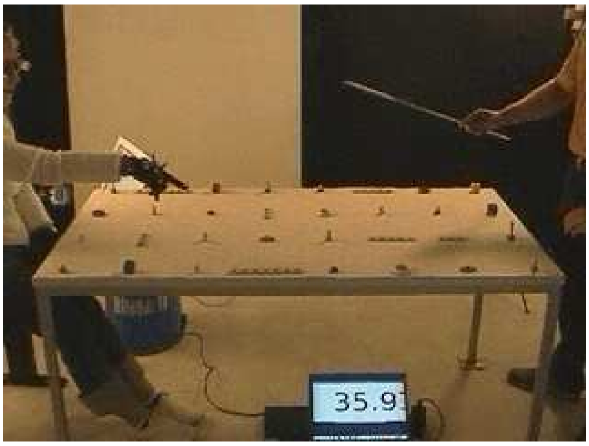


Figure 1: The experimental set-up: The DG sits to the left of the table, the OI stands to the right and has a pointer. The system time needed for synchronizing tracking and recording is displayed on a monitor.

Figure 2: Special gloves

2.1 The Realization of the Experiments

The identification game gets instantiated in two variations, differing in the communicative channels (speech and gesture) the DG is allowed to use:

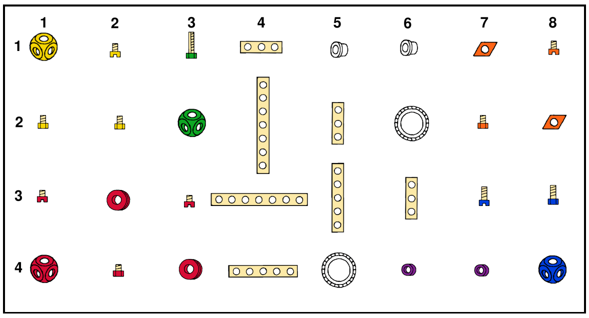


Figure 3: The experimental domain is divided up into eight rows and four columns. It covers an area of 70 × 140 cm. The DG is positioned to the left, the OI to the right of the domain.

- speech plus gesture (S+G Trial);
- gesture only (G Trial).

In each subsetting the DG has to get the object of each of the 32 identification games from the display on the monitor (roughly) in front of him. The order of the objects has been fixed in a presetting. In order to abstract over potential sequence effects, different object presettings have been randomly generated which are iterated over the subsettings and over the whole experimental runs.

The flowchart of the interaction. The interaction between DG and OI is highly restricted to avoid uncontrollable negotiation processes between the interactants. It consists of three formalized steps:

1. Demonstration by DG (bimodal or gestural, according to current subsetting);
2. Interpretation and identification by OI with a pointer only (the referent remains in it's place);
3. Feedback by DG.
   The feedback is restricted to "Ja" (*yes*) in the successful case (accept) and to "Nein" (*no*) in the unsuccessful case (denial). In both cases the identification game terminates and the participants move on, starting with the DG selecting the next object from his display.

2.2 Annotation

46 of the recorded experimental subsettings, 23 with and 23 without speech, enter into analysis. That makes a total of 1472 (46 × 32) demonstrations.

Annotation of the video data has been carried out making use of two software tools, Anvil and Praat. The audio tool Praat[3] was used for the transcription of spoken language, the video films were annotated with the multimodal annotation tool Anvil[4]. Since the concern of the study is pointing, annotation is restricted to DG's first move, that is, to the demonstration act. Annotation is done on several layers (of course, annotating speech is restricted to the S+G Trials):

gesture.phase [*preparation*, *stroke*, *retraction*]; structuring gesture motion according to the trinity established by (McNeill, 1992).

gesture.handedness [*left*, *right*]; for two-handed gestures both values are specified simultaneously.

speech.transcription DG's speech transcribed at the level of words.

speech.number The number of words used in DG's move.

speech.quality [*shape*, *color*, *function*, *position*, *proxy*]; "semantic categories" that are referred to in an utterance (the last one labels taxonomically unspecified nouns, NPs or determiners, like "Ding" (*thing*) or "Das" (*that*) or "Dies Teil" (*this thing*)).

move.referent unique name of object.

move.success [*yes*, α], if the OI could successfully identify the object. Name α of erroneously chosen object otherwise.

Our research interest is the precision of pointing – operationalized in terms of the pointing cone. Accordingly, only those gesture tokens enter into analysis which are purely deictic (showing, e. g., no iconic traits). Furthermore, the success (or failure) of a move should depend on exactly one gesture. We implement this two-step filter in annotation layers, on which annotators have to make suitable decisions:

gesture.validity [*yes*, *no*]; is the gesture a purely deictic one?

move.validity [*yes*, *no*]; Is the game's gesture valid and does the gesture include exactly one stroke?

As a preliminary test procedure for the reliability of the annotation scheme the interrater-agreement between three raters' annotations of one video on the most versatile layers, namely `speech.quality` and `gesture.validity`, has been calculated. With a value of $\mathrm{AC}_1 = 0.9$ for semantic categories and a value of $\mathrm{AC}_1 = 0.85$ for gesture classification, both ratings prove to be quite consistent.[5]

2.3 Processing Tracking Data

The geometrical and temporal information assembled in the tracking data files is processed to deliver quantitative models of pointing. Since we have the orientation and the exact position of the DG's head ("cyclop's eye") and the exact position of the index finger as well as of the referred object, we are able to represent pointing beams as vectors. Based on careful qualitative observations of the subjects' pointing behavior, we assume two different yet plausible ways of anchoring and orienting a beam: Firstly, origin and orientation may be given exclusively by the index finger (*index finger pointing*, IFP); secondly, the beam can be anchored in the (tip of the) index finger, but the orientation is determined by projecting a beam from the cyclop's eye (point between the eyes of the DG) through the anchor (*gaze finger pointing* GFP). Thus GFP models the presumed influence of gaze on pointing in a strict way. The "true pointing vector" (if there is such a thing) probably is somewhere in the middle between the extremes defined by GFP and IFP and might be reconstructable by interpolating the two. Using our IADE (*Interactive Augemnted Data Explorer*) framework (Pfeiffer et al., 2006), a tool for recording, analysis and (re-)simulation of multimodal data, both pointing beams can be visualized in simulation videos, as shown in Figure 4. The (extreme) case shown exemplifies that both kinds of pointings can diverge a great deal. The picture also shows the idealized beam. Idealized beams are the straight lines connecting the pointing vector's anchor with the point in space inhabited by the object referred to. Comparing the GFP and IFP beams with their ideal counterparts delivers a measure of pointings' "faultiness". As error estimates we employed two gauges, angular and orthogonal deviation. *Prima facie*, angular de-

---

[3] `http://www.praat.org/`

[4] `http://www.dfki.de/~kipp/anvil/`

[5] Identity of ratings cannot be ascribed to chance on a risk level of $\alpha = 0.01$. AC_1 is the first order agreement coefficient developed in (Gwet, 2001). Most of the other layers have been evaluated extensively in a precursor study.

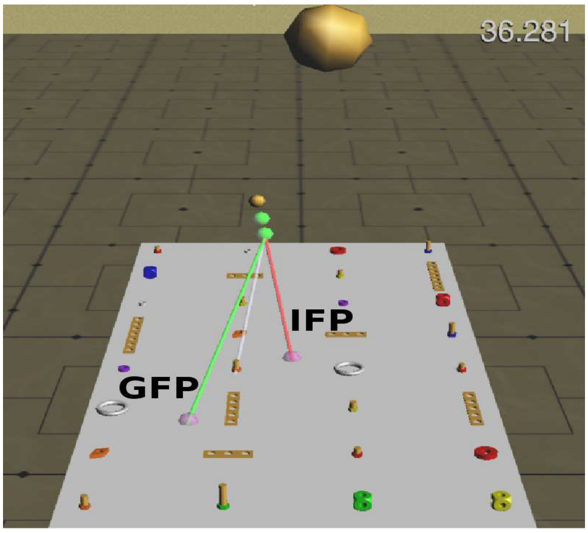


Figure 4: Simulating IFP, GFP, and the idealized pointing beam in between.

viation is more suitable since angles are distance-independent. Angular deviation is calculated as the angle γ spanning between the simulated and the ideal pointing vector. A schematic depiction is given in Figure 5. However, given short distances between anchors and objects, even small variances result in a high angular deviation. As a comparative value, deviations are measured directly on a meter scale in terms of orthogonal deviation. It is given by the distance between the object's point in space and its orthogonal projection P onto the (prolongation of the) simulated beam.

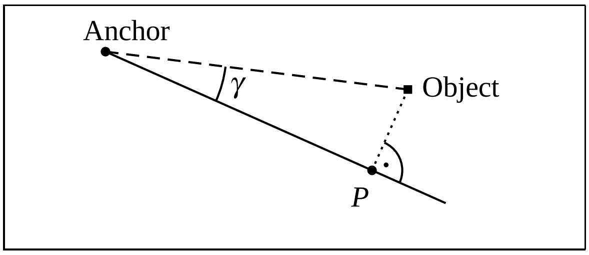
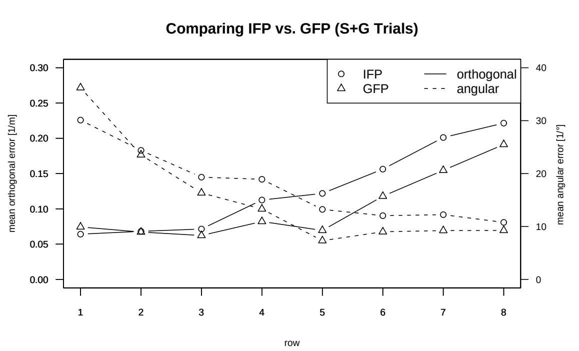
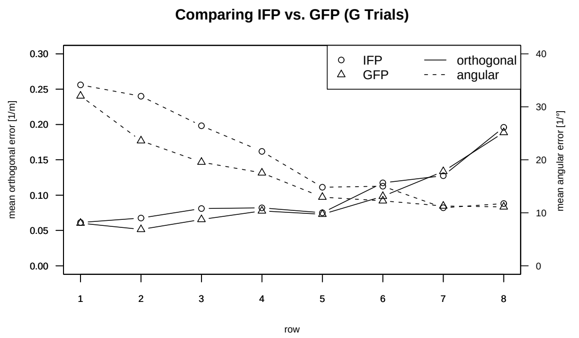


Figure 5: Error estimates for pointing beams: $\gamma =$ angular deviation, $\overline{P\ \text{Object}}$ = orthogonal deviation.

3 Some Results



(a) Deviations in S+G Trials



(b) Deviations in G Trials

Figure 6: Comparing IFP and GFP by means of orthogonal and angular deviation over the rows of the domain.

Given the outlined measurements we can compare IFP and GFP in terms of preciseness. Plotting the means of their deviations (both orthogonal and angular) against the associated row, the measured IFP and GFP values exhibit a similar envelope, as can be seen from Figure 6, and thus do not permit a preference in either direction. As expected from the calculations of the error gauges explained above, angular deviation decreases with increasing distance. In opposition, orthogonal deviation rises from row to row. Demonstrations fail their targets – sometimes even by a lot. What do they aim at instead? Plotting the intersection points of tracked demonstrations with the tabletop over the rows of the domain, we get a visual pattern forming "clouds": The impacts of pointing vectors, from IFP as well as from GFP, are distributed around the object to be indicated. The farther the target lies, the more blurred is the shape of the associated scatter-plot, ranging from near circles in the first row to broad and fuzzy regions in the last one. Representative for all plots, Figure 7 shows IFP in G Trials. The ommited ones look quite similar. The plot is based on all DGs' demonstration acts, which, for each object, are averaged by their median. This ensures that each gesture token, be it a long or a short one, makes the same (viz. one) contribution. To make the "clouds"-issue clearer, the areas which are hit by GFP beams stemming from both the S+G and the G Trials are displayed as a bagplot – a bivariate generalization of a boxplot (Rousseeuw et al., 1999) – in Figure 8. The inner

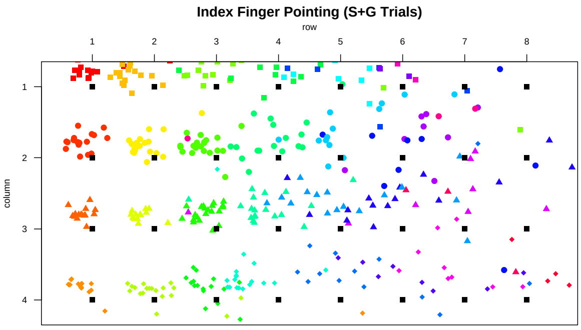


Figure 7: Medial intersections of IFP beams with tabletop in G Trial.

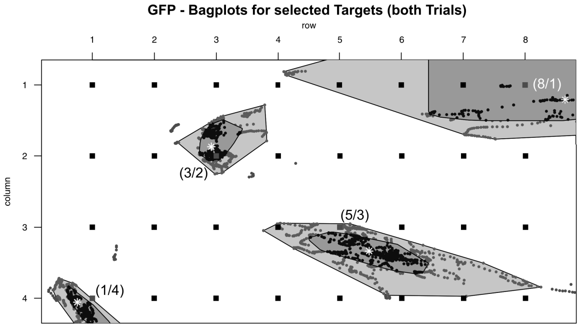


Figure 8: Areas of selected GFP beams. The star marks the median of the cloud, the inner hull frames the data distributed around it.

hull covers 50 % of the data distributed around the “depth median”. Using this representation, it can be nicely seen how the clouds grow and get lengthier from row one onwards. In this respect, clouds already exhibit cone-like properties. Those distribution patterns will serve as a basis for us to extrapolate the delineation of the pointing cone from the data (in addition to other parameters and findings of our study – cf. (Kranstedt et al., 2006b, subsec 3.3.4)).

The growing of the clouds may be due to two effects: Firstly, the mean variation of pointing vectors increases naturally with distance; secondly participants *systematically and intentionally* point over the domain when referring to an object in row eight. Thus, they are using what can be called a *gestural hyperbole*. That this behavior is indeed governed by a successful strategy can be seen from Figure 9: The number of identification errors in the G Trials decreases clearly in the last row as compared to the seventh row. There it can also be seen that the participants could identify all objects in the first three rows. The number of failures increases rapidly from the fifth row onwards. Since there are nearly no failures in the S+G Trials we ignore them here.

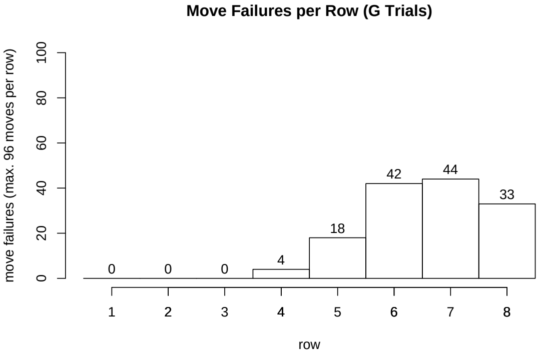
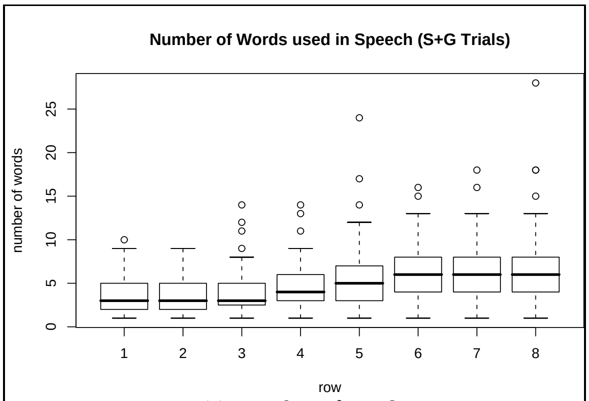
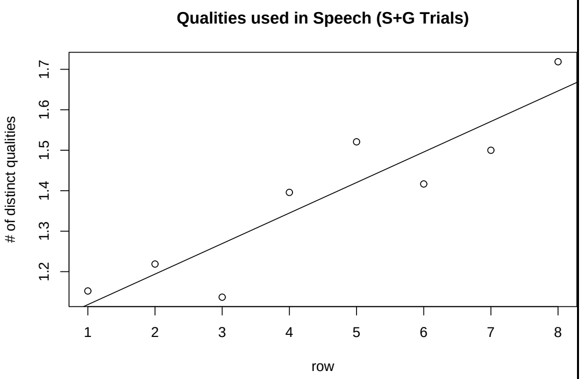


Figure 9: Frequency of identification failures per row.

Considering the S+G Trials, we find two tendencies: 1. The farther away an object is, the more words accompany the gesture; 2. The farther away an object is, the more semantic categories are used to accompany the gesture. Both regularities are depicted in Figure 10.



(a) Number of words



(b) Number of semantic categories

Figure 10: The increasing complexity of co-gestural speech over the rows of the domain.

Since we know about the gestures’ loss of discriminatory power wrt distance – this is evident from the findings presented in Figures 7, 8, and 9 – the tendencies can be ascribed to balancing efforts. The DG compensates his pointing at distance with verbal contributions if he is allowed to, as is the case in the S+G Trials. This can be corroborated by contrasting failures in the G Trials (Figure 9) with the number of words in the S+G Trial (Figure 10(a)): The row in which the number of words increases coincides with the row where the failures increases – both phenomena show up between the fourth and the fifth row. A related increase is shown by the regression line in Figure 10(b) which indicates that the averaged frequency of semantic categories used in referring to objects in the different rows rises from one in row one to nearly two in row eight. Since usually one word expresses a single semantic category, this finding implies that speakers have to use more words if they do not employ a gesture. Indeed, we gained the same result in earlier studies where we expressed it the other way round: *Gestures save words*. It shows the semantic significance of pointings; when referring to objects with gesture and speech people need less words than in referring by speech alone.

Summary of Results. Given a dense domain made up of concrete, equally distributed objects like the one presented here, our findings suggest that pointings can successfully demonstrate objects in the pointer's proximity. However, they seem to do so by delimiting the area the object lies in. The delimiting capacity of pointings diminishes in distance. There, the gesture's spatial cues have to be enhanced by verbal descriptions. The findings are as follows:

- *Pointing is a highlighting (and not a referring) device*. The beams do not meet their targets, they rather encircle them. However, these "clouds" become blurred in the distance. This finding might replace the object-pointing/region-pointing distinction in our earlier work;

- *Pointing breaks down with distance*. Given the density in our setting, pointing starts to get error-prone somewhere between 60.25 and 77.75 cm, which are the distances of the fourth and the fifth row, respectively, measured from DG's end of the table;

- *Distance-dependence of gesture vs speech portions*. Pointings do not permit to single out an object on their own. Distal pointings are accompanied by more complex verbal descriptions. The latter are "more complex" in both numerical and semantic respects. This could not be rendered significant in our earlier studies, but is now in accordance with (van der Sluis and Krahmer, 2004);

- *Partitioning of the pointing domain*. Our earlier investigations suggested a tripartition of the domain into a proximal, a middle, and a distal area. The results presented here suggest a structured domain, too. However, structuring according to the increasing descriptive data would divide the domain into two areas, split somewhere in the middle.

4 On Demonstration: Relating Empirical Data to Theory

If we want to establish a logic of demonstration, we have to deal with at least two questions: Firstly, "What is the logical form (LF) of a demonstration accompanying some expression such as a pure demonstrative or a definite description?" And secondly, "Which models M will be adequate to go with this LF and to provide truth conditions, a suitable notion of entailment and the like?" Since we have to discuss very fundamental things here, we do not want to go into matters of speech act theory, dynamic semantics or sophisticated dialogue theory. Note also that these issues are different from multi-modal integration matters (cf. (Lücking et al., 2006)). For ease of reference, we abbreviate demonstrations, more precisely, their stroke, using ↘.

In order to deal with the LF problem and the M one in a down-to-earth manner let us first recapitulate the empirical findings (referred to as Oi below) which of course do not partition matters into LF-related and M-related ones by themselves. For the start of the discussion we take models M as tupels comprising *inter alia* a domain D.

O1. Empirical domains are structured with respect to DG's proximal and distal relations to targets. Actually, a parameter or index DG should be supplanted by IF or GF-relations indexed by DG;

O2. Demonstrations do not, as a rule, hit their targets;

O3. Demonstrations single out sets of objects rather than single objects;

O4. Proximal demonstrations are distributed around their targets;

O5. Distal demonstrations can encompass other objects besides their target;

O6. The farther away the object demonstrated, the more words accompany the demonstration;

O7. There is a phenomenon of indirect interpretation dubbed *gestural hyperbole*.

How can we account for O1,...,O7? Let us first turn to LF. Here O2 and O3 seem to be of prime importance. Since the extension of a demonstration is not an individual but a region, represented as a set, the LF of a demonstration must not be modelled with a constant but with a one-place predicate. Doing this, a problem arises concerning pure demonstratives like 'this' and their concomitant demonstrations, since $\searrow$ and 'this' are of different type (*predicate* vs. *term*). However, supported by our empirical data and in a way opposed to tradition, we can argue that $\searrow$ does not contribute to the term 'this' as such (and whatever might be used in its place) in *e. g.* 'This is nice' but to the utterance as a whole. As a consequence, we might aim at $[\lambda x(\searrow(x) \wedge nice(x))this]$ to represent the meaning 'This [thing] is demonstrated and is nice'.

Obviously, O1, O4, O5 point into a similar direction and lead us onto issues related to M: While in our setting the extension of a demonstration in the proximal region encompasses only a single object, in the distal region (or in more dense domains) there might be more. This we can accommodate by adding a spatial structure to the model: The model contains a function assigning a coordinate to every object in the domain. Hence we get *distances* between DGs and objects and can do justice to the domain's density. The extension of a demonstration is determined by DG's position, the direction of his pointing, and some pointing domain (in our setting idealized as objects on a surface). To this end, DG's context c determines, *inter alia*, his index finger coordinate (functioning as the anchor point), denoted by c_{IDG}, and the coordinates of his eyes (for orienting the vector in case of GFP), denoted by c_{GDG}. For every gesture occurrence $\searrow_i$ in the context, there is a list of coordinates $[p]_i$ describing the relevant spatial properties of the pointing hand, denoted by $c_{\searrow_i}$.[6] In addition, the pointing domain is represented as a surface s, also part of the context, and denoted by c_s. $\searrow_i$'s intension fixes its extension for every pointing context depending on c_{IDG}, c_{GDG}, $c_{\searrow_i}$, and c_s. It is represented as the function $f : \langle c_{\mathrm{IDG}}, c_{\mathrm{GDG}}, c_{\searrow_i}, c_s \rangle \mapsto Ext(\searrow_i)$ which determines the pointing predicate's extension for all pointing contexts. f is defined in terms of the chosen pointing model, *i. e.* IFP or GFP. So, there is a choice between the two functions *IFP* and *GFP* yielding for every tuple $\langle c_{\mathrm{IDG}}, c_{\mathrm{GDG}}, c_{\searrow_i}, c_s \rangle$ a possibly different pointing cone.[7] The geometrical intersection of this cone with the surface s (*e. g.* the table) yields a region. The collection of the objects in this region *is* the extension $Ext(\searrow_i)$. Moreover, f has the characteristics indicated by the empirical findings, *i. e.* it assigns a smaller extension to pointings in the proximal region and larger extension to pointings in the distal region, extensions having fuzzy borders. It should be clear that from DG's context c a presumably fuzzy partitioning of the domain D into a proximal and a distal subdomain can be reconstructed (*e. g.* that part of the table is distal where there is more than one object in every region pointed at).

Assuming such, the truth conditions for a DG's utterance 'This is nice' amount to ' 'This $\searrow_1$ is nice' is true in context c iff there is exactly one object $o \in D$ such that $o \in Ext(\searrow_1)$ and o is nice.'

If we decide the issue this way, what is going to happen in cases of pointings into the distal region? Well, their felicity will depend on the density of the domain and the meaning of the linguistic information going with the demonstration, which should perhaps have the force of a definite description. This accounts precisely for O5 and O6. If an expression *cum* demonstration turns out to be false wrt the proximal or the distal region, we have to consider a solution along Gricean Pragmatics using the Quality Maxim. The same holds true for the more dramatic O7 cases of indirect interpretation, which are always false on a literal reading.

In sum, if we follow the arguments suggested by the empirical data, we have to separate demonstration from referring, which goes against the prevalent philosophical tradition represented by work from Wittgenstein, Davidson or D. Kaplan. Instead of ending up with two referring terms for

---

[6] The conceptualisation follows here the work on pure indexicals such as 'I'.

[7] Where c_{GDG} plays no rule for IFP.

the example above, one for ‘this’ and the other one for ↘ to be related by identity, we get an additional predication, a context-dependent subset of D. In a sense, the consequences of the “type shift” of demonstrations from individuals to sets are less dramatic than trying to do without such a shift. Doing without the shift would mean to consider demonstrations as pure referring entities and to treat their non-satisfaction in a neo-Gricean way, perhaps along the lines of Levinson’s *Presemantic Pragmatics* (Levinson, 2000).

5 Outlook

To determine the parameters defining function f which assigns extensions to demonstrations in a given context, we have to fix a model for the pointing cone. So a main task in the near future is to derive the delineation of the cone from the empirical data. The concept of a cone and our findings fit well with processing paradigms of pointing represented in (developmental) psychology and linguistics where the function of demonstration is *inter alia* seen in “focusing the attention” (of the addressee). Here as well as in Human Computer Interaction the cone is part and parcel of a precise model for pointing gestures.

However, the empirical findings reported above are difficult to reconcile with traditional philosophical and linguistic theories of demonstration. Therefore we want to compare them to stipulations dealing with demonstration by Wittgenstein, D. Davidson, and D. Kaplan, where the main focus will be “Which paradigmatic cases of demonstration did philosophers found their theories on?”

References

George Butterworth and Shoji Itakura. 2000. How the eyes, head and hand serve definite reference. *British Journal of Developmental Psychology*, 18:25–50.

George Butterworth. 2003. Pointing is the royal road to language for babies. In Sotaro Kita, editor, *Pointing: Where Language, Culture, and Cognition Meet*, chapter 2, pages 9–33. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, Gaithersburg (MD).

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006a. Deixis: How to determine demonstrated objects using a pointing cone. In Sylvie Gibet, Nicolas Courty, and Jean-Francois Kamp, editors, *Gesture in Human-Computer Interaction and Simulation*, pages 300–311. Springer, Berlin.

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006b. Deictic object reference in task-oriented dialogue. In Gert Rickheit and Ipke Wachsmuth, editors, *Situated Communication*, pages 155–207. Mouton de Gruyter, Berlin.

Stephen C. Levinson. 2000. *Presumptive Meanings*. MIT Press, Cambridge, MA.

Andy Lücking, Hannes Rieser, and Jens Stegmann. 2004. Statistical support for the study of structures in multi-modal dialogue: *Inter*-rater agreement and synchronization. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog ’04—Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue*, pages 56–63, Barcelona.

Andy Lücking, Hannes Rieser, and Marc Staudacher. 2006. Multi-modal integration. Brandial’06.

David McNeill. 1992. *Hand and Mind—What Gestures Reveal about Thought*. Chicago University Press, Chicago.

Thies Pfeiffer, Alfred Kranstedt, and Andy Lücking. 2006. Sprach-Gestik Experimente mit IADE, dem Interactive Augmented Data Explorer. In *Dritter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR*, Koblenz. Accepted paper.

Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. 1999. The bagplot: A bivariate boxplot. *The American Statistician*, 53:382–387.

Ielka van der Sluis and Emiel Krahmer. 2004. The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the ICSLP*.

Ludwig Wittgenstein. 1958. *The Blue and Brown Books—Preliminary Studies for the “Philosphical Investigations”*. Harper & Row, New York.

Verbal or visual? How information is distributed across speech and gesture in spatial dialog

Kirsten Bergmann, Stefan Kopp
Artificial Intelligence Group
University of Bielefeld
P.O. 100131, 33501 Bielefeld, Germany
{kbergman, skopp}@techfak.uni-bielefeld.de

Abstract

In spatial dialog like in direction giving humans make frequent use of speech-accompanying gestures. Some gestures convey largely the same information as speech while others complement speech. This paper reports a study on how speakers distribute meaning across speech and gesture, and depending on what factors. Utterance meaning and the wider dialog context were tested by statistically analyzing a corpus of direction-giving dialogs. Problems of speech production (as indicated by discourse markers and disfluencies), the communicative goals, and the information status were found to be influential, while feedback signals by the addressee do not have any influence.

1 Introduction

In spatial dialog like in direction giving, humans make frequent use of speech-accompanying gestures. By ”gesture” we mean expressive movements of the hands and arms while speaking. According to (McNeill, 2005) there are four semiotic dimensions within these gestures, as there are iconicity, metaphoricity, deixis and temporal highlighting (beats). Iconic features of gestures present visual information about concrete referents, while metaphoric features refer in the same way to abstract referents. Deictic features point to concrete or abstract referents within the external space, and beats are small and fast movements that structure utterances. One often finds several of these features mixed in the same gesture. This paper focuses on gestures that have their major dimensionality in iconicity and deixis, and we present a study that investigates how information is distributed across these gestures and their concomitant speech.

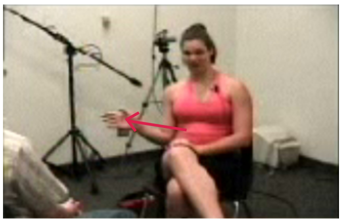


Figure 1: Gesture accompanying the utterance ”take a right” as an example for gestural redundancy.

Gestures are temporally coordinated with speech as well as closely related to the content of the verbal utterance they accompany (McNeill, 1992). The semantic synchrony of both modalities can be thought of as a continuum of co-expressivity, with gestures encoding completely the same aspects of meaning as speech on one extreme. Although both modalities express information in their specific way, we refer to this as *redundancy*. Figure 1 gives an example for redundant meaning in speech and gesture. The utterance ”take a right” contains an action (”take”) and a direction (”right”). Both aspects are expressed as well by the accompanying dynamic gesture made to the right. That is, these two features are communicated redundantly by speech and gesture. At the opposite extreme of the continuum there are gestures encoding aspects that are not uttered verbally, in other words these gestures *complement* speech. In figure 2 an illustrating example for complementarity is given. The direction giver talks about an entrance and visualizes the entrance by gesture. The major content conveyed by speech is the existence and function of an entity, namely

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 90–97.

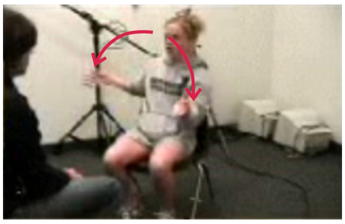


Figure 2: Gesture accompanying the utterance ”it’s the entrance” as an example for gestural complementarity.

being the entrance. Without the accompanying gesture the recipient’s mental representation of the entrance could take different shapes, but the gesture visualizes the arch-shaped architecture of this specific entrance. So the specification of the entrance’s shape is a complementary feature of the speech-accompanying gesture. Interestingly, there seems to be a 50:50 distribution of redundant and complementary gestures (Cassell et al., 2000; Cassell and Prevost, 1996), and even the blind distribute semantic components across the modalities (Iverson and Goldin-Meadow, 1998). The question is when people gesture at all, how they distribute information across speech and gesture. What are the influencing factors? So far, research has not been able to give any satisfying answers on this. McNeill (1992) contends that representational gestures are more likely to be used for newsworthy concepts. Cassell and Prevost (1996) analyzed manner-of-motion verbs and accompanying gestures using semantic features to distinguish between redundant and complementary gestures. They found rhematic information with a focus marking newness or contrast resulting mainly in complementary gestures, while thematic information with a focus marking contrast is accompanied mainly by redundant gestures. Yan (2000) studied gestures from a house description experiment using semantic features to classify redundant and complementary gestures. He developed a hierarchy of rules that managed to predict 60% of the gestures. His major findings are that the introduction of single/multiple object(s) is accompanied by complementary gestures, while redundant gestures are used to localize objects. Bavelas et al. (2002) report findings suggesting that gestures are used to compensate for problems of verbal encodability. Kita and Özyürek (2003) found cross-linguistic variations in iconic gestures, indicating that gestures are shaped simultaneously both by spatial properties of the referents and the way the spoken language packages information. Furthermore, Melinger and Levelt (2004) found first direct evidence that the decision to gesture influences decisions about what is explicitly mentioned in speech or is omitted.

The aim of this study is to find factors that can explain the observed occurrence of redundant and complementary gestures. For this purpose we have included both, meaning itself and the wider dialog context in the analysis. A level for comparing the semantics of speech and gesture has to be established firstly. The following steps aim at problems of verbal encodability as well as different kinds of feedback that signal understanding or non-understanding. Moreover, the particular communicative goals, as Denis (1997) identified them in route directions, as well as the information status might have an impact on the co-expressivity of speech and gesture as well.

2 Method

Our corpus analysis takes place within the scope of a study done at the Northwestern University in Chicago (Kopp et al., 2004). In the following, this study is described briefly, supplemented by a description of the annotation scheme developed for our purpose.

2.1 Participants

28 undergraduates (11 males and 17 females) participated in the experiment as direction givers. All of them were native speakers of English. They got the task to describe a route across Northwestern University’s Campus to another person they thought was unfamiliar with the campus.

2.2 Materials

Ten different routes existed, each of them starting at the building where the experiment took place, and connecting five locations on the campus.

2.3 Procedure

Each direction giver got a list of ten routes and was asked to sort out those ones she/he did not feel comfortable to give directions for. Among the remaining routes one was selected randomly. In order to guarantee comparable conditions, the participant was instructed to make her-/himself familiar with the route by walking it. Afterwards she/he was seated face-to-face with the direction follower. They were instructed to make sure that the direction follower understood the directions and would be able to find the way on her/his own. Audio- and videotapes were taken of each dialog. For the videotape, four synchronized camera views were recorded.

2.4 Coding

Some annotation has been done in the scope of other studies (Kopp et al., 2004; Kopp, 2005). This includes the transcription of the direction giver's words and the segmentation of the occuring iconic and deictic gestures. Moreover, the gesture morphology has been annotated, that is *hand shape*, *hand orientation* and *hand location*. The latter includes shape and extent of the trajectory, which is used to judge the gesture's semantics in the following.

In the scope of our own corpus analysis a total of 1508 gestures out of 10 different dialogs were annotated by two coders using the tool *Praat*[1] to transcribe the words of the direction follower, and the multimodal annotation tool *Anvil* (Kipp, 2004). The following levels of annotation have been added to the corpus: (1a) speech semantics, (1b) gesture semantics (2) problems of verbal encodability, (3) dialog acts, (4) communicative goals, and (5) information status.

2.4.1 Semantics of speech and gesture

The central annotation levels are gesture and speech semantics. In a first step, the lexical affiliate of each gesture, i.e. the word(s) deemed to correspond most closely to a gesture in meaning, has been determined (Schegloff, 1984). For each utterance one or more semantic features (SFs) were annotated both for the gesture and its lexical affiliate. Judging the semantics of speech and gesture is not an easy task. Because of their underspecifity gestures can not be interpreted without looking at their verbal context. Therefore, the risk of circularity is given, when gesture and speech semantics are overhastily equated (McNeill, 2005). To devoid this circularity, we first determined the idea unit underlying the multimodal utterance. Based on this, we judged the semantic contributions of both modalities. An example of the procedure of semantic interpretation is given in figure 3. The utterance refers to ”Cook Hall”, and the underlying idea unit encloses information about the appearance and location of this entity. The information that the referent is a building with a rectangular front emerges from the photo, while a look at the map reveals that ”Cook Hall” is on the right side of the route being described. With this additional information one can fasten down the distribution of information: Verbally the direction giver introduces an entity as ”big building” which is to the ”right”. According to this, we assign the following SFs to the verbal utterance: ENTITY, SIZE and RELATIVE POSITION. The accompanying gesture visualizes the shape of the building. Additional information about the relatively large extent of the gesture is adopted from gesture morphology. Thus we annotate the SF categories SHAPE and SIZE for this gesture, not relative position since the gesture is made in front of the speaker and not to his right.

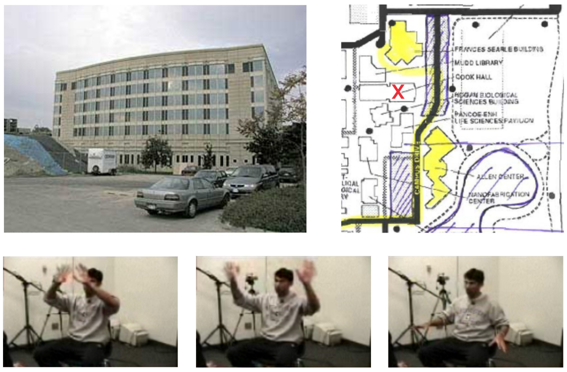


Figure 3: Gesture accompanying the utterance ”you're gonna see a big building to your right” with additional information about the referent in the form of map and photo.

For the overall set of SFs, semantic categories developed by Jackendoff (1983) have been modified depending on the domain of spatial discourse. The following categories adequately cover the semantics of both speech and gesture in our corpus, given with the rules used to annotate speech semantics:

- ENTITY: Streets, paths, buildings, signs etc.
- RELATIVE POSITION: Prepositions characterize information about the spatial position of entities, e.g. ”on your left” or ”behind the parking lot”.
- ACTION: Information about actions, verbally conveyed by motion verbs like ”walk”, ”go”, ”head”, ”follow” etc.
- DIRECTION: Directional information concerning actions is realized verbally with adverbs like ”left/right” or

---

[1] http://www.praat.org

”north/south/west/east”.

- PATH: There are three variants of paths: (1) bounded paths, characterized by prepositions like ”from” or ”to”, (2) paths along a reference object, characterized either by verbs like ”pass” or by prepositions like ”along”, ”through” or ”around”, and (3) paths running relative to a reference object, characterized by verbs like ”follow” or by prepositions like ”on”.

- SHAPE: Words like ”circular” or ”zig-zaggy” are annotated as shape.

- SIZE: Adjectives like ”huge” or ”small” are coded as size.

- AMOUNT: An amount of entities can be verbalized by numerals or by words like ”several” or multiple”.

- PROPERTY: Other properties of entities, except size and shape.

Concerning the meaning of gestures the same categories are used. The first decision to be made is applied to the dynamics of each gesture. A gesture can be either dynamic or static. Dynamic gestures include a trajectory between starting point and target point, while static gestures only consist of a posture at a target position. In the latter case either RELATIVE POSITION, SIZE or AMOUNT are taken into consideration. Typically, positioning gestures are done with one hand, while sizes are visualized with both hands, but in case of doubt the (verbal) context is decisive. If two entities are localized, AMOUNT is annotated additionally. For dynamic gestures there is a wider range of possibilities. In a first step one has to distinguish gestures referring to actions and gestures referring to entities. For the latter ones the SFs SHAPE, SIZE, AMOUNT and PROPERTY are considered. Supportive for the coder is a look at the gesture morphology where gesture shapes may be found (Sowa, 2006). If the gesture conveys a SHAPE, typically the trajectory or the inner sides of the hands form it. SIZE can be found in a dynamic gesture as well, because sometimes a ”scaling” movement refers to the size of entities. Moreover, the morphology clearly contains information about the extent. Typically, AMOUNT is assigned to a dynamic gesture if it refers to more than two entities. In these cases RELATIVE POSITION is annotated as well. PROPERTY is used if any properties of entities except the above ones are visualized, e.g. smoke out of a chimney. If the gesture refers to an action, we annotated the SF ACTION in either case. In addition, either DIRECTION or PATH are conveyed. Directional gestures are pointing gestures, visualizing the direction of an action, while paths are visualized with a ”sweeping” movement of the hands. Sometimes the SHAPE of the path is depicted additionally.

2.4.2 Verbal encodability problems

We coded two different characteristics for problems of verbal encodability: discourse markers and disfluencies. Both kinds of characteristics have been coded for their occurrence (either within the particular gesture’s lexical affiliate, or directly before it). A special case of discourse markers are hedges, which are defined as ”words whose job it is to make things more or less fuzzy” by Lakoff (1972). ”kind of”, ”sort of”, ”somehow”, ”like” etc. are considered to be *more fuzzy hedges*. Disfluencies reflect production problems coming along with spontaneous speech. According to Shriberg (1999) the following features are coded as disfluencies: (1) filled pauses (”uh”, ”um”), (2) repetitions (”the the”), (3) repairs (”that’s called Cook Buil- Cook Hall”), and (4) false starts (”and then you gonna may- once you get to the end of the building”).

2.4.3 Dialog acts

Following the annotation scheme DAMSL (Dialog Act Markup on Several Layers) by Allen and Core (1997), we analyzed how the co-expressivity of speech and gesture is influenced by (non-)understanding signals. In DAMSL, forward looking functions state how an utterance constrains the future actions or beliefs of the hearer, and affects the discourse. We used the utterance tags *Statement*, *Influencing-addressee-future-action* and *Info-request*. Backward looking functions indicate how the current utterance relates to the previous dialog. We coded the utterance tags *Acknowledge* (”okay”, ”aha” etc.), *Repeat-Rephrase*, and *Completion* as understandings signals, and *Answer* referring to the forward looking info-requests.

2.4.4 Communicative goals

In terms of the communicative function of a dialog act, according to Denis (1997) two major components can be identified, as there are actions/instructions and striking points along the route, so-called landmarks. Based on this, Denis develops several categories of communicative goals that can be distinguished in route directions. Our segmentation of these categories in the corpus is based on the preceding annotation of forward looking functions. Utterances tagged as *Statement* or *Influencing-addressee-future-action* were assigned to the following categories:

- *Reorientation*: Instruction to change the orientation, e.g. ”turn right”
- *Locomotion*: Instruction aimed to reduce the distance between the actual position and the destination, e.g. ”go straight on”
- *Action+Landmark*: Instruction combining action and landmark, e.g. ”cross X”, ”turn left at X”, ”go past X”
- *Landmark*: Reference to landmark without localization or further description
- *Landmark with spatial orientation*: Localization of a landmark, e.g. ”there’s a road in front of you”
- *Landmark description*: Non-locating description of landmarks, e.g. ”it’s a big pink colored building”

2.4.5 Information status

Finally, we coded the information status for each SF using the following states: *new* for SFs introduced in the dialog, *evoked* for SFs already given verbally and *evoked by gesture* for SFs already given only by gesture.

In general, annotation-based corpora depend on subjective judgements of the coders, and reliability of these judgements is mandatory. We reached a mean Kappa value of κ=0.774 (SD=0.101) indicating substantial agreement among the two coders on a test set of about 20% of the corpus. Especially judging speech and gesture semantics with a set of categories is always difficult and approximative, and one could imagine more or less categories. We established our category set iteratively in order to adequately cover the relevant meaning, while at the same time ensuring reliability.

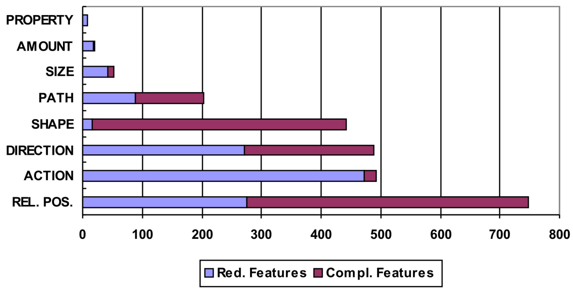


Figure 4: Distribution of the different kinds of redundant/complementary SFs.

3 Results

In the course of judging the gesture semantics, each gesture got assigned between one and five SFs: 51,1% of the gestures have one SF, 31,8% of them have two SFs, and 17,1% have three or more SFs. Among these SFs, 48.63% are redundant while 51.38% are complementary to the accompanying speech. This distribution supports earlier findings by Yan (2000) and Cassell and Prevost (1996) on a level of semantic features. In terms of gesture-wise consideration, one finds 31.7% of the gestures being completely redundant, that is they do not have any complementary SFs. Another 38.9% of the gestures do not have any redundant SFs and therefore are exclusively complementary. Finally 29.7% of the gestures have both redundant and complementary parts. Figure 4 summarizes the number of times that different types of SFs occur in gestures.

The first analysis of the corporal data concerns problems of verbal encoding that become apparent in discourse markers and disfluencies. If there are any discourse markers in speech, there is a significantly higher proportion of complementary SFs in the accompanying gestures (χ^2=13.625, df=2, p=0.001). In addition, the frequency of redundant SFs is decreased in these cases (χ^2=24.279, df=2, $p<0.001$). Concerning redundancy the same findings hold for disfluencies. Gestures accompanying disfluent utterances also have a significantly lower proportion of redundant SFs (χ^2=6.813, df=2, p=0.033), while there is no correlation of disfluencies and complementarity (χ^2=2.128, df=2, p=0.345). Compared to the overall temporal occurrence of gestures in our corpus, gestures accompanying discourse markers or disfluencies occur more frequently before their lexical affiliate (e.g. ”there is a [little... like...] kind of an alley”).

Further analysis has been done concerning the influence of the direction follower’s feedback on the distribution of information across the modalities. This feedback manifests either in interposed questions or in understanding signals. Regarding interposed questions there is neither a significant relationship between the resulting answers of the direction giver and redundancy in gesturing (χ^2=3.272, df=2, p=0.195), nor is there any influence of backward-looking utterances and complementarity (χ^2=1.604, df=2, p=0.448). Regarding positive feedback of the direction follower, the time passed since the last understanding signal may be relevant for judging the influence of these signals on co-expressivity of speech and gesture. The following time-intervalls have been tested: 0-4.99s, 5.00-9.99s, 10.00-19.99s, 20.00-29.99s and $>$30.00s. Across all intervall lengths, we did not find any significant influence of utterances tagged as *Acknowledge*, *Repeat-Rephrase* or *Completion* on the number of redundant SFs of speech-accompanying gestures (χ^2=7.079, df=8, p=0.528), nor on the frequency of complementary SFs (χ^2=8.325, df=8, p=0.402).

Furthermore, we analyzed the influence of communicative goals on the frequency of gesturing in general. The majority of annotated communicative goals is accompanied by exactly one gesture (76.3%), while 10.9% do not have any accompanying gestures, and 12.9% are accompanied by two or more gestures. Nevertheless this distribution depends on the kind of communicative goal (see figure 5). Descriptions of actions without any reference to landmarks (*Reorientation*, *Locomotion*) as well as utterances of the category *Landmark* do have one accompanying gesture in the majority of cases. *Landmark descriptions* are more often uttered without gesturing, while *landmarks with spatial orientation* tend to go with two or more gestures.

In addition, we tested the influence of communicative goals on the co-expressivity of speech and gesture inference-statistically.

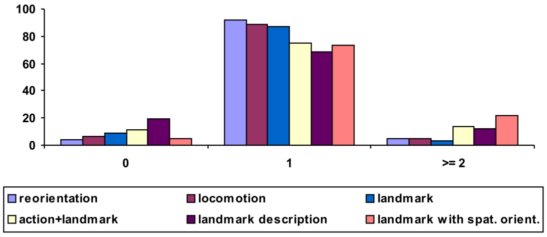


Figure 5: Frequency of gestures per communicative goal.

- *Reorientation*: If the direction giver instructs the direction follower to change the direction, the accompanying gestures are characterized by a significantly higher proportion of redundant SFs (χ^2=227.998, df=2, p$<$0.001). ACTION and DIRECTION are the types of SFs found in these gestures. The number of complementary SFs is decreased in case of reorientations (χ^2=46.578, df=2, p$<$0.001). PATH, DIRECTION and RELATIVE POSITION are the kinds of SFs that are used complementarily.

- *Locomotion*: Concerning the number of redundant SFs in speech-accompanying gestures locomotions are similar to reorientations. The number of redundant SFs in gestures accompanying utterances tagged as *Locomotion* is significantly higher than expected (χ^2=54.303, df=2, p$<$0.001). Again, ACTION and DIRECTION are found to be used most frequently. Regarding the influence of locomotions on the number of complementary SFs in speech-accompanying gestures, there is no significant relationship between those two variables (χ^2=2.029, df=2, p=0.363).

- *Action+Landmark*: Concerning the redundancy in gestures accompanying utterances of the kind *Action+Direction*, two or more redundant SFs occur more often than expected (χ^2=98.904, df=2, p$<$0.001). They are usually of the kinds ACTION and DIRECTION as in the case of locomotions and reorientations, but also of the kinds PATH and RELATIVE POSITION. The proportion of complementary SFs is increased in this category (χ^2=26.179, df=2, p$<$0.001). Especially one complementary SF is used relatively often.

- *Landmark*: In this category the frequency of redundant SFs is significantly decreased, (χ^2=106.632, df=2, p$<$0.001), while the number of complementary SFs is higher than expected (χ^2=46.423, df=2, p$<$0.001). RELATIVE POSITION and SHAPE are found to occur most frequently in the gestures when the direction giver mentions a landmark.

- *Landmark description*: The same findings as for landmarks hold for landmark descriptions. The proportion of redundant SFs in accompanying gestures is lower than expected (χ^2=88.432, df=2, p=0.001) and the proportion of complementary SFs is higher than expected (χ^2=33.582, df=2, p$<$0.001). Moreover, the SFs RELATIVE POSITION and SHAPE are also the ones used most often.
- *Landmark with spatial orientation*: In the case of landmarks with spatial orientation there is a large number of gestures with RELATIVE POSITION as the only redundant SF (χ^2=110.852, df=2, p$<$0.001). Gestures with more than one redundant SF occur rarely. Concurrently, the frequency of complementary SFs is lower than expected (χ^2=79.427, df=2, p$<$0.001). If there are any complementary SFs they are of the kind RELATIVE POSITION or SHAPE.

To sum up, one may say that actions are described with speech-accompanying gestures that have more redundant SFs, while the proportion of redundant SFs is decreased when conveying information about landmarks. Concerning complementarity there are more such SFs than expected in gestures that belong to the categories *Landmark*, *Landmark description* and *Action+Landmark*. Less complementary SFs can be observed when referring to landmarks with spatial orientation.
Concerning the influence of the information status of the SFs, the only found correlation exists for the category ENTITY. The redundancy of gestures accompanying the introduction of entities is decreased, while utterances referring to evoked entities are accompanied by gestures with a higher proportion of redundant SFs than expected (χ^2=13.012, df=2, p=0.001). Moreover, the frequency of complementary SFs is slightly increased in case of new entities, while evoked entities are accompanied by gestures with less complementary SFs (χ^2=4.480, df=2, p=0.106).

4 Discussion

To our knowledge, this study is the first one analyzing the influence of dialog context and communicative goals on the distribution of information across speech and gesture. Our analysis of the direction giving dialogs reveals three major factors influencing the co-expressivity of speech and gesture, while others were found not to do so. First, problems concerning verbal encoding have an effect on the distribution of meaning, leading to more complementary and less redundant SFs in gestures. This goes together with the results of Bavelas et al. (2002) who found more non-redundant gestures when people had to describe pictures that were hard to encode. It seems as if people compensate for such verbal problems by adding complementary information to gestures. Second, the co-expressivity of speech and gesture is influenced by communicative goals. Instructions are accompanied by gestures with more redundant SFs, while gestures referring to landmarks are characterized by more complementary SFs. When giving directions, instructions are really important for the direction follower to find her/his way, especially reorientations and actions referring to landmarks. For this reason it would make sense to convey this information redundantly. However, Beattie and Shovelton (2006) recently found, that speakers tend to convey salient information gesturally. One could argue that at least in case of reorientations it would be difficult to have complementary SFs beyond ACTION and DIRECTION in gesture, but in fact we found gestures with SHAPE, PATH or RELATIVE POSITION as complementary SFs. Nevertheless, the number of complementary SFs is decreased significantly and information about actions, directions and sometimes paths is conveyed redundantly instead. Concerning the larger number of complementary SFs when referring to landmarks, one should think of the particular strengths and weaknesses of both modalities. Shapes and positions can often be easier visualized with hands and arms, than uttered verbally. In these cases the risk going along with complementary meaning in gesture, that is being overlooked by the dialog partner, is accepted. In the category *Landmark with spatial orientation* the localization is conveyed by speech, and in consequence there are less complementary SFs in the accompanying gestures. In the same sense one can interpret the found relationship between communicative goals and the use of gestures in route directions in general. Actions and landmarks have one accompanying gesture in the majority of cases. Descriptions of landmarks are not necessarily accompanied by any gesture.

Within landmark descriptions there may be contents, e.g. colors, that can not even be visualized. In contrast, gestures occur more often when entities are set in relation to one another, as in *Landmarks with spatial orientation*.

Third, the introduction of entities goes along with slightly reduced redundancy and increased complementarity in gesturing. So findings of Yan (2000) are supported tendentiously, but the influence can only be observed for entities, not for other kinds of SFs.

Finally, no influence could be found for feedback signals of the dialog partner, but there are at least two aspects relativising these results. First, the direction followers were not really unfamiliar with the campus. So their interposed questions do not reflect real understanding problems. In fact, the questions were of the kind ”what is the color of the building?” or ”how long does it take to get from here to there?”. Second, and even more important is the fact that only verbal signals of understanding have been annotated. Because of the video quality it was not possible to code nonverbal signals of feedback, although there is no doubt that such signals like head movements or facial mimics are equally good for signaling understanding or non-understanding.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialogue Act Markup in Several Layers.

Janet Bavelas, Christine Kenwood, Trudy Johnson, and Bruce Philips. 2002. An Experimental Study of When and How Speakers Use Gestures to Communicate. *Gesture*, 2:1:1–17.

Geoffrey Beattie and Heather Shovelton. 2006. When size really matters. *Gesture*, 6:1:63–84.

Justine Cassell and Scott Prevost. 1996. Distribution of Semantic Features Across Speech and Gesture by Humans and Computers. In *Proceedings of the Workshop on Integration of Gesture in Language and Speech*.

Justine Cassell, Matthew Stone, and Hao Yan. 2000. Coordination and Context-dependence in the Generation of Embodied Conversation. In *First International Conference on Natural Language Generation*.

Michel Denis. 1997. The Description of Routes: A Cognitive Approach to the Production of Spatial Discourse. *Current Psychology of Cognition*, 16:409–458.

Jana M. Iverson and Susan Goldin-Meadow. 1998. Why People Gesture When They Speak. *Nature*, 396:228.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press: Cambridge, MA.

Michael Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Boca Raton: Florida.

Sotaro Kita and Asli Özyürek. 2003. What Does Cross-Linguistic Variation in Semantic Coordination of Speech and Gesture Reveal?: Evidence for an Interface Representation of Spatial Thinking and Speaking. *Journal of Memory and Language*, 48:16–32.

Stefan Kopp, Paul Tepper, and Justine Cassell. 2004. Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 97–104, New York, NY, USA. ACM Press.

Stefan Kopp. 2005. The Spatial Specificity of Iconic Gestures. In Klaus Opwis and Iris-Katharina Penner, editors, *Proceedings of KogWis05. The German Cognitive Science Conference*, pages 112–117. Basel: Schwabe.

George Lakoff. 1972. Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. In P.M. Perantean, J.N. Levi, and G.C. Phares, editors, *Papers from the 8th Regional Meeting: Chicago Linguistics Society*, pages 183–228.

David McNeill. 1992. *Hand and Mind - What Gestures Reveal about Thought*. University of Chicago Press: Chicago.

David McNeill. 2005. *Gesture and Thought*. University of Chicago Press: Chicago.

Alissa Melinger and Willem J.M. Levelt. 2004. Gesture and the Communicative Intention of the Speaker. *Gesture*, 4:119–141.

Emanuel A. Schegloff. 1984. On some gestures' relation to talk. In J. M. Atkinson and J. Heritage, editors, *Structures of Social Action*, pages 266–298. Cambridge University Press.

Elisabeth Shriberg. 1999. Phonetic Consequences of Speech Disfluency. In *Proceedings of the International Congress of Phonetic Sciences*, volume 1, pages 619–622.

Timo Sowa. 2006. *Understanding Coverbal Iconic Gestures in Shape Descriptions*. Akademische Verlagsgesellschaft Aka: Berlin.

Hao Yan. 2000. Paired Speech and Gesture Generation in Embodied Conversational Agents. Master's thesis, MIT, School of Architecture and Planning.

Cross Recurrence Quantification Analysis of indefinite anaphora in Swedish dialog. An eye-tracking pilot experiment

Philip Diderichsen
Lund University Cognitive Science
philip.diderichsen@lucs.lu.se

Abstract

A new method is used in an eye-tracking pilot experiment which shows that it is possible to detect differences in common ground associated with the use of minimally different types of indefinite anaphora. Following Richardson and Dale (2005), cross recurrence quantification analysis (CRQA) was used to show that the tandem eye movements of two Swedish-speaking interlocutors are slightly more coupled when they are using fully anaphoric indefinite expressions than when they are using less anaphoric indefinites. This shows the potential of CRQA to detect even subtle processing differences in ongoing discourse.

1 Introduction

There exists an extensive literature within linguistics on the topic of referring expressions and their discourse function (Ariel, 1990; Gundel, Hedberg & Zacharski, 1993; Lambrecht, 1994; Prince, 1981; Walker & Prince, 1993). Almost everyone agrees that pronouns need to be 'in focus', i.e. highly mentally accessible, to be used felicitously, while full indefinite noun phrases, at the other end of various types of givenness scales or hierarchies, do not need to be accessible to the same degree.

These issues have been extensively studied within linguistics, and also in many reading experiments within psycholinguistics. But, mainly for technical reasons, it has not been as thoroughly studied whether the theories also hold for unconstrained spoken conversation. This study is a first step to do just that. In the spirit of Trueswell and Tanenhaus (2005), it will be attempted to bridge the methodological gap between psycholinguistics and the more qualitative conversation analysis tradition, in order to evaluate hypotheses about the mental status of entities referred to using anaphora.

A specific anaphor type will be targeted here, namely indefinite *one*-anaphora (Dahl, 1985). Contrasting examples of a 'regular', definite pronoun, and an indefinite *one*-anaphor are shown in example (1) and (2) below.

(1) A: I heard that Ahmed bought a Ferrari.
B: That's right! I saw **it** outside the stadium yesterday.

(2) A: I heard that Ahmed bought a Ferrari.
B: Really? I saw **one** outside the stadium yesterday.

This relatively rare type of pronoun is chosen for its unique set of features, which intersect at two extremes of certain givenness scales (Gundel et al.s *Givenness Hierarchy*, for instance). These anaphora are at the same time pronominal and indefinite noun phrases. Therefore, accounts like the Givenness Hierarchy do not account well for this type of expression. Gundel et al. propose that referents in focus — typically expressed using pronouns — must necessarily be uniquely identifiable and 'type identifiable' as well. Indefinite pronominal expressions like *one*-anaphora are in focus[1] as well as type identifiable, but not uniquely identifiable. How can this be? Either, indefinite *one*-anaphora are not really in focus, or the Givenness Hierarchy is unable to account for them.

---

[1] — as demonstrated in examples (1) and (2): the meaning of *one* depends on an antecedent just as much as *it* does.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 98–105.

It will be shown that more anaphoric indefinites are indeed more in focus than less anaphoric indefinites. This supports a new integrated account of givenness with multiple independent cognitive dimensions, see figure 1.

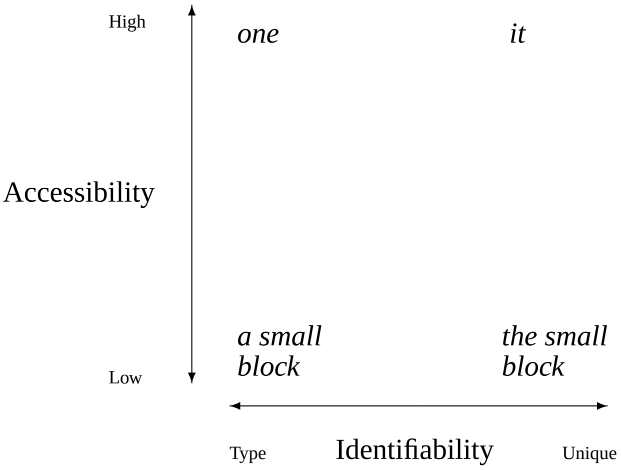


Figure 1: Two-dimensional model of cognitive statuses licencing referential form. Assumed accessibility in the listener licences pronominalization and assumed identifiability in the listener licenses definiteness.

The data are collected using an adaptation for dialog of the 'visual world' eye-tracking paradigm popularized by Michael Tanenhaus and colleagues (Tanenhaus et al., 1995; Tanenhaus & Trueswell, 2005). This enables one to follow ongoing spoken language processes without interfering too much with the task. Eye-tracking is an ideal technique to use to investigate whether pronouns really are 'in focus' for the listener, as is claimed in most accounts. Eye movements presumably reflect attentional states in regard to a visual scene more transparently than any other measure currently available. It will be possible to determine when an entity referred to is in visual focus relative to the utterance of a certain type of referring expression, and if there are systematic differences in the amount of attention paid to the referents of various anaphor types.

The data analysis is inspired by the work of Richardson and Dale (2005), who employed cross recurrence quantification analysis (CRQA) to show that speakers' and listeners' eye movements are coupled when producing and listening to monologs about a picture of the cast of a sitcom. They found that the highest recurrence of gaze patterns occurs at a lag of 1650 ms in the listener's gaze pattern relative to the speaker's. Hadelich and Crocker (2006) found somewhat longer 'eye-eye spans' in conversational dialog, defined as the time between the onset of a speaker's last fixation on an object before mentioning it and the onset of a listener's first fixation on it after hearing it mentioned. The eye-eye spans ranged from about 1700 ms to about 2000 ms, and were shorter the more narratively grounded (i.e. given) the expressions were. Richardson and Dale state that results like these are in the approximate range of the combined results from eye-tracking studies of isolated production (Griffin & Bock, 2000) and comprehension (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). Griffin and Bock find an eye-voice span of about 900 ms, Tanenhaus et al. reported voice-eye lags of 145 to 250 ms after the offset of the uniquely specifying word in an expression, depending on distractors and pragmatic context. Furthermore, Chambers et al. (2002) found lags of 350-400 ms after the onset of target words when processing was not facilitated by pragmatic context. But in addition to this, several studies have shown shorter 'voice-eye spans' than the 400-800 ms mentioned by Richardson and Dale, and sometimes even predictive eye movements, depending on the context (Allopenna et al., 1998; Kamide et al., 2003).

On the basis of the widely held view that pronouns must be 'in focus', it is hypothesized that more anaphoric expressions will be associated with higher recurrence rates. It is also hypothesized that the listener's eye movements will parallel the trajectory of the speaker for longer stretches of time when more anaphoric expressions are used.

2 Materials and method

2.1 Participants and experimental setup

The experimental task was a version of the picture copy task (see for instance Gullberg et al. (1997)), implemented as a virtual building block task. One person told another to build a simple 'space invaders'-style pixel mosaic figure from a representation of the finished figure on the screen.

Four native speakers of Swedish volunteered for the experiment. They were assigned to two conversational dyads. Each participant acted as instructor once and as constructor once, yielding 4 conversations in all. The instructor had visual access to a representation of the constructor's screen.

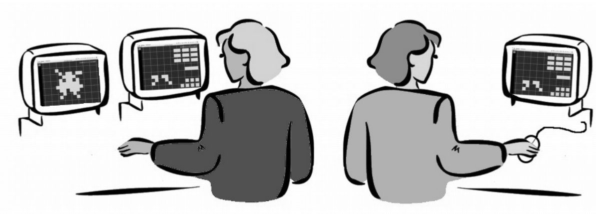


Figure 2: The experimental setup. The instructor (on the left) can see a copy of the constructor's (on the right) screen image. Both participants have their eye movements tracked during the task.

The instructor's task was to freely describe the target figure in such a way that the constructor would be able to build an exact copy of it. The constructor was free to respond. No constraints with respect to choice of expressions or strategy were put on either instructor or constructor.

As for the techical setup, the instructor had two computer screens in front of him or her, both Apple 19" flat panel LCDs. One of them was connected to an Apple G4 PowerMac, and had the sole purpose of displaying the target figure. The other was connected to an Apple G5 PowerMac, and mirrored the constructor's display. The constructor had one display, also a 19" flat panel LCD. This display was also connected to the G5. Only the constructor was allowed to use the mouse, which was used to build the figure by dragging and dropping blocks.

Both conversation partners were wearing bicycle helmets, each mounted with an SMI eyetracker, a head camera, and the Polhemus head tracking system. The sound side of the interaction was captured by two Sony ECM-66B tie clip microphones. Each conversation was preceded by a 13-point calibration procedure.

2.2 Data analysis

The occurrence of singular referring expressions in a specified part of the interactions were counted. Only the referring expressions used to refer to blocks presumed to be 'unused' (cf. Prince's term), i.e. available for use to the right on the constructor's screen, were counted. It is a widely held assumption that definite noun phrases require uniquely identifiable referents, and since there are not many of these among the unused blocks, singular definite expressions were expected to be used seldom.

The referring expressions were classified with respect to ellipsis, definiteness value, and modification. Ellipsis is when the head noun is missing from a noun phrase. This category includes pronouns. Full noun phrases are those that have a head noun. There are two definiteness values: *definite* and *indefinite*. Modification is classified into the two categories *unrestricted* and *restricted*. Unrestricted noun phrases are those that have no modifiers, or at most non-descriptive and uninformative ones. An example of a modified, but unrestricted noun phrase would be *en till sån*, lit. 'one more such'. Restricted noun phrases are those that have descriptive and informative modifiers (restricting the meaning of the noun phrase), such as *en till sån mellanlång* 'one more such half-long'. As can be seen from this example, restricted noun phrases may include non-descriptive modifiers as well as descriptive ones. These examples of unrestricted and restricted noun phrases are all elliptical, but full noun phrases can also be either unrestricted or restricted, whether they are modified or not. This sometimes depends on the amount of information in the head noun. For instance, an expression like 'a/the block' is an unrestricted full NP, while *en fyrar* 'a four-block'[2] is considered a restricted full NP.

After tabulating the singular referring expressions it was clear that there were almost exclusively indefinite noun phrases, and that most of them lacked a nominal head (see table 1 below). The indefinite, elliptical noun phrases could be subdivided further into two groups, unrestricted and restricted elliptical expressions. The unrestricted ellipsis group can be considered fully anaphoric, whereas the restricted group resemble full noun phrases because of the informative modifier(s), and are therefore less anaphoric. Importantly, most types of singular indefinite noun phrase begin with (or consist of) the same word, namely *en* 'a/one', in the Scandinavian languages[3]. This means that one cannot always know from the first word in a referring expression whether it is a pronominal form or not.

Eye movements where measured in a time window of 3 seconds before and after the onset of the word *en*. These eye movement sequences were compared against each other using cross recurrence quantification analysis (CRQA) (Marwan & Kurths, 2002; Marwan, Thiel & Nowaczyk, 2002).

---

[2] *fyrar* is a nominalization of the numeral *fyra*, 'four'.

[3] Just as is the case with definite full noun phrases and definite pronouns, e.g. *den* 'it/the/that'

CRQA is a method of non-linear data analysis originally used within physics, and consists of a number of quantitative measures carried out on cross recurrence plots (CRPs). CRPs are a way of visualizing the coupling between the time series from two different dynamical systems. CRPs are produced by plotting black and white dots in a coordinate system where both axes are time axes, one for each of the time series being compared. For each point on the x axis, a dot is plotted on each point of the y axis: black if the state of one system is close to the state of the other system (i.e. *recurrent*), white if it is not. In this way, each point in time of one time series is compared to all points in time of the other time series. The state of each system is in this case simply defined as the two-coordinate gaze position of each interlocutor. Example CRPs are shown in figure 3 below.

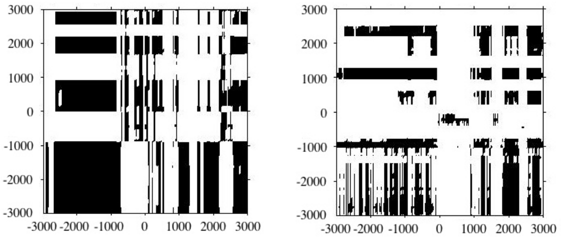


Figure 3: Cross recurrence plots. The plot on the left is from the unrestricted group, the one on the right from the restricted group. Both axes are time axes centered at the onset of *en* 'a/one'. The unrestricted CRP is more dense than the restricted one, which was generally the case in the data.

Certain parameters of the recurrence measures have to be specified. The dimensionality *m* of the underlying phase spaces was set to 2, since the eye movements play out as two-dimensional dots in a plane. The parameter τ represents a time delay between the two time series. This was set to 0. This means that in the basic analyses presented below, the recurrence between two syncronized gaze patterns is measured, see figure 4 below. The parameter ϵ is a distance threshold. When the two gaze positions are within this threshold, they are counted as recurrent. This parameter was set to 70 mm, which roughly corresponds to the height of one cluster of blocks in the 'unused' area (there were 3 such clusters in all: long, half-long, and small blocks[4]). Finally, the time window size was fixed at 6 seconds — 3 seconds before, and 3 seconds after the onset of the word *en*. In this way, the amount of recurrence can be assessed at a delay of up to 3 seconds in either direction.

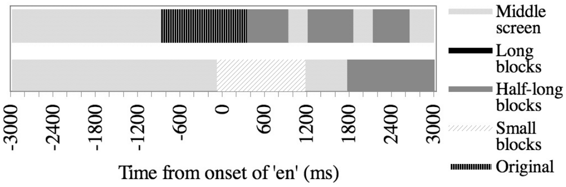


Figure 4: Two synchronized gaze pattern time series centered around the onset of the word *en*. The upper scarf plot shows the instructor's gaze pattern, and the lower one the constructor's. The referring expression used by the constructor in this case is *en mellanlång* 'a medium-long [one]'.

If one of the interlocutors revisits extended portions of the other interlocutor's gaze path, this behavior will result in extended diagonal lines in the CRP, since looking at the same positions in the same temporal order shows up as dots on a diagonal line. Two measures related to diagonal CRP lines were relevant: Mean length of all diagonal lines in the plot (denoted L), and length of the longest diagonal line in the plot (denoted L_{max}). L and L_{max} were measured for each gaze pattern pair associated with a referring expression (i.e. for each CRP) and averaged. The diagonal line length measures were relevant because it was expected that the more accessible referent type would be looked at faster after being mentioned, and be more likely than the less accessible referent type to already be in visual focus at the time of the onset of the referring expression. Thus, longer diagonal lines were expected in the unrestricted group, but it was not known at which time lag. Therefore, the CRQA methods are suitable, since they quantify these measures in the whole +/– 3 second time window. The raw percentage of recurrence (i.e. the percentage of black dots in the whole CRP, denoted *RR* for 'recurrence rate') was also measured. This measure gives an impression of how much the two interlocutors look at the same positions, not necessarily in the same order.

The eye movement measures were only carried out on two of the four dialogs. The other two unfortunately had to be discarded because of a calibration error.

---

[4] The other two areas were the original figure on the instructor's screen and the emerging copy in the middle of the constructor's screen.

Finally, one dialog was scrutinized more closely in order to get an impression of the eye-voice, voice-eye, and eye-eye spans in the eye movement data. The video editing program Final Cut Pro was used to analyze the data frame-by-frame.

3 Results

3.1 Linguistic data

The referring expressions used to refer to unused blocks were counted. The results are tabulated below.

	Indefinite		**Definite**		
	unrestr.	restr.	unrestr.	restr.	Total
Ellipsis	**10**	**31***	**4**	**3**	**48**
Full NPs	**0**	**20**	**0**	**1**	**21**
Total	**10**	**51**	**4**	**4**	**69**

*) 3 of these are uttered by the constructor

Table 1: Singular referring expressions used to refer to unused blocks in dialog 1a, 1b, 2a, and 2b. (All but three of them were uttered by the instructor.)

The first thing that one notices is that there are more indefinite than definite expressions. This is not surprising, given that unused entities are usually referred to using indefinite noun phrases (Prince, 1981).

There are more elliptical expressions than full noun phrases. Since all the entities talked about are blocks, it is not informative to include nouns like ‘block’ or the like in the expressions. Many of the elliptical expressions have restrictive modifiers, however, and this makes them less anaphoric than the unrestricted ones.

The unrestricted indefinite elliptical expressions from two dialogs were compared to the restricted ones in terms of eye movement recurrence (see table 2). This is the focus of the next section.

Elliptical indefinites		
Unrestricted	Restricted	Total
9	**25**	**34**

Table 2: The small subset of data analyzed for eye movement recurrence. The data are from dialogs 2a and 2b. The other two dialogs had to be excluded because of corrupted eye movement data.

3.2 Eye movement data

A frame-by-frame analysis of the elliptical indefinites in dialog 2b was carried out. It turned out that the speaker did not always fixate the target category (i.e. the next block type to be mentioned) on the mirror of the constructor’s screen prior to uttering an instruction, but equally often looked at the original figure for information on the next block. This is of course not too surprising after all, since the task was to copy an existing object, not building a new one. In fact, all but one of the 20 instructions given by this particular instructor were preceded by a gaze on the original. All 20 gaze patterns had a very similar structure. Typically, a few gazes back and forth between the original figure and the copy were followed by a gaze at the target block type, sometimes followed by gazes back and forth between the copy and the target block. The onset of the looks at the target were located around time 0, i.e. the onset of *en* ‘a/one’, ranging from –920 ms to 1240 ms. On average, target gazes associated with unrestricted expressions started a bit earlier (–80 ms) than gazes associated with restricted expressions (–23 ms). Data were too sparse to verify if this difference was significant.

It was clear that the onset of the first target gaze of the speaker should not serve as the point of departure of eye-voice measures, since this first gaze often followed after the onset of *en*, and in fact sometimes seemed to be triggered by the constructor manipulating the target rather than by any language production processes, as shown in figure 5.

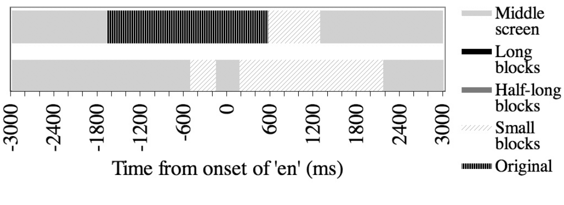


Figure 5: The instructor (upper scarf) looks at the target after, and perhaps triggered by, the constructor (lower scarf). The referring expression used by the constructor in this case is *en liten till* ‘a small [one] more’.

Instead, the eye-voice span was measured from the onset of the last gaze on the original figure before time 0. The voice-eye span was measured from 0 to the onset of the first gaze on the target after 0. If a target gaze started before 0 and crossed it, a value of 0 was recorded[5]. Combined, these two figures yield the 'eye-eye spans' in the data, see table 3.

	Eye-voice	**Voice-eye**	**Eye-eye**
Unrestr.	1280 ms	376 ms	1656 ms
Restr.	1643 ms	557 ms	2200 ms

Table 3: Eye-eye spans for dialog 2b.

These data closely parallel those of Hadelich and Crocker (2006), and the unrestricted measure is almost identical to Richardson and Dale's (2005) point of maximal recurrence.

It was measured how much attention was paid to targets on average. The constructor spent almost exactly equally much attention on the target areas in unrestricted (1000 ms) and restricted cases (1011 ms). The listener spent 1888 ms gazing at unrestricted targets, and 2489 ms on restricted targets.

Thus, the constructor in this particular dialog on average looked at the target earlier, and fixated it for a shorter time in connection with unrestricted expressions.

On the assumption that the speaker looks approximately equally much at the intended referent type in both groups, a difference in gaze patterns might materialize because the listener could already have the intended referent type in focus in the case of unrestricted anaphora (indeed, this could be the speaker's principal reason for using a fully anaphoric form), but less so in the case of restricted anaphora. Put differently, unrestricted anaphora might be more accessible to the listener, yielding the shorter voice-eye span observed. This alone would not necessarily affect the recurrence rate, however. If only the voice-eye lags differ, overall recurrence might not differ. On the other hand, if listeners look at the intended referent for different durations depending on the type of expression, as suggested by the analysis of dialog 2b above, then the overall recurrence rate (*RR*) would be affected. The *RR* increases with the similarity of the distribution of attention on the different areas of interest in the two gaze patterns, regardless of the order of the gazes. The speaker in dialog 2b spent equal amounts of attention on the targets in the two groups. The amount of attention spent on the unrestricted expression referents by the listener was lower than the amount spent on the restricted referents, and thus more similar to the speaker's amount of attention on the target referents. If this is a general pattern, a higher *RR* would be expected in the unrestricted case.

The *RR* results confirm this hypothesis. There is a significant difference in the expected direction in overall recurrence (t(32) = 1.76, $p < 0.05$ one-tailed). See figure 6.

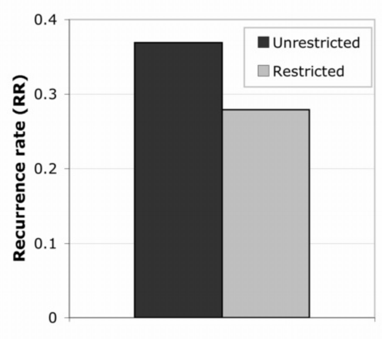


Figure 6: *RR* of the two types of referring expressions. The difference is significant ($p < 0.05$).

This result indicates that the two interlocutors look more at the same positions overall during the 6 second time window when a fully anaphoric expression is used. The frame-by-frame analysis suggests that this comes about because the more anaphoric (i.e. unrestricted), and presumably more accessible forms require less attention from the listener.

The recurrence rate gives a measure of how much the two interlocutors look at the same things, but it is not sensitive to the order of the fixations. The measures of the diagonal lines in the CRPs remedy this. A diagonal line means that the same areas are fixated a number of data points in a row by the two interlocutors, possibly at a lag. If the interlocutors look at the same areas at the same *time*, the line appears on the x=y diagonal (called the *line of incidence*, or LOI).

---

[5] Others have analysed this kind of data from a point of departure 200 ms after 0 in order to only consider eye movements that could plausibly have been driven by the linguistic form considered (Chambers et al., 2002). Here, however, target fixations seem to occur only in connection with the utterance of relevant instructions, and predictive eye movements are included because of the very possibility of pronouns being uttered in contexts where the listener might be assumed to already have the intended referent 'in focus'.

Since the frame-by-frame analysis showed that the unrestricted target gazes of the listener were more similar than the restricted ones to those of the speaker, it was expected that the gaze patterns of speaker and listener would be similar for longer stretches of time in the unrestricted group, yielding longer diagonal lines in the CRP.

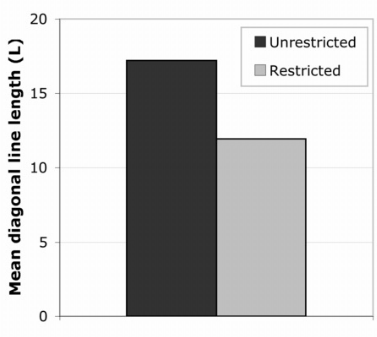


Figure 7: L of the two types of referring expressions. The difference is significant (p $<$ 0.01).

The results for L and L_{max} both show significant differences in the expected direction (L: t(32) = 2.88, p $<$ 0.01 one-tailed; L_{max}: t(32) = 3.06, p $<$ 0.01 one-tailed). See figures 7 and 8.

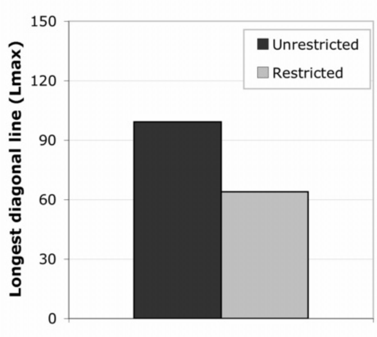


Figure 8: L_{max} of the two types of referring expressions. The difference is significant (p $<$ 0.01).

These results suggest that the gaze positions of the instructor and the constructor recur for longer stretches in connection with the unrestricted expressions, just as expected.

4 Discussion

The results presented above are very promising, in that it has been shown on the basis of very limited data that there are subtle differences in eye movement trajectory recurrence depending on the use of minimally different types of referring expressions. This is consistent with the idea of fluctuations in the moment-by-moment activation of concepts in common ground motivating linguistic forms of varying anaphoricity. The results constitute the next step towards the documentation of well-known and widely accepted assumptions about anaphor processing in ongoing discourse.

It was shown that more anaphoric indefinite expressions lead to more recurrence overall, and longer uninterrupted stretches of tandem eye movement patterns.

This pattern of results generally supports the hypothesis that accessibility licences anaphoricity, possibly in a graded manner. The fact that these results come from a study of different types of indefinite forms suggests that accessibility may indeed be independent from identifiability, as sketched in figure 1 above.

To be fair, the threshold ϵ, set to 70 mm, is relatively large. It comprises almost one fourth of the screen height. This means that the two gaze points can be quite far apart and still be counted as recurrent. On the other hand, the threshold corresponds to the height of each cluster of unused block types, so in order to capture all gaze pairs that fall within the same category of blocks, a threshold of this size is necessary. Therefore, this threshold size is justified.

Relevant future work will obviously be to carry out a full-scale version of this experiment, and to create new versions of it specifically designed to target other expression types (e.g. singular definite referring expressions).

In conclusion, the CRQA methods show considerable promise as a toolbox for the quantitative study of ongoing anaphor use in relatively natural conversation.

5 Acknowledgements

Thanks to Kenneth Holmqvist, Anders Sjöström, Johan Dahl, Marcus Nyström, Richard Andersson, and Joost van de Weijer of the Humanities Lab at Lund University, and all the (other) participants in the study for their time and efforts to help.

References

Allopenna, P. D., Magnuson, J. S. & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*, 419–439.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London and New York: Routledge.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H. & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*, 30–47.

Dahl, D. (1985). *The structure and function of one-anaphora in English*. Bloomington, Indiana: Indiana University Linguistics Club.

Griffin, Z. & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11(4)*, 274–279.

Gullberg, M., Morén, J. & Stenfors, I. (1997). The verbal transmission of visual information: An experimental study. *Lund University Linguistics Department Working Papers* (pp. 113–131).

Gundel, J. K., Hedberg, N. & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307.

Hadelich, K. & Crocker, M. W. (2006). Gaze alignment of interlocutors in conversational dialogues. In *CUNY 2006*.

Kamide, Y., Scheepers, C. & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from german and english. *Journal of psycholinguistic research*, *32(1)*, 37–55.

Lambrecht, K. (1994). *Information structure and sentence form: topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Marwan, N. & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*, 299–307.

Marwan, N., Thiel, M. & Nowaczyk, N. (2002). Cross recurrence plot based synchronization of time series. *Nonlinear processes in Geophysics* (pp. 325–331).

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics*. New York: Academic Press.

Richardson, D. C. & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*, 39–54.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Tanenhaus, M. K. & Trueswell, J. (2005). Eye movements as a tool for bridging the language-as-product and language-as-action traditions. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, Learning, Development, and Conceptual Change chapter 1, (pp. 3–37). Cambridge, MA, USA: MIT Press.

Trueswell, J. C. & Tanenhaus, M. K. (Eds.). (2005). *Approaches to studying world-situated language use. Bridging the language-as-product and language-as-action traditions*. Learning, Development, and Conceptual Change. Cambridge, Massachusetts: The MIT Press.

Walker, M. A. & Prince, E. F. (1993). A bilateral approach to givenness: A hearer-status and a centering algorithm.

Multi-modal Integration for Gesture and Speech

Andy Lücking **Hannes Rieser** **Marc Staudacher**
Bielefeld University, CRC 360 “Situated Artifical Communicators”, B3
{andy.luecking,hannes.rieser,marc.staudacher}@uni-bielefeld.de

Abstract

Demonstratives, in particular gestures that “only” accompany speech, are not a big issue in current theories of grammar. If we deal with gestures, fixing their function is one big problem, the other one is how to integrate the representations originating from different channels and, ultimately, how to determine their composite meanings. The growing interest in multi-modal settings, computer simulations, human-machine interfaces and VR-applications increases the need for theories of multi-modal structures and events. In our workshop-contribution we focus on the integration of multi-modal contents and investigate different approaches dealing with this problem such as Johnston et al. (1997) and Johnston (1998), Johnston and Bangalore (2000), Chierchia (1995), Asher (2005), and Rieser (2005).

1 Introduction

In this paper we are concerned with the multi-modal integration of pointing gestures (called *gestures* hereafter) and speech. Gestures can be used to refer to objects present in the actual situation like apples or tables. It is also possible to point at objects not present in the actual situation as when giving directions or placing discourse referents into the gesture space (see McNeill, 1992). We confine ourselves to the former and provide crucial data for speech-gesture-integration below. We take these data as evidence for the claim that gestures are essentially linguistic.

A striking characteristic of the speech-gesture-interplay is that demonstratives (determiners, exophoric pronouns and place adverbs) *require* a gesture to co-occur with them. We represent a gesture’s stroke with the symbol ‘↘’, statements of acceptability are displayed as in (1) where ‘#’ stands for “not acceptable”.

(1) a. Grasp ↘ this bolt!
b. #Grasp this bolt!

Example (1-a) is well-formed while (1-b) is not, since the gesture is missing. In a related construction (replacing ‘this’ by ‘the’), the use of a gesture is not required as the pair (2-a) and (2-b) shows.

(2) a. Grasp the bolt (on the table).
b. Grasp ↘ the bolt (on the table).

A feature left implicit in the format chosen to represent gestures and their co-present speech in (1) are the temporal relationships between them. Tokens of words and gesture can overlap in various ways. If we use a linear string representation of both words and a gesture’s stroke with a precedence reading, different possible stroke positions give rise to different acceptability judgements. In other words, synchronisation matters. In case the stroke starts and ends before the onset of the accompanying utterance, as in (3-a), the multi-modal utterance has to be dismissed as being not acceptable. The same holds for strokes altogether following their affiliated speech, as in (3-d).[1] We conclude from these data that gestures have syntactic properties.

(3) a. #↘ Grasp this bolt!
b. Grasp ↘ this bolt!
c. Grasp this ↘ bolt!
d. #Grasp this bolt ↘!

Gestures also have semantic properties as the following example shows. Suppose a situation *s* where two candies, a red one and a green one, are lying side by side. Whether an utterance of ‘This ↘ candy is red.’ is evaluated as true or false in *s* depends on which candy is pointed at in *s*. Besides truth conditional effects, there is empirical evidence that gestures have *rich information content*. Lücking et al. (2004) found that the number of words used in a verbal description was less if the description was accompanied by a deictic gesture. Thus the finding suggests that gestures contribute content that otherwise would have to be expressed verbally.

Moreover, gestures relate to pragmatic phenom-

---

[1] However, we are able to interpret such utterances – presumably by pragmatic, i.e. inferential processes.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 106–113.

ena. For example, it is not possible to substitute a verbal constituent for a deictic gesture in a *null context*, as in (4):[2]

(4) #He grasps ↘.

Note that example (4) can be rendered acceptable if a suitable object can be *accommodated*. Such a multi-modal utterance is also acceptable if it is uttered in a suitable context. For example, suppose a combat of gladiators in a Roman arena. The emperor decides whether they will live or die by pointing at them and (presumably) uttering (5-a) or (5-b), respectively. Given the supposed context, the utterances are acceptable.

(5) a. ↘ *missum*! (off he go!)
b. ↘ *iugula*! (cut his throat!)

In dialogues, a gesture can be used to realize a dialogue move. In (6) a piece of conversation between A and B is given, where B's gestural answer is acceptable. Its acceptability seems to be parasitic on the structure of question-answer-pairs and Gricean maxims.

(6) A: Where is the salt?
B: ↘

So, since gestures have syntactic, semantic, and pragmatic properties, they are just like words.

2 Interface Problems

If gestures are essentially linguistic, (formal) linguistic theories should account for them. From this point of view, current theories have a descriptive and explanatory gap and are in this sense deficient. Consequently, something new has to be taken account of. How shall we theorize? – In this section we discuss some *interface problems*.

The first point, however, relates to *theory change positions*. The question is whether a new kind of theory is required or an existing theory should be extended. Different answers are possible.

The *syntax enhancer* proposes a multi-modal theory not differing substantially form current ones. The enhancer thinks that syntax should be changed in such a way that gestures are accounted for, and then looks for changes in semantics and pragmatics.

In opposition, the *syntax radical* proposes to develop a new kind of a multi-modal theory differing from the current ones in a substantial way. Properties of current theories need not be preserved. The radical thinks current syntax should be replaced by a new kind which can account for gestures from the outset and then looks for an apt semantics and pragmatics.

The *pragmasemantics enhancer* has the same attitude towards a multi-modal theory as the syntax enhancer has. However, he thinks that gestures should be accounted for in semantics and pragmatics as opposed to syntax. The enhancers seem to be more conservative than the radical.

Each position has its price. By regarding gestures as linguistic, we change our existing concepts, notably some overly restrictive *concept of meaning*. Meaning is then no longer that which is or can be said but something else. A gesture cannot properly be "said". However, it seems that the richer concept of meaning still shares many properties with the traditional one.

The first problem relates to a consequence of the different positions, namely to the *point of integration*. The syntacticians propose either to approximate gestures to some existing linguistic category or to propose a new one for which combination rules are stated. The pragmasemanticists, on the other hand, will say that gestures are part of the linguistic context which is used to interpret an utterance. So, integration is some kind of syntactic combination (e.g. multi-modal subcategorization) and/or context-dependence.

The second problem relates to *descriptive adequacy*. There is an important difference between describing mono-modal information and multi-modal data. The former, e.g. speech, has a temporal order in such a way that for every two information bits one precedes the other. There are no overlap- or part-of-relations. In contrast, the information bits of the latter allow for such relations since the data is distributed across the different channels (such as sound and vision). Should an adequate description of multi-modal data take care of this? – This depends on the description's aim. For example, if an agent system is developed, multi-modal output planning might be important. Then questions of timing matter and, arguably, time should be explicitly represented. If the aim is doing semantics, however, only as much description is required as to describe the correct

---

[2] However, some readers might have different intuitions. We would like to point out that its acceptability might be due to the valence of the transitive verb predicting an argument at the level of its logical form which might be linked to ↘.

satisfaction conditions. In this paper we take a linguistic view on the matter and have chosen a linear representation.

A related problem is the *linearization problem*. The question is whether all data descriptions have to linearize in the sense that the information bits in the representation have to be in a linear order. It seems to us that, when doing semantics in a type-logical-style, the data must always be linearized since every information bit in the representation can only be either a functor or an argument.The point is illustrated by the algebraic set-up of multimodal integration below (tr is a translation mapping speech S and gesture G data to a type-logical intermediate-language IL. IL is interpreted by I in the semantics M.):

$$\begin{tikzcd}
S \arrow[dr] & & \\
& IL \arrow[r, "I"'] & M \\
G \arrow[ur, "tr"] & &
\end{tikzcd}$$

The fourth problem relates to *constructability*. It consists in providing a construction mechanism for logical forms of multi-modal utterances. If we have semantic aims, we want to have a systematic means to extract the right forms from multimodal utterances. It should be possible to construct the intended logical form. Though, depending on the theory change position, what is needed can be quite different.

3 Approaches to integration

Having covered the ground for a review, we quickly chart out the proposals. A summary of the approaches is presented in Table 1.

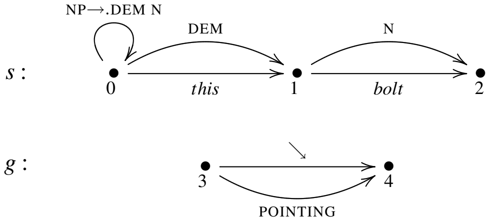
3.1 Johnston et al. HPSG (1997, 1998)

In course of the (military) software engineering project QuickSet, Johnston et al. (1997) developed an architecture that integrates input coming on different channels by means of unification of typed feature structures. The system provides a multimodal interface allowing its user to give directives simultaneously by voice (speech) and pen (gesture) input. Both speech and gesture are assigned attribute-value matrix (AVM) representations by speech and gesture recognizers. Since a conventional HPSG grammar is merely extended to account for multi-modal utterances, it is an approach of a syntax enhancer.

Users interacting with the QuickSet system can point (at X) and by doing so they introduce a certain point in space represented as a latitude-longitude coordinate pair. This locational function of pointings is captured in the following representation showing that the semantics (content) of an object of category (cat) *spatial_gesture* is a definite point in space:

$$\begin{bmatrix} \text{cat}: & \textit{spatial\_gesture} \\ \text{content}: & \begin{bmatrix} \text{fsType}: \textit{point} \\ \text{coord}: \; \textit{latlong}(\text{x,y}) \end{bmatrix} \end{bmatrix}$$

The AVM-grammar formalism rests on a *multimodal chart parser*. A multi-modal chart extends a conventional chart in that the former covers channel-crossing edges defined in terms of sets of identifiers of gestural (g) and speech (s) terminals:



Possible multicharts:

multichart 1: {[s,0,1], [g,3,4]}
multichart 2: {[s,1,2], [g,3,4]}
...

The basic rule allowing to "bridge" between the modalities is the *basic integration scheme*:

$$\begin{bmatrix}
\text{lhs}: & \begin{bmatrix} \text{cat}: & \textit{comm} \\ \text{modality}: & \boxed{2} \\ \text{content}: & \boxed{1} \\ \text{time}: & \boxed{3} \end{bmatrix} \\
\text{rhs}: & \begin{bmatrix} \text{dtr1}: & \begin{bmatrix} \text{cat}: & \textit{loc\_comm} \\ \text{modality}: & \boxed{6} \\ \text{content}: & \boxed{1}[\text{loc } \boxed{5}] \\ \text{time}: & \boxed{7} \end{bmatrix} \\ \text{dtr2}: & \begin{bmatrix} \text{cat}: & \textit{spat\_gest} \\ \text{content}: & \boxed{5} \\ \text{modality}: & \boxed{9} \\ \text{time}: & \boxed{10} \end{bmatrix} \end{bmatrix} \\
\text{cnstr}: & \begin{Bmatrix} \text{overlap}(\boxed{7},\boxed{10}) \vee \text{follow}(\boxed{7},\boxed{10},4) \\ \text{total-time}(\boxed{7},\boxed{10},\boxed{3}) \\ \text{assign-modality}(\boxed{6},\boxed{9},\boxed{2}) \end{Bmatrix}
\end{bmatrix}$$

The AVM for the integration scheme is stated very closely to a CFG-rule of the form lhs → rhs; the right-hand side (rhs) is made up of two constituents, namely dtr1 and dtr2. Thus, mapping the rule to a tree, they are the daughters of their mother constituent on the rule's left-hand side (lhs). The rhs-part of the AVM-structure is made up of a verbal located command (*loc_comm*; in QuickSet this can be, *e. g.*, "sandbag wall") and a spatial gesture. The gesture determines the location value

	Johnston HPSG et al. (1997, 1998)	**Johnston and Bangalore FSM (2000)**	**Chierchia (1997)**	**Asher SDRT (2005)**	**Rieser LTAG (2005)**
Motivation	Human Computer Interaction	Human Computer Interaction	Anaphora, context-dependent Quantifiers	Anaphora, anchoring of deictic NPs	Extended use of language, meaning, gestures as signs
Type of theory	Syntax, semantic representation	Syntax, semantic representation	Semantics, formal pragmatics	Semantics, formal pragmatics	Syntax, semantics, pragmatics
Type of grammar	Constraint-based (HPSG)	CFG	–	–	LTAG
Pointing representation	AVM-structure for locations	Object constants	Pragmatic indices	Externally anchored discourse referent	Set of object constants
Point of integration	Pointing introduced via subcategorization	Nouns, translation to semantics	Semantically underspecified quantifier representation	Presuppositional SDRS, underspecified discourse relation	Extended valence of relations in multi-modal interface
Strengths	Multi-modal chart parser	Highly efficient FSM parser	local extension of existing theory	Pointing in dialogue	Speech and gesture interaction in interface
Weaknesses	Restricted grammar, little linguistic motivation	No linguistic motivation, hardly extendible	No syntactic gesture representation	No syntactic gesture representation	No general mechanism to build interface from standard grammar

Table 1: Cluster of Approaches to Multi-modal Integration

of the word's content. The mother structure (lhs) then is a complete multi-modal command. The cross-channel integration is constrained by a set of restrictions given as the value of the feature cnstr (short for constraints). Most notably, *co-occurrence constraints* are expressed as temporal requirements, see the use of tags $\boxed{7}$ and $\boxed{10}$.

As it stands, the basic integration scheme licenses only multi-modal structures that consist of a speech portion and exactly one accompanying gesture. A more general framework that corrects this limitation is the extension given by Johnston (1998) where integration is handled via multi-modal subcategorization, analogous to the (lexicalist) treatment of complementation in HPSG. To this end, a new feature sbct (for subcat) is introduced whose elements can be recursively discharged by a *subcat combination scheme*, a generalized version of the basic integration scheme. Leaving the restricted QuickSet grammar but still remaining in the spirit of the grammar of Johnston and colleagues, the feature sbct can be used to capture that demonstratives require a co-occuring pointing gesture. A determiner like "this" is incomplete, that is, being of category *sub_dem*, unless it combines with the subcategorized gesture to build a proper AVM of category *dem*, as licensed by the subcat combination scheme. Applied to the example sentence "*Grasp this* ↘ *bolt!*", the "*this* ↘ "-part gets modelled as follows:

$$\left[\begin{array}{l}
\text{lhs}: \left[\begin{array}{l}\text{cat}:\ \textit{dem}\\ \text{cont}: \boxed{3}\end{array}\right]\\
\text{rhs}: \left[\begin{array}{l}
\text{dtr1}: \left[\begin{array}{l}
\text{cat}:\ \textit{sub\_dem}\\
\text{cont}: \boxed{3}\left[\begin{array}{l}\text{obj}: \left[\text{fsT}: \textit{exist\_there}\right]\\ \text{loc}: \left[\begin{array}{l}\text{fsT}:\ \textit{point}\\ \text{coord}: \boxed{1}\end{array}\right]\end{array}\right]\\
\text{sbct}: \boxed{2}\left[\begin{array}{l}\text{cat}:\ \textit{spatial\_gesture}\\ \text{cont}: \left[\begin{array}{l}\text{fsT}:\ \textit{point}\\ \text{coord}: \boxed{1}\end{array}\right]\end{array}\right]
\end{array}\right]\\
\text{dtr2}: \boxed{2}\left[\text{cont}: \left[\text{coord}: \boxed{1}\right]\right]
\end{array}\right]
\end{array}\right]$$

The demonstrative has a locational "gap" that gets filled by the gesture it subcategorizes for. The complete sentence is then projected as usual.

3.2 Johnston and Bangalore FSM (2000)

Johnston and Bangalore (2000) propose a multi-modal context-free grammar (CFG) to handle integration. Their parser implementation uses well-understood finite-state techniques.[3] Moreover, the translation to logical form is a product of concatenation. Thus, it is simple and highly efficient. As will become apparent, the position taken by this approach is the one of a syntax radical.

The multi-modal input, speech and gesture, is assumed to be distributed across different chan-

---

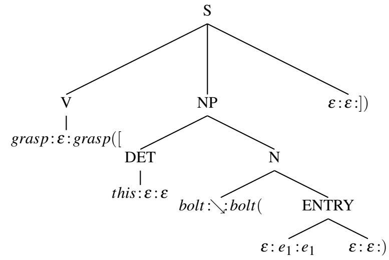
[3] Thus only the regular part of the CFG may be used.

nels. In order to use (mono-modal) context-free techniques, the reading of symbols in different channels is regarded as a single read operation of a complex, structured symbol. Each part of the structure relates to a symbol in a channel.

A *multi-modal CFG* is a tuple $\langle N,T,P,S\rangle$ where N is the set of *nonterminals* such as S, V or NP. T is the set of *terminals* of the form $W\!:\!G\!:\!M$ where W denotes a verbal symbol (e.g. 'bolt'), G denotes a gesture symbol (e.g. '$\searrow$') and M denotes a meaning expression, *e. g.* '$bolt($'.[4] All symbols of W, G and M are elements of the symbol alphabets Σ_W, Σ_G and Σ_M, respectively. Each alphabet contains the empty element ε. A special feature of the gesture alphabet Σ_G is that it consists of two disjunct subsets: One subset contains all gesture symbols, the other one contains all event symbols. As usual, P is the set of *productions* and S denotes the start symbol. Other notions such as *derivation* are the standard ones for CFG.

Gesture symbols and event symbols have different roles. An occurrence of the former indicates the presence of a gesture whereas the occurrence of the latter is used as a reference to entities referred to by a gesture. These event symbols label buffer elements of a finite buffer. The buffer is used to keep track of all assignments between gesture occurrences and the entities referred to by those gestures.

Using a modification of the sample grammar provided by Johnston and Bangalore (2000), the structure of (1-a) is:



The ENTRY node triggers the buffer mechanism which assigns e_1 a name of the entity referred to by the gesture occurrence, *e. g.* $obj1$. The meaning string $grasp([bolt(e_1)])$ is constructed by a top-down/left-to-right traversal through the tree. It is the result of the concatenation of the M-parts of every traversed terminal.

We think that their proposal is highly interesting to produce efficient parsers. However, it doesn't seem to be a good way to write *linguistically motivated grammars*. For example, sentences S also contain a verbal ε symbol. N is assigned a branching structure with ENTRY as its right node, and so on. Due to these *ad hoc* structures, the grammar is not easily extendible. The translation to logical forms is weird. There is no basic translation for 'this'. Basic translations for most syntactic expressions are not well-formed semantic expressions, *e. g.* '$bolt($' or '$)$'. Thus it fails to be admissible on any standard account of semantic translation. For the same reason, an incremental interpretation is not possible.[5] Last but not least, using the "tape"-metaphor of automata theory, accumulation of gesture symbols on the tape leads to difficulties. Consider the unacceptable example (3-a). Since the gesture symbol is already on the tape, the N rule can be applied and thereby (3-a) is licensed by the grammar for (1-a).

3.3 Chierchia (1995)

Chierchia sketches a way to handle multi-modal integration in his renowned book *Dynamics of meaning* (Chierchia, 1995). He proposes to modify the translation of definites of the form *the N* to account for indexically used definites. Since he uses no syntactic representation for gestures and extends grammar conservatively, he is a pragmasemantics enhancer.

He locates the place where one should modify in the *representation of definites*.[6] In a first take he views a definite as a (partial) function from properties (represented as sentential functions) to the unique object that satisfies them, if there is such an object. He does so by introducing them formally as iota-terms of the form $\iota x\phi$ with the following semantics: If x is of type e and ϕ is of type t, then $[\![\iota x\phi]\!]^g = u$, where u is the unique object such that $[\![\phi]\!]^{g[x/u]}$, otherwise $\perp$ (read as "undefined").

However, a sentence like 'You, grasp the bolt.' being translated as $grasp(you, \iota x bolt(x))$ has an infelicitous use when the iota-term property is not satisfied. Chierchia's remedy is to analyse such utterances as utterances of the sentence 'You, grasp the bolt pointed at.' or 'You, grasp the bolt we are looking at.' which are translated as $grasp(you, \iota x(R(o,y,x) \wedge bolt(x)))$ and $grasp(you, \iota x(R(y,x) \wedge bolt(x)))$, respectively. In the translations o designates a location and y the speaker. The predicate R is interpreted as *is*

---

[4] The meaning expression is indeed written as `b-o-l-t-left parenthesis`.

[5] Arguably, it is possible using Lambda-terms.

[6] Note that the presentation of Chierchia's approach is simplified in that no possible world semantics is used.

pointed at ... by in the first translation and as *is looking at* in the second one.

Chierchia generalizes the translation of definites of the form *the N* to $\iota x(R(y_1, y_2, \ldots, y_n, x) \wedge N(x))$ where $R, y_1, y_2, \ldots, y_n$ are free variables. The context has to assign R an n-place function from the values of the pragmatic indices $y_1, y_2, \ldots, y_n$ to N's denotation. The indices $y_1, y_2, \ldots, y_n$ are taken to be part of the logical form.

So, the propositional part of (1-a) is analysed as $grasp(you, \iota x(R(o,y,x) \wedge bolt(x)))$ in a context where R, y and o are as before.

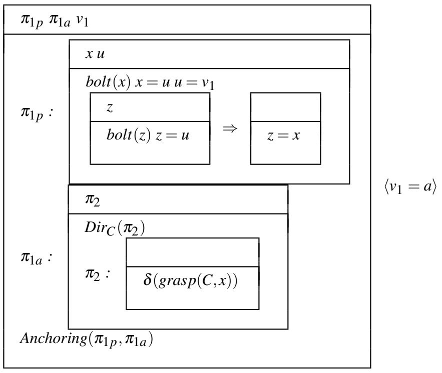
Chierchia's proposal is interesting, since it is a conservative extension of existing theories. The change is local and restricted to the logical form of definites. Though, does it amount to a satisfactory account of integration? We think not, since his proposal neglects the syntactic properties of gestures. They are not given a syntactic representation and only appear in the context. Moreover, there is no explicit integration mechanism. It is not clear how information given by gestures is used to construct the assignment for R, y and o.

3.4 Asher SDRT (2005)

Asher (2002, 2005)[7] sketches how SDRT can be extended to account for gestures. SDRT is a discourse representation theory modelling the semantics/pragmatics-interface. The theory itself is not committed to a particular grammar formalism and hence not to any specific syntax. Thus, the integration problem is approached by way of a pragmasemantic enhancer.

According to the SDRT account of definites, presuppositional information is distinguished from asserted information. The presupposed information of newly introduced definite NPs cannot simply be accommodated since an arbitrary object satisfying the conditions would not do for deictically used definites. It is proposed that definite descriptions introduce an underspecified relation, called *bridging relation*, between the referent and some other contextually given object, set to identity by default. In other words, such definites have to be *anchored* to some object in the non-linguistic context. Anchoring involves a *de re* attitude towards the object, some sort of *knowing how* needed to solve the conversational goals of the speaker. Anchoring requires linking an agent A's epistemic attitude to conversational goals. If an *Anchoring* relation between the presupposition of a definite ψ and some element in the discourse context exists for the agent A, he is supposed to have a computable means of getting to the referent of ψ from the present non-linguistic context of utterance under some given purpose ϕ; to capture this, a notion of *path* is defined. If the anchoring function of a deictically used definite is accepted by the participants in dialogue, they are assumed to mutually believe that the definite picks out the same object for them. Hence, anchoring amounts to coordination or alignment.

Applying Asher's new SDRT proposal to (1-a), the result of an apt multi-modal integration strategy under best-update is:



SDRT itself says nothing about multi-modal integration, though, it is part and parcel of it: The conceptual information of the gesture occurrence consists in the external anchoring of the discourse referent v_1 to the object a, written as $\langle v_1 = a \rangle$. The presupposed information of (1-a) is represented in π_{1p}, the asserted one in π_{1a}. The bridging relation between x and u is resolved to identity and thus $x = u$. We assume a speech act theory style imperative semantics. Consequently, Dir_C is to be read as 'C is commanded that ...' and $\delta(grasp(C,x))$ in π_2 is the action commanded, namely that agent C grasp x. Finally, the *Anchoring* relation holds between π_{1p} and π_{1a}. Thereby, x in π_2 is externally anchored to a.[8]

Asher's proposal is unique with regard to discourse modelling. However, gestures have no syntactic representation. It is not clear how the multi-modal input is integrated. While the DRT construction algorithm can be used in principle, the construction problem remains unsolved in practice. For SDRT provides, "out of the box", neither

---

[7] We are grateful to Nicholas Asher for having taught us SDRT in the years 2003-2005 and for letting us work with unpublished SDRT material, especially Asher (2005).

[8] Observe that *Anchoring* is a subordinating discourse relation. Thereby, x in π_{1p} is accessible to x in π_2.

an anphora resolution mechanism nor a construction algorithm for SDRSs.

3.5 Rieser LTAG (2005)

In Rieser (2005)'s LTAG approach integration of demonstrations is handled by a grammar based interface. If, from the point of view of function, demonstrations are considered as words of a special kind, acting in a way like names, the lexicon of the interface has to be extended. It will encompass demonstration forms. In a similar way, syntax rules have to be added which allow for the combination of pointing and verbal expressions. Finally, gestural and verbal meanings have to be integrated in a compositional way. Here, interface modelling is based on LTAG (Joshi, 2004). This approach counts as a syntax radical. For verb valences in the interface are different from the usual ones. As a consequence, the denotation of verbs is also different.

Extension of LTAG Syntax We need additional structure in order to accommodate $\searrow$ and its positions. The LTAG-format used works with a set of trees anchored by terminal elements and two rules, substitution and adjunction. Adjunction will not be used, it is mentioned here for reasons of generality. $\searrow$ is considered as a terminal.

The relevant LTAG fragment is displayed in Fig. 1. (a) is the subject-less imperative rule. Elementary trees (b) and (b$'$) do service for two distinct pointing positions. They express that pointing is needed. (c) says that pointings function in a sense like NPs. (d) and (e) follow canonical CFG-rules.

Syntax-semantics Integration In order to achieve the syntax-semantics-integration we decorate all terminals with appropriate type-logical formulas. β-conversion is needed in order to model compositionality.[9]

$$\begin{aligned}
\text{grasp!}: \quad & \lambda\Theta\lambda\Pi\lambda u(\Pi(\Theta(\lambda y\lambda v F_{dir}(grasp(u,v) \\
& \wedge (v=y)))))) \\
\text{bolt}: \quad & \lambda x bolt(x) \\
\searrow: \quad & \lambda P_{\searrow}.P_{\searrow}(a), \ldots \\
\text{this}: \quad & \lambda P\lambda Q.Q(\iota x(P(x)))
\end{aligned}$$

How do we arrive at the representation of *grasp!*? – The reasoning is as follows: The verb *grasp* needs two argument slots, since it is transitive. In the interface fusing together the definite description and the pointing which goes with it, we need an identity condition linking the object argument of grasp and the variable for the object pointed at. Since the pointing functions like an NP, we must use the formula interpreting *grasp!* in order to get the correct bindings.

Computation of Meaning for (1-a) As can be seen from (1)-(8) the account is compositional:

$$\begin{gathered}\lambda\Theta\lambda\Pi\lambda u(\Pi(\Theta(\lambda y\lambda v F_{dir}(grasp(u,v) \wedge (v=y))))) \\ \lambda P_{\searrow}.P_{\searrow}(a)\end{gathered} \tag{1}$$

$$\begin{gathered}\lambda\Pi\lambda u(\Pi(\lambda P_{\searrow}.P_{\searrow}(a) \\ (\lambda y\lambda v F_{dir}(grasp(u,v) \wedge (v=y)))))\end{gathered} \tag{2}$$

$$\lambda\Pi\lambda u(\Pi(((\lambda y\lambda v F_{dir}(grasp(u,v) \wedge (v=y)))(a)))) \tag{3}$$

$$\begin{gathered}(\lambda\Pi\lambda u(\Pi((\lambda v F_{dir}(grasp(u,v) \wedge (v=a)))))) \\ (\lambda P\lambda Q.Q(\iota x(P(x)))\lambda x bolt(x))\end{gathered} \tag{4}$$

$$\begin{gathered}(\lambda\Pi\lambda u(\Pi((\lambda v F_{dir}(grasp(u,v) \wedge (v=a)))))) \\ (\lambda Q.Q(\iota x(bolt(x))))\end{gathered} \tag{5}$$

$$\begin{gathered}(\lambda u((\lambda Q.Q(\iota x(bolt(x)))) \\ (\lambda v F_{dir}(grasp(u,v) \wedge (v=a)))))\end{gathered} \tag{6}$$

$$\begin{gathered}(\lambda u(((F_{dir}(grasp(u,(\iota x(bolt(x)))) \\ \wedge((\iota x(bolt(x))) = a))))\end{gathered} \tag{7}$$

$$F_{dir}(grasp(you,(\iota x(bolt(x)))) \wedge ((\iota x(bolt(x))) = a) \tag{8}$$

(8) is the result from β-reducing u in (7) with the indexical *you*.

4 Open Research Problems

The research on multi-modal integration (MMI) is still in its infancy. Therefore, basic empirical, methodological and theoretical issues have been hardly discussed. In this section we want to comment upon the following issues on the research agenda: Inverting the methodology for multimodal integration, motivation for a dynamic semantics approach, separation of presuppositional and assertional information, underspecification, restrictions of $\searrow$ under embedding, interaction of $\searrow$ with subsentential utterances, and interaction with iconic and emblematic gestures. These topics are treated in turn below.

As can be seen from the HPSG-approach and the LTAG proposal for MMI verbal expressions receive primary status on the modelling side and gesture is then added. One could also use the converse methodology, giving gesture primary relevance and adding language. It would drastically change the semantic role of gesture.

All the approaches discussed in this paper used static semantics, except Asher and Chierchia.

---

[9] In a way the strategy taken is similar to Lewis (1970) interpretation of PSG with formulas of intensional logic.

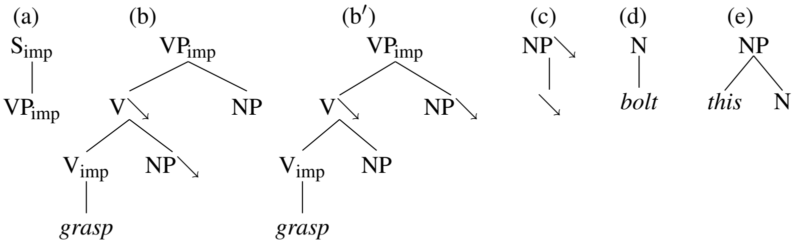


Figure 1: LTAG Fragment

However, there are good reasons for a dynamic semantics account. Such a dynamic account can also be fruitfully applied to complex demonstrations functioning as antecedents (7-a) or as anaphora (7-b):

(7) a. You may have ↘ this piece of cake. It tastes awful.
b. You are looking for a sweet? Take ↘ this strawberry tart.

What one can learn especially from the SDRT account is that presuppositional and assertional information should be separated. As a consequence, the use of rhetorical relations seems mandatory.

As example (3) shows, stroke positions can appear at various places in the utterance, stroke is, to borrow a linguistic term, polymorphic. A detailed reconstruction of this effect in grammars would have to result in a plethora of rules. Therefore an underspecification account of "stroke-syntax" seems to be more advisable. A similar argument goes for the pairing of stroke positions and functions of stroke. If position determines function, as seen from a type-logical perspective, we have many functions, which, however, are perhaps not distinct from the semantic point of view. This is hard to model.

Empirical and modelling problems arise, if strokes appear with deeply embedded material. This may give rise to ambiguities concerning attributions of stroke which is relevant with regard to truth conditional considerations. In the situated communication data used here, pointing comes frequently with subsentential utterances. A study of this effect using SDRT as the descriptive frame was started in (Lücking et al., 2006).

We know from our experimental studies that pointings tend towards "iconization". It is not clear as yet, how these effects should be modelled. One interesting aspect is how to represent iconic gestures and how to deal with compositionality matters in the interface of demonstration and iconicity.

Acknowledgments

Our work on SDRT was supported by the CRC "Situated Artificial Communicators," project "Deixis in Construction Dialogue" (DEIKON) at Bielefeld University, funded by the German Research Foundation (DFG). Thanks to the anonymous reviewers whose remarks were helpful for improving our paper.

References

Nicholas Asher. Deixis, Binding and Presupposition. Forthcoming in: Festschrift for Hans Kamp, 2002.

Nicholas Asher. Bielefeld Lectures on SDRT, 2005.

Gennaro Chierchia. *Dynamics of meaning: anaphora, presupposition and the theory of grammar*. University of Chicago Press, Chicago [a.o.], 1995.

Michael Johnston. Unification-based Multimodal Parsing. In *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics – Volume I*, pages 624–630, Montreal, Quebec, August 1998. ACL.

Michael Johnston and Srinivas Bangalore. Finite-state Multimodal Parsing and Understanding. In *Proceedings of the 18th Conference on Computational Linguistics – Volume I*, pages 369–375, Saarbrücken, July 2000. ACL.

Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman, and Ira Smith. Unification-based Multimodal Integration. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pages 281–288, Madrid, July 1997. ACL.

Aravind K. Joshi. Starting with complex primitives pays off: Complicate locally, simplify globally. *Cognitive Science*, 28(5):637–669, 2004.

David K. Lewis. General semantics. *Synthese*, 22:18–67, 1970.

Andy Lücking, Hannes Rieser, and Jens Stegmann. Statistical Support for the Study of Structures in Multi-Modal Dialogue. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04*, pages 56–63, Barcelona, 2004.

Andy Lücking, Hannes Rieser, and Marc Staudacher. SDRT and Multi-modal Situated Communication. ESSLLI, to appear, 2006.

David McNeill. *Hand and Mind—What Gestures Reveal about Thought*. Chicago University Press, Chicago, 1992.

Hannes Rieser. Pointing and grasping in concert. In Manfred Stede, Christian Chiarcos, Michael Grabski, and Luuk Lagerwerf, editors, *Salience in discourse: multidisciplinary approaches to discourse*, pages 129–139. Nodus Publikationen, Münster, 2005.

Acquiring words across generations: introspectively or interactively?

Zoran Macura
Department of Computer Science
King's College, London
The Strand, London WC2R 2LS
`zoran.macura@kcl.ac.uk`

Jonathan Ginzburg
Department of Computer Science
King's College, London
The Strand, London WC2R 2LS
`jonathan.ginzburg@kcl.ac.uk`

Abstract

How does a shared lexicon arise in population of agents with differing lexicons, and how can this shared lexicon be maintained over multiple generations? In order to get some insight into these questions we present an ALife model in which the lexicon dynamics of populations that possess and lack metacommunicative interaction (MCI) capabilities are compared. We ran a series of experiments on multi-generational populations whose initial state involved agents possessing distinct lexicons. These experiments reveal some clear differences in the lexicon dynamics of populations that acquire words solely by introspection contrasted with populations that learn using MCI or using a mixed strategy of introspection and MCI. The lexicon diverges at a faster rate for an introspective population, eventually collapsing to one single form which is associated with all meanings. This contrasts sharply with MCI capable populations in which a lexicon is maintained, where every meaning is associated with a unique word. We also investigated the effect of increasing the meaning space and showed that it speeds up the lexicon divergence for all populations irrespective of their acquisition method.

1 Introduction

A key feature of natural language is metacommunicative interaction (MCI)—utterance acts in which conversationalists acknowledge understanding or request clarification. The need to verify that mutual understanding among interlocutors has been achieved with respect to any given utterance—and engage in discussion of a clarification request if this is not the case—is one of the central organising principles of conversation (Clark, 1996). However, hitherto there has been little work on the emergence and significance of MCI meaning.

What significance does MCI have for linguistic interaction within a community? Pretheoretically, they serve as a device for ensuring a certain state of equilibrium or lack of divergence gets maintained within a linguistic community. The plausibility of this speculation can be assessed by converting it into more concrete questions such as the following:

(1) a. Given a community A where clarification requests do not get expressed, and community B where they do, how do the two communities evolve with respect to vocabulary drift.

b. How does this vocabulary drift change once a gradual turnover of community members is introduced?

In previous work we have shown how language converges for different types of populations in a mono-generational model (Ginzburg and Macura, in press). We also compared the performance of mono-generational and multi-generational populations and showed how the introduction of infants and mortality in the model affects the lexicon dynamics (Macura and Ginzburg, in press). In this paper we take a closer look at multi-generational populations, in particular the effect of varying meaning space—number of different plants in the environment—on the results.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 114–121.

In the next section we describe the computational model, including how gradual turnover of agents is implemented. In Section 3 we present the experiments and assess the validity of the proposed model. Finally, in Section 4, we draw some conclusions.

2 The Model

In our previous work we have shown how language converges for different types of populations within a single generation (Ginzburg and Macura, in press). In this type of model there is no generational turnover of agents and the transmission of language is horizontal, whereas the communication is between adult agents of the same generation (e.g. Steels (1998)). In multi-generational models such as the iterated learning model (e.g. Kirby et al. (2004); Smith (2005)) language is vertically transmitted from one generation to the next, where the adult agents are allowed to speak to the child agents only. So in these models there is no horizontal communication (i.e. between adults of the same generation).

We present a model which implements both horizontal (adult-adult) and vertical (adult-child) language transmission (see Vogt (2005) for a similar approach). The model contains an ALife environment in which the lexicon dynamics of populations that possess and lack MCI capabilities are compared. The environment is modelled loosely after the Sugarscape environment (Epstein and Axtell, 1996), in that it is a spatial grid containing different plants. Plants can be perceived and disambiguated by the agents. Agents walk randomly in the environment and when proximate to one another engage in a brief conversational interaction concerning plants visible to the agents.[1]

In the next section we look at the communication protocol in more detail, followed by a closer look at the implementation of generational turnover.

2.1 Communication

Agents can talk about the plants in the environment by making syntactically simple utterances—essentially one consisting of a single word. Every agent has an internal lexicon which is represented by an association matrix (see Smith (2005) for a similar approach). The lexicon stores the association scores for every meaning–representation pair (i.e. plant–word) based on individual past experiences. Agents don't have an invention capability therefore are only able to talk about the plants that they have a representation for.

Communication is a two sided process involving an intrinsic asymmetry between speaker and addressee: when talking about a plant in his field of vision, the speaking agent necessarily has a lexical representation of the plant (a word with the highest association score for the plant chosen as the topic), which he sends to the hearing agent. There is no necessity, however, that the addressee agent is able to interpret this utterance. If unable to do so (meaning that the hearing agent doesn't have the word in her lexicon, or that the plant it associates with the word is not in her context) the way that the agent tries to ground it depends on the agent's type.

Three types of communicative agents exist in the model; agents capable of making a clarification request (CR agents), agents incapable of doing so (introspective agents), and hybrid agents that use both CRs and introspection.

An introspective agent learns the meanings of words through disambiguation across multiple contexts. Upon hearing a word the agent looks around her and for every plant in her context (field of vision) she increases its association score with the word heard. This strategy is akin to the cross-situational statistical learning strategy used by inferential agents in Smith (2005), and to selfish learners in Vogt and Coumans (2003).

A CR agent on the other hand can resort to a clarification request upon hearing a word. If hearing the word for the first time (no associations with the word in her lexicon) or if there are no plants in her context, a clarification request is raised. Otherwise the agent checks the plants in her context and if there is a mismatch between her internal state and the context (agent thinks that the word heard refers to a plant not in her context) she again resorts to raising a clarification request. The speaking agent answers this clarification request by pointing to the plant intended, after which the hearing agent increases the association score of the word heard with the pointed plant. However, if the perceived plant is in her context then the hearing agent only reinforces its association score with the word heard.

---

[1] An agent's field of vision consists of a grid of fixed size originating from his location. Hence proximate agents have overlapping but not identical fields of vision.

A hybrid agent has a capability of either using the CR strategy or the introspective strategy. The agent only resorts to a clarification request if she cannot ground the word heard (there are no plants in her context or there is a mismatch between her internal state and the context). When hearing an unknown word and having some plants in the context the agent follows the introspective strategy.

After updating her lexicon[2] the hearing agent chooses the plant with the highest association score for the word heard. If this perceived plant matches with the speakers intended plant then the conversational interaction is deemed as a success. Neither agent is given any feedback on the outcome of their conversational interaction (see Smith (2005) for a similar approach).

2.2 Generational Turnover

A typical approach when modelling a multi-generational population is the introduction of mortality and child agents. The iterated learning model (Kirby et al., 2004) is an example of a multi-generational model where the language transmission is vertical (i.e. from one generation to the next). In such models the adult agents are always the speakers and child agents are always the hearers. The agents play a number of language games, which defines the length of a generation. At the end of a generation, the adults are removed from the model, the children become the new adults, and new children are introduced. This way of implementing generational turnover in the iterated learning model and other multi-generational models (e.g. Vogt and Coumans (2003)) is very rigid.

We propose a multi-generational model which is more realistic and resembles closer a human community (e.g. a tribe). In order to extend the mono-generational model described in (Ginzburg and Macura, in press) into a multi-generational model, there is a need to introduce a gradual agent turnover. This is done by introducing mortality. Every agent has a maximum age which is set randomly when the agent is born, and it lies in the range of $\pm 20\%$ from agent to agent. Upon reaching his maximum age the agent dies. Thus it is very unlikely that the whole adult population dies out at the same time as the adult agents are of different ages and have different maximum ages.

In order to keep the population size stable, we also introduce natality. So for every agent that dies a new infant agent is born to a random adult agent in the model. The infant agent inherits the parent's type (introspective, CR or hybrid). Infants have an empty lexicon, with no knowledge of the meaning space or the word space. Each infant follows the parent around and is only able to listen to the parent's dialogues with other agents. In fact an infant only hears the dialogues in which her parent is the speaker. So the assumption here is that an infant learns only the words uttered by her parent. An infant cannot be a speaker and learns exclusively by introspection. Every infant agent has an adulthood age which is set randomly and is about a sixth of the agent's lifespan. When reaching the adulthood age the infant stops following her parent and becomes an adult, meaning that it is able to walk around independently, engage in dialogues with other adult agents and become a parent. An infant can die only if her parent reaches the maximum age and dies.

This multi-agent model implements both vertical and horizontal language transmission as adult agents can communicate with each other as well as parent agents can communicate with their children. There is no clear distinction of when a generation starts and ends, like in the other multi-generational models, because there is continual agent turnover which makes calculating the results more intricate (see Section 3).

3 Experimental Results

This section describes different setups and experiment results for the model described in Section 2. In order to test the questions raised in (1) we ran several experiments in which agents posses distinct lexicons, and clarification requesting (CR) and introspective capabilities.

Before creating a population of agents, the environment is created containing 40 different plants (which represent 40 different meanings). There are three instances of every plant and they are randomly distributed in the environment.

The population in the simulations described here is made up of 40 agents that are also randomly distributed in the environment at the start. 20% of the initial population is made up of infants (i.e. 8 infant agents). Agents form two different communities each of whose members initially share a common lexicon. The initial com-

---

[2] Only the hearing agents update their lexicons after a conversational interaction.

munity lexicons are distinct from each other (in that no meaning has the same representation associated with it). Agents can be either of the same or different type within the community. Apart from the differences in the initial lexicons and types between the agents, all other properties are the same.

Once the simulation starts the agents begin walking randomly in the environment. At every time step agents' age increase and each agent moves to a random position in the environment. After moving an agent looks for other agents (that fall into his field of vision). If an agent sees another agent then two of them enter a dialogue where the 'see-er' is the speaker and the 'seen' is the addressee. After a dialogue the agents continue walking in a random direction. When an agent reaches his maximum age he dies and a new infant is born.

The performance of the model is based upon these behaviours which are collected at regular intervals in a simulation run:

- *Lexical Accuracy*: the population average of correctly acquired words. A word is said to be correctly acquired if it is associated with the same meaning as in either of the two initial lexicons.
- *Meaning Coverage*: the average number of meanings expressible by the overall population. There is no requirement that the meanings have correct associations.
- *Word Coverage*: the average number of words expressible by the population (correctness not taken into account).
- *Communicative Success*: the percentage of successfully completed conversations. A successful conversation is when the intended meaning by the speaker matches the perceived meaning by the hearer.
- *Method of Acquisition*: the percentage of conversational interactions that follow the introspective strategy or the CR strategy.
- *Distinct Lexicons*: the total number of distinct lexicons in the population. A lexicon is distinct only if there is no other lexicon in the population with which it shares all plant-word associations, so even if two or more lexicons have 19 out of 20 same plant-word associations they are regarded as distinct.
- *Lexical Convergence*: the percentage of agents sharing a lexicon. Agents share a lexicon if and only if all the plant-word associations are the same in their respective lexicons. Lexical convergence of 1 implies that all the agents use the same words for every plant in their lexicons.

The initial conditions and model parameters affect the above behaviours in complex ways. To determine what consequences arise when a single parameter is manipulated there is a need to control all other parameters and keep them constant whilst only manipulating the parameter being investigated.

Each parameter has a default value throughout the experiments, unless it is being investigated. The default and investigative values are shown in Table 1. In this paper we investigate the effect of increasing the meaning space on the lexicon dynamics of different populations.

Parameter	Default	Investigative
population size	40	-
adulthood age	5000 ±1000	-
max age	30000 ±5000	-
meaning space	40	20, 40, 60

Table 1: Default and investigative parameter values used during the experiments.

We ran four types of experiments with different population make-ups, namely introspective populations, CR populations, hybrid populations and mixed populations (made up of both introspective and CR agents in a 1:1 ratio). For all different experiments, 10 trial runs were carried out for statistical analysis.

In the first set of experiments the default parameter values as shown in Table 1 were used (Sectction 3.1). Then experiments with varying meaning space (Section 3.2) were carried out in order to get some insight into how it affects the outlined behaviours.

3.1 Multi-generational Experiments

The population in these experiments is kept constant to around 40 agents at any moment in time and the ratio of adults to infants is roughly 3:1. The agent life span is limited to around 30,000 ticks (±20%). Results were taken at every 20,000 ticks. The simulation is stopped when it reaches 2 million ticks, which means after around 70 generations.

The lexical accuracy initially drops very sharply for every population (Figure 1). At the beginning of the simulation there are a total of 80 words in the population (40 words from each community). As the words compete with one another there is a point when one word becomes dominant for a given plant and the majority of agents start using it. Thus the other competing words for the same meaning are used less frequently. The fact that the infant agents only learn the words uttered by their parents makes it very unlikely that the infrequently uttered words will pass to the next generation. After about three generations (100,000 ticks) the lexicon stabilises for the CR and hybrid populations, whilst for the mixed and introspective populations it keeps diverging.

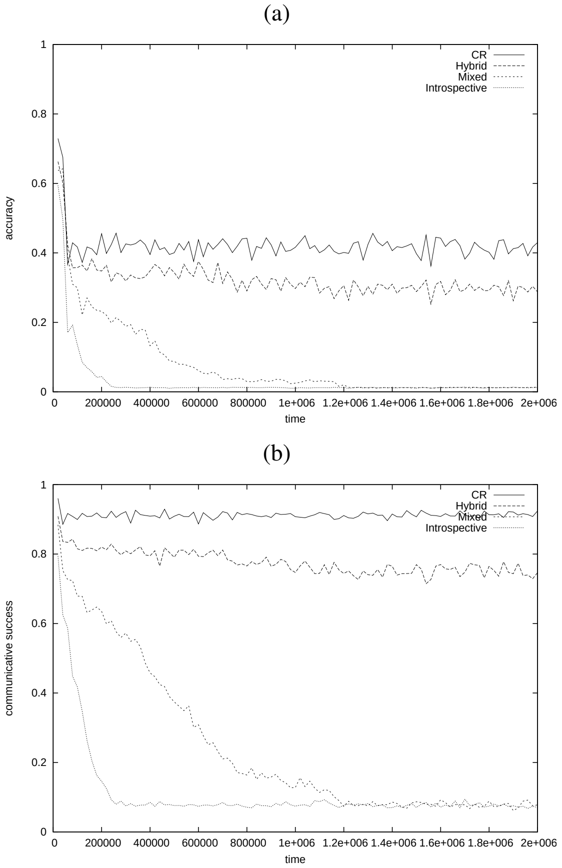


Figure 1: (a) Lexical accuracy, (b) Communicative success.

The reason for this stabilisation in CR and hybrid population can be explained by looking at Figure 2. Figure 2(a) shows that the meaning coverage for different populations is stable (all of them are able to express nearly every meaning). The word coverage however drops rapidly along with the lexical accuracy, as seen in Figure2(b). This is an indication that only the dominant words are surviving. Once the word coverage drops to around 50% the lexicon stabilises. Around 40 different plants are expressible by the population at this stage, so every plant is associated with one word. These words can be successfully passed onto the next generation as they are used with greater frequency.

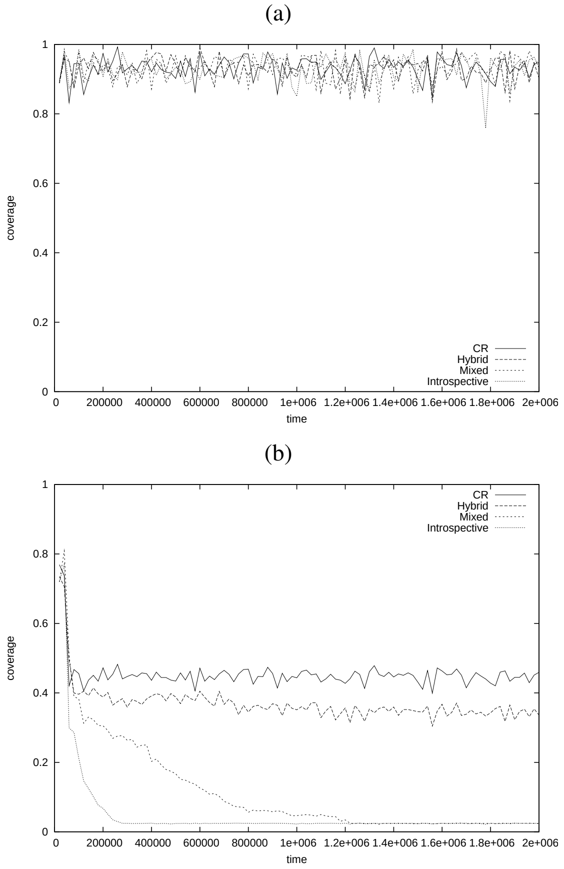


Figure 2: (a) Meaning and (b) Word coverage.

This is not the case for the mixed and introspective populations. The lexicon keeps diverging very rapidly and eventually reaches nearly 0% convergence (very few words have the association with the same plants as in the initial lexicon). Looking again at Figure 2 explains why this happens. The word coverage also drops very sharply, where in the end only one word is known by the whole population. The meaning coverage is comparable with other populations (where are all able to express nearly all the plants) so it is easy to see that every plant in the population is associated with this single word. The divergence is considerably slower in the mixed population than in the introspective.

The communicative success is in turn affected by the lexical accuracy as can be seen in Figure 1. The reason is that the higher the lexical accuracy is, the more similar the lexicons are between the agents in the population. Thus the more plant-word associations the agents share the more successful communications they are likely to have. Note that even though the lexicon is diverging at a fast rate initially, the agents in CR and hybrid populations are still able to communicate successfully about different plants.

The percentage of conversational interactions where introspective or CR strategy has been employed is shown by Figure 3. It can be seen that the populations in which CRs can be expressed (CR, hybrid and mixed) perform much better than the ones in which CRs can't be expressed (introspective). An interesting observation is that the clarification strategy in the mixed populations raises for the first 80,0000 ticks and then levels off. This is because as the lexicon is steadily diverging the agents are less successful in communicating, leading to more CRs being raised.

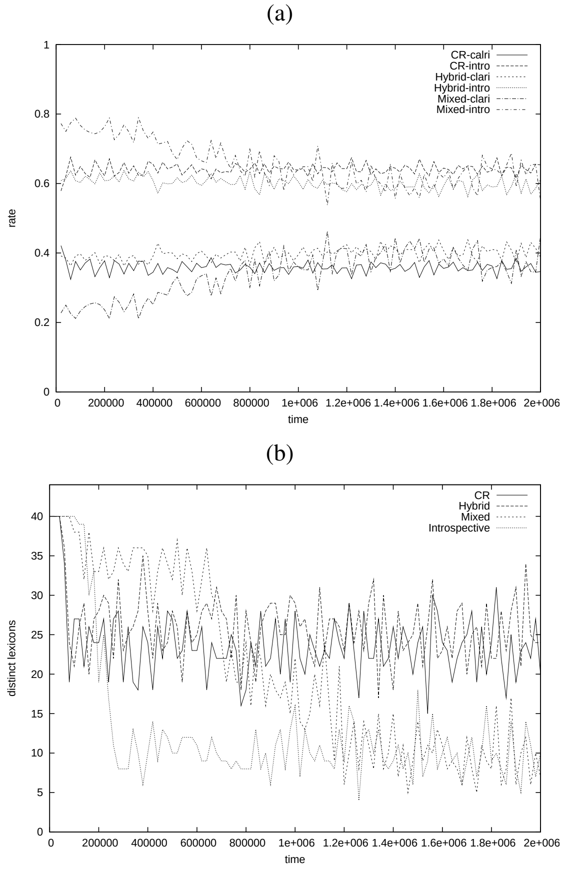


Figure 3: (a) Method of acquisition, (b) Distinct lexicons.

None of the populations converge to a single common shared lexicon (Figure 3(b)). One reason for this is that infant agents often have incomplete lexicons which differ from other agents, and this brings up the number of distinct lexicons. Another reason derives from the way common lexicons are calculated. Two or more agents are said to share a common lexicon if and only if all the plant-word associations are the same in their respective lexicons. But as there are 40 meanings it is very unlikely that all the agents will have the same associations. Thus, even though they might share the majority of the associations their lexicons are considered as distinct. We can induce from Figure 1 that the convergence for CR and hybrid populations is high where between 80% to 95% of the plant-word associations are shared. The lower number of distinct lexicons in the introspective and mixed populations might suggest that they have converged to a common lexicon. Strictly speaking, this is true, but as we have shown one word is used for representing every plant so the majority of agents converge to the same lexicon containing only this single word.

3.2 Meaning Space Variation

In this set of experiments we manipulate the meaning space—the number of different plants in the simulation. Increasing the meaning space involves increasing the differentiation among types of plants. The actual number of tokens remains constant (i.e. 120 plants). Thus when the meaning space is 20 there are six instances of each plant in the environment, whilst when the meaning space is 60 there are only two instances.

The effect of increasing the meaning space is similar for the different types of populations, thus we only present the results of CR populations. Figure 4 shows that increasing the meaning space from 20 to 60 causes a fall of around 20% in both the lexical accuracy and communicative success.

Meaning coverage is affected to a lesser extent but there is still a slight drop as the meaning space increases (Figure 5(a)). Word coverage, however, drops more significantly (Figure 5(b)). One reason for this is that as the meaning space increases the actual number of plants stays constant (e.g. for meaning space = 60 there are only two instances of each plant type in the environment). Therefore the agents are less likely to talk about all different plants as they encounter each one infrequently.

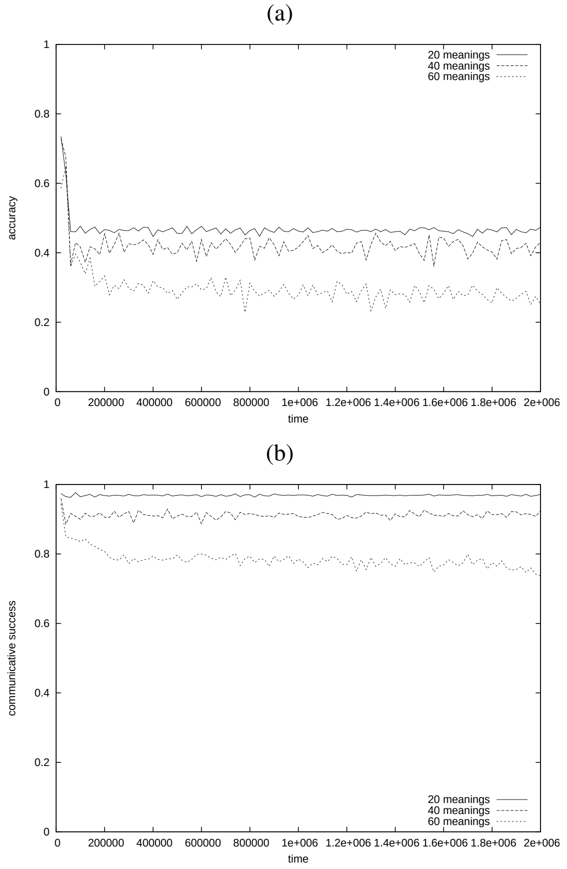


Figure 4: (a) Lexical accuracy, (b) Communicative success.

The percentage of clarification requests increases as more plant types are introduced (Figure 6(a)). The reason for this is, presumably, that there is greater uncertainty as to the referent of a word heard. This uncertainty rises as more plants are introduced, causing the agents to resort to clarification requests more often. Figure 6(b) shows that the number of distinct lexicons also rises as the meaning space increases: as there are more possible meanings it is less likely that agents will have the same association for all the meanings in their lexicons.

4 Conclusions and Future Work

In this paper we have discussed how metacommunicative interaction (MCI) serves as a key component in the maintenance of a linguistic interaction system. We ran a series of experiments on multi-generational populations in which lexicon dynamics of the populations that posses and lack MCI capabilities were compared.

We showed that limiting life span of agents in the multi-generational model raised some clear

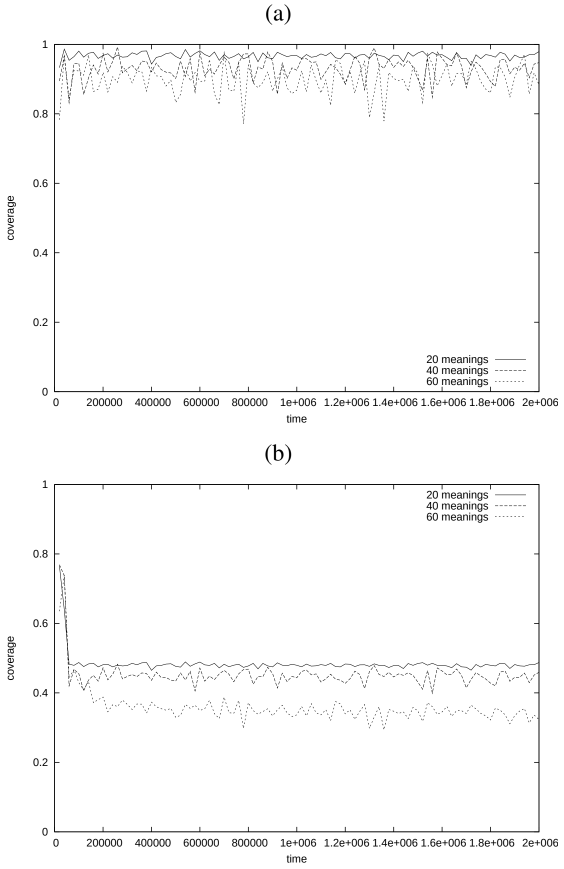


Figure 5: (a) Meaning and (b) Word coverage.

differences in the lexicon dynamics between the MCI capable and incapable populations. The main effect demonstrated is that in the introspective (and to a lesser extent mixed) populations the lexicon diverges continually, ending up with a situation where every agent in the population uses the same word to represent every plant in the environment. On the other hand MCI capable populations are able to maintain the lexicon, and the adult agents converge to a common lexicon.

We also investigated the effect of increasing the meaning space and showed that it speeds up the lexicon divergence for all populations irrespective of their acquisition method.

While this confirms our initial theorising, much work remains to butress it as a fundamental dividing line between MCI-ful and MCI-less populations. In our current experiments we are seeing that increasing the maximum age of agents improves the lexicon stability and convergence. Further work needs to be done in order to get more insight into this issue.

A crucial issue, which given space considerations we can only discuss here telegraphically, is

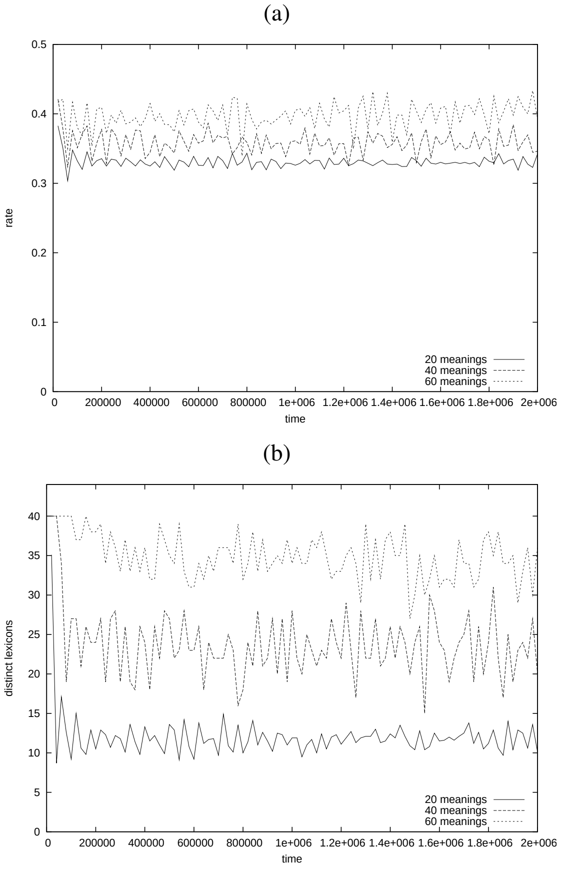


Figure 6: (a) Method of acquisition, (b) Distinct lexicons.

the relevance of the current simulation to *real human language use*.[3] There are a variety of simplifications in the current set up, possibly the grossest one is that agents employ a language lacking any sort of syntactical complexity. This would in turn lead to a massive increase in the size of the (potential) meaning space. While it is certainly an interesting and important extension to the current work, it is at least worth pointing out why such a move need not alter the current results beyond recognition. Recent corpus research on the distribution of clarification requests (see e.g. Rodriguez and Schlangen (2004); Purver (2006)) makes it clear that the lion's share of CRs in human conversation concern clarification of reference, of deixis, and of mishearing. Moreover, there is no evidence for CRs that concern syntactic ambiguity (e.g. attachment or scope). Hence, even in real human language the main communication difficulties seem to center on referential or lexical uncertainty.

Acknowledgements

This work is supported by an EPSRC quota award to Macura. We would like to thank two reviewers for Brandail for their very helpful comments.

References

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

Joshua M. Epstein and Robert Axtell. 1996. *Growing Artificial Societies: Social science from the bottom up*. MIT Press.

Jonathan Ginzburg and Zoran Macura. in press. Lexical acquisition with and without metacommunication. In Caroline Lyon, Chrystopher L. Nehaniv, and Angelo Cangelosi, editors, *Emergence of Communication and Language*. Springer Verlag, London.

Simon Kirby, Kenny Smith, and Henry Brighton. 2004. From ug to universals: linguistic adaptation through iterated learning. *Studies in Language*, 28(3):587–607.

Zoran Macura and Jonathan Ginzburg. in press. Lexicon convergence in a population with and without metacommunication. In E. Tuci P. Vogt, Y. Sugita and C. Nehaniv, editors, *Proceedings of the Third International Symposium on the Emergence and Evolution of Linguistic Communication (EELC06)*, Lecture Notes in Computer Science, Berlin. Springer-Verlag.

Matthew Purver. 2006. Clarie: Handling clarification requests in a dialogue system. *Research on Language and Computation*, to appear.

Kepa J. Rodriguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-orientaded spoken dialogues. In *Proceedings of the Eighth Workshop on Formal Semantics and Dialogue*.

Andrew D. M. Smith. 2005. The inferential transmission of language. *Adaptive Behavior*, 13(4):311–324.

Luc Steels. 1998. The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1(2):169–194, October.

Paul Vogt and Hans Coumans. 2003. Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation*, 6(1), 1.

Paul Vogt. 2005. On the acquisition and evolution of compositional languages: Sparse input and the productive creativity of children. *Adaptive Behavior*, 13(4):325–346.

---

[3] As emphasized by a Brandial reviewer.

Clarifying spatial descriptions: Local and global effects on semantic co-ordination.

Gregory J. Mills
Interaction, Media and Communication
Research Group
Department of Computer Science
Queen Mary, University of London
London E1 4NS
`gj@dcs.qmul.ac.uk`

Patrick G. T. Healey
Interaction, Media and Communication
Research Group
Department of Computer Science
Queen Mary, University of London
London E1 4NS
`ph@dcs.qmul.ac.uk`

Abstract

A key problem for models of dialogue is to explain the mechanisms involved in generating and responding to clarification requests. We report a 'Maze task' experiment that investigates the effect of ‘spoof’ clarification requests on the development of semantic co-ordination. The results provide evidence of both local and global semantic co-ordination phenomena that are not captured by existing dialogue co-ordination models.

1 Introduction

Perhaps the shortest possible clarification question is attributed to Oscar Wilde[1]. After sending a telegraph to his Parisian literary agent enquiring about the sales figures of his latest novel he received the response that sales of the book were indeed favourable. Wilde's subsequent telegraph was the single-character "?", to which the agent responded with the equally terse "!". Wilde could, of course, have formulated his clarification question differently, potentially leading to different patterns of response by his agent:

A: The sales are favourable.

1. W: What?
   A: Sales are favourable
   A: Sales are better than expected

2. W: How good ?
   A: 300 pounds
   A: More than your previous book

3. W: Favourable?
   A: Yes. 612 copies
   A: No. Incredible

Clarification requests (henceforth CRs), such as (1)-(3) above, are used to signal potential problems with the interpretation of a previous utterance. They are thus central to maintaining co-ordination in dialogue, as they serve the purpose of bringing the conversation "back on track" (Schegloff, 1992) when inter-subjectivity is threatened.

An account of the mechanisms underlying the use of different CR’s, and their effects on the interaction is essential for an adequate understanding of dialogue and important for the practical goal of creating more natural, robust dialogue systems. However, empirical investigations of CRs have generally been limited to post-hoc analysis of corpora. For practical reasons it is difficult to achieve the levels of control necessary to support experimental manipulations of CR’s. This has made it difficult to compare the effects of different CR’s on conversational trajectories or subsequent semantic co-ordination.

This paper develops an experimental technique described in Purver et al. (2003). It combines an experimental chat-tool with a version of the maze game developed by Garrod & Anderson (1987). This enables the introduction of artificial 'probe' clarification requests into participants' dialogue without causing overt disruption to the conversation. By manipulating the type of probe CR used we can investigate their relative

---

[1] This has also been attributed to Victor Hugo, seeking the opinion of his publishers on his latest manuscript, and is possibly apocryphal.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 122–129.

impact on semantic co-ordination in the maze game (cf. Garrod & Anderson, 1987; Garrod & Doherty 1994).

2 Clarification Requests

2.1 Levels of misunderstanding

A common thread running through the CR taxonomies proposed by Schlangen (2004), Gabsdil (2003) and Purver (2003), is that different CRs access different levels of understanding within some form of action hierarchy or 'ladder' (Clark, 1996; Allwood, 1995). An example ladder is:

Level 4. Action recognition.
Level 3. Meaning recognition.
Level 2. Utterance recognition.
Level 1. Securing Attention.

So, for example, the "Favourable?" CR above might typically request further specification of the meaning (level 3) whereas the "What?" CR might typically request clarification of what the initial utterance was (level 1) (although see Drew, 1997). Communication is only fully complete if understanding is secured at all levels (although see Allwood, 1995).

Hearer's choice of clarification type can thus signal the information required for them to reach a higher level of understanding. CRs such as (1) above typically signal low co-ordination as they give fewer clues about the nature of the problem or expected response (Schlangen 2004) than CRs such as (3), which requests further specification of what 'favourable' could mean in this context.

This ordering trades on a pragmatic expectation that people normally design their CR's to give as much information as possible about their current level of understanding. Although "what's" can be used to clarify at higher levels the expectation is that people should produce CR's that signal the highest level of understanding currently available to them. In the collaborative model this is formulated as the "strongest initiator rule" (Clark and Schaefer, 1989) which posits that a "[hearer] ought to index the parts he did hear, or the parts he didn't hear, and request help" in reaching the higher levels of comprehension (see Drew, 1997 for criticism of this formulation).

2.2 Semantic co-ordination

'Ladder' approaches can thus categorize and rank sources of problematic understanding, however they are, in effect, 'semantically neutral'. The different levels of the hierarchy don't address the potential for different forms of co-ordination that depend on semantic differences. This possibility is illustrated by data from the Maze game (Garrod & Anderson, 1987).

In the Maze game (see below) participants are faced with a recurrent problem of describing locations to each other. Over time their spatial descriptions shift from predominantly instance-bound 'Figurative' (Figural/Path descriptions described below) versions that depend on the specific configuration of the current maze to more 'Abstract' (Line/Matrix described below) approaches that invoke a relatively systematic underlying model of the maze that abstracts away from each instance and generalises across instances more easily.

There is evidence that interaction mechanisms, and not simply task experience, play a specific role in this shift toward more 'Abstract' description schemes. Even where task experience is equivalent, pairs' preference for 'Abstract' or 'Figurative' schemes differs depending on the opportunities they have had for interaction (Garrod and Doherty, 1994; Healey, 1997). There is also evidence that participants can only develop a shared 'Abstract' scheme following a prior stage of co-ordination with a 'Figurative' scheme (Healey & Mills, 2006; Healey, et. al. in press). Drawing on data from a different task, Schwartz (1995) argues that the shift towards abstraction emerges as a result of general collaborative processes that are unavailable to solitary speakers.

The experimental evidence thus suggests that differences in choice of description type correspond to differences in the degree of semantic co-ordination developed between dialogue participants. However, these shifts in semantic co-ordination are not readily explained by existing accounts of dialogue co-ordination. For example, they are not due to the kind of 'contraction' of referring expressions observed in many definite reference tasks (see e.g. Clark, 1996). The 'Abstract' descriptions are not reduced versions of Figurative descriptions, they involve a change in the underlying semantic model of the maze that participants are using (Garrod and Anderson, 1987). Also, as Garrod (1999) argues, local entrainment / priming mechanisms of the kind incorporated into the interactive alignment model (Pickering and Garrod, 2004) are conservative and not equipped to address global trends towards abstraction or innovations in description type. When people change schemes, the interactive alignment prediction is that the most frequently used (i.e., primed) prior scheme will predominate. A global shift towards a new scheme thus requires us to consider alternative co-ordination mechanisms.

Similarly, there is a general expectation in both Purver et al's (2003) and Rodriguez & Schlangen's (2004) models that interlocutors will modify their original utterance in response to a CR. However, there are no mechanisms for predicting what kinds of semantic change occur in response.

In summary, there is evidence that interaction contributes directly to the development of semantic co-ordination. Prima facie it seems likely that clarification requests play a key role in this process (cf. Clark, 1996; Pickering and Garrod 2004; Healey and Mills 2006). However, existing models of dialogue do not provide clear ways of interfacing between patterns of clarification request and possible semantic changes that might occur as part of the response. One reason for this is that it has not been possible to systematically investigate the effects of different kinds of CR on dialogue co-ordination.

Two basic empirical questions that arise then are a) whether there is a direct connection between the occurrence of CR's and semantic co-ordination and b) whether there is a connection between the 'level' of CR and the form of semantic co-ordination.

To address these questions, a "Maze Game" experiment was set up using a text-based chat tool. The basic rationale of the experiment was to test the effects of different CR types on the form and content of participants' responses. Before introducing our specific hypotheses we explain the experimental methods in more detail.

3 Methods

The experiment employs a modified version of the "Maze Game", devised by Garrod and Anderson (1987). This task creates a recurrent need for pairs of participants to produce location descriptions. These descriptions can be reliably classified into four broad categories (see below), thus enabling the indexing of semantic co-ordination between participants (Garrod & Anderson, 1987; Garrod & Doherty 1994).

To support turn-level experimental manipulations of the dialogue, a chat-tool technique is used that engages participants in artificial clarification sequences. Before giving details of the procedure we first describe the implementation of the maze game and chat-tool used in the experiment and then explain the generation of these clarification requests.

3.1 The maze game application

The maze application is written in Java and displays a simple maze consisting of a configuration of nodes that are connected by paths to form grid-like mazes (see Fig 1). The mazes are based on a 7x7 grid and are selected to provide both grid-like and asymmetric instances.

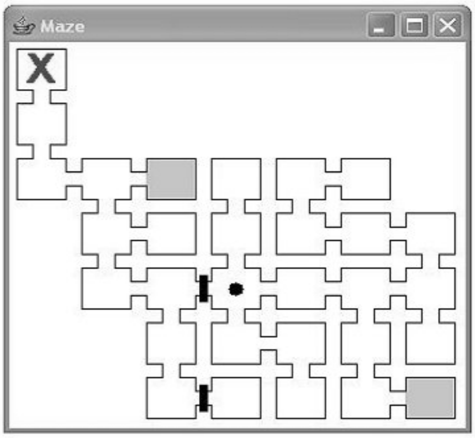


Figure 1: Example maze configuration. The solid black circle shows the player's current position, the cross represents the goal point that the player must reach, solid bars the gates and grey squares the switch points.

Subjects can move their location markers from one node to another via the paths. Each move is recorded and relayed to the server where it is time-stamped and stored. The game requires both subjects to move their location markers from a starting location to a goal that is marked with a cross. Although the maze topology is the same for both subjects, each subject has a different starting location and goal, neither of which are visible to the other subject. They are also not able to see each other's location markers.

Movement through the maze is impeded by gates that block some of the paths between nodes. These gates can be opened by the use of switches (grey coloured nodes). The locations of switches and gates are different on each maze and not visible to the other subject. Whenever a subject moves to a node that is marked as a switch on the other's screen, all of the other subject's gates open. All the gates subsequently close when they move off the switch.

This constraint forces subjects to collaborate: in order for participant (A) to open their gates, A has to guide participant B onto a node that corresponds to a switch that is only visible on A's screen. Successful completion of a maze (when both reach their respective goals) therefore requires subjects to exchange descriptions of their location and the locations of gates, switches and goals. Each new maze has a new configuration, starting points, gates and switches.

3.2 The Chat Tool

All communication takes place via a custom-built java chat tool similar to desktop messaging applications. The display is split into an upper window, a status bar and a lower window. The upper window displays the ongoing conversation, and the lower window is used for typing. All key presses are time-stamped and stored for later analysis. The status bar is a prominent single line of text that is controlled by the server and is similar to the status bar of proprietary messaging tools that display the activity status of the other conversant.

3.3 The Chat Server

In addition to relaying turns between participants, the server monitors the content of the turns in order to generate artificial clarification requests that appear, to participants, to originate from each other.

The server compares each turn with a lookup table of location descriptions obtained from a previous corpus of 10000 maze game turns (Healey & Mills, 2006), combined with rules for detecting misspellings and non-standard "txt" conventions. This ensures that CR's are generated only on turns containing spatial descriptions.

Each clarification request generated by the server is preceded by the other participant's chosen nickname, followed by a colon, and is dynamically modified to mimic spelling and typing speed.

To provide a manipulation of CR type the two classes of CR were selected; Reprise Fragments (‘Frags’) that echo a word from the target turn and ‘Whats’ (e.g., “what?” or “sorry?”) that query the turn as a whole. These are the two most common forms of CR in ordinary dialogue (Purver et. al. 2003) and they provide two different levels of clarification. Reprise fragments involve direct re-use of a word from the turn and imply that the rest of the turn was understood. By contrast ‘Whats’ suggest that there were global problems finding a sense for the turn (but see also Drew 1997).

‘Frags’ (High co-ordination): Repetition of a single fragment of the location description.

‘Whats’ (Low co-ordination): What? Huh? Sorry? Ehh? Uhh? Where?

Participants' responses to the probe CR’s are captured by the server. The probe CR and the response are displayed only in the participant’s own chat-window. After receiving a response to the CR, the server sends one of the following acknowledgement turns to the recipient: “ok”; “k”; “ok right” and resumes relaying subsequent turns as normal. During the ‘fake’ CR exchange the server monitors whether the other participant starts typing. If this occurs, an error message is displayed and further text-entry is prevented until either the CR sequence is finished or a predefined time-out threshold is reached. To ensure error messages do not cue the interventions, a small number of random error messages are also introduced at other points in the dialogue.

3.4 Subjects

21 pairs of native English speaking subjects were recruited, 23 male and 19 female, from undergraduate students. They were recruited in pairs to ensure that they were familiar with each other. Only subjects who had some previous experience of using internet chat software such as ICQ or Microsoft Messenger were selected for the experiment. Each subject was paid £10.00 for participating in the experiment.

3.5 Procedure

Pairs of subjects were seated in separate rooms in front of a desktop PC. On each PC a window containing the maze (same configuration but different features see Fig 1) and a chat-tool window are displayed. Subjects were asked to select a nickname to be used in identifying chat turns and then wait for further instructions.

Subjects were told that the experiment was investigating the effects of a novel chat-tool and computer game on how people interact with each other. They were informed that their interaction would be recorded anonymously for subsequent analysis. Subjects were advised that they could request the log to be deleted and were free to leave at any time but would still receive payment in full.

They were given a written description of the maze game and told that the experiment involved solving twelve mazes. No information was given about the CRs generated by the server. At the end of the experiment the full nature of the experimental interventions was explained.

Twelve mazes were presented in random order to each pair. Artificial clarification requests on turns that involved spatial descriptions were introduced throughout the experiment with a 5% probability of any turn being clarified – consistent with findings from Purver et al. (2003) and Schlangen (2004) that approximately 5% of dialogue turns are CRs.

The experimental group were thus exposed to a within-subjects manipulation of CR type (‘Frag’ vs. ‘What’). A control group of 12 additional pairs, recruited from the same undergraduate population, followed the same procedure but without the manipulation of probe CR’s.

Experimental Hypotheses:

1. The introduction of artificial CR’s will interfere with semantic co-ordination
2. More severe problems will cause more disruption (i.e., ‘Whats’ will cause more disruption than ‘Frags’)
3. People will systematically shift to more Figurative forms of semantic co-ordination (Figural / Path descriptions below) where problems occur.

4 Results

Overall, 246 clarification requests were artificially generated by the server: 109 ‘Frags’, 128 ‘Whats’ and 9 CRs generated for turns that did not contain spatial descriptions. These non-spatial clarifications were excluded from further analysis. On debriefing, nobody in the experimental group reported detecting that the probe CRs did not originate from their partner.

4.1 Description Types

Both the target turns used by the server to generate CRs and subjects' responses were classified according to the criteria developed by Garrod and Anderson (1987). This categorizes location descriptions into four basic classes corresponding to different underlying mental models of the maze:

Figural: a heterogeneous category of relatively concrete descriptions that draw on some specific element of the overall configuration of particular features to identify a target location.

A: “right above the sticking out bit at the top”

Path: involves identifying a route to be traversed through the maze to the target location. Path descriptions are sensitive to the specific layout of boxes and connections in the maze.

A: “From middle go up 1, 2 right, 1 down”

Line: classifies the maze into a set of line elements corresponding to rows, columns or diagonals. The target line is described first, followed by the target box as a position along it.

A:“In the bottom box, 2nd column from right”
A:“The third row, fifth to the left”

Matrix: introduces a Cartesian coordinate system with locations identified via the specification of two vectors either as rows and columns or in terms of numbers or letters for each axis.

A : “My switches are at 4,6 5,4 . I'm on 3,4”
A: “I'm in the 3rd row, 4th column”

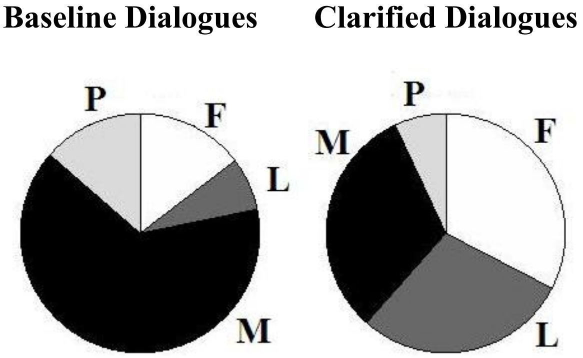


Fig 2: Global distribution of description types in baseline (control) condition and in dialogue queried with clarification requests (F = Figural, P = Path, L = Line, M = Matrix)

4.2 Distribution of description types

Figure 2 above illustrates the contrast in the global distribution of description types in the baseline control condition and in dialogue that is periodically interrupted with artificial clarification requests. The difference in use of description types is reliable (Multinomial Regression: $Chi^2_{(3)}$=276, p=0.00). The results show that the probe CRs significantly disrupt co-ordination in the experimental group. The largest category of description type in the experimental group is Figurative whereas in the baseline control group the Matrix descriptions predominate.

To check whether co-ordination was still developing over time (but to a lower level) in the experimental group the distribution of description types used in target turns the first four games was compared with those used in the last four games (see Fig 3).[2] This showed that there was still a significant shift in the use of description types over time (Multinomial Regression: $Chi^2_{(3)}$ =15.1, p=0.00) with participants migrating from 'Figurative' descriptions (Figural/Path) towards 'Abstract' (Line/Matrix) in the later games. This suggests that semantic co-ordination was still developing but at a significantly slower rate than in the control group.

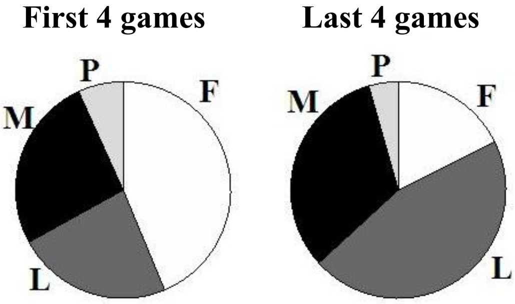


Fig 3: Global distribution of description types in first four and last four games in the experimental (CR) group.

In order to test the effects of the two CR types on the way responses were formulated a focused comparison of the distribution of description types in the responses immediately following the 'What' vs 'Frag' CRs was made. This showed no reliable difference (Multinomial Regression $Chi^2_{(3)}$ =1.68, p=0.64).

To provide an additional test of the third hypothesis –namely that people systematically shift to more 'Figurative' description types as a way of resolving co-ordination problems- we examined the relationship between the description type used in the target turn produced by a speaker and the spatial description type they produced in their response to the probe CR. Out of a total of 142 spatial description responses 101 (71%) responses used the same description type as the target. Of the 29% (41) that changed description type 14 (34%) involved a shift from 'Figurative' (Figural/Path) to 'Abstract' (Line/Matrix) whereas 27 (66%) involved a shift in the opposite direction. Overall, responses to the CR's predominantly used the same description type but where a change occurred it was more likely to involve a change to a more Figurative description type.

4.3 Other Measures of CR Effects

Times from the log files were used to provide two further comparisons of responses to the 'Frag' and 'What' CR's. Firstly, turn completion time – the time from the onset of typing of a response to its completion. A one-way analysis of variance revealed reliable differences between CR types, ($F_{(1,235)}$=6.5, p= 0.01). Overall, participants took longer to formulate their responses to 'Whats' than to fragments, taking an average of 18 seconds to respond to the former, and 25 seconds to respond to the latter.

The second measure of response time used was typing-onset time: the time between the onset of an intervention and the initial onset of typing the response. A one-way analysis of variance showed no effect of CR type ($F_{(1,235)}$= 0.32, p = 0.57).

In order to provide a measure of the indirect disruption caused by a CR, data from the log files was used to calculate the number of turns between receipt of a CR by a participant and the next turn in which they produced a spatial description (see 5.1 below). A one-way analysis of variance revealed significant differences between CR types ($F_{(1,93)}$=8.46, p=0.02). Overall, 'Frags' caused less disruption (average 3.2 turns before next description) than 'Whats' (average of 5.3 turns before next description).

The log files were also analysed for number of 'deletes' or edits that occurred in the construction of a turn prior to sending it. Although there was no reliable evidence of a relationship between edits and description type in responses ($Chi^2_{(1)}$ =0.881, p= 0.35) there was a reliable relationship between edits of the target turn and description type of the subsequent response to the CR ($Chi^2_{(1)}$ = 9.9, p=0.002). If there were no edits in the target turn there were more 'Abstract' responses (44 Matrix/Line vs. 23 Figural/Path). If the target was edited prior to sending there were fewer 'Abstract' responses (32 Matrix/Line vs. 43 Figural/Path).

5 Discussion

The global distribution of description types reported here (see also Healey and Mills, 2006) replicates the patterns of use observed in spoken Maze game studies (Garrod and Anderson, 1987; Garrod and Doherty, 1994; Healey, 1997). In particular, the pattern of migration from relatively concrete descriptions (Figural/Path) that depend on the specific details of each maze, towards more abstract description types (Line/Ma-

---

[2] Target turns only were selected for this analysis as these would be furthest from the immediate influence of the artificial CRs.

trix) that invoke schemata that generalize across instances (see above) is the same in both modalities.

The advantage of using chat tools with the maze game is that it makes it possible to carry out context sensitive, turn-level experimental manipulations of dialogue; in the present experiment the manipulation of probe CR's. This allows us to address the question, raised in the introduction, of whether CR's have a direct effect on patterns of semantic co-ordination. The experimental results presented above provide strong evidence for such a connection.

While pairs in the control group converge on the 'Matrix' scheme, those exposed to CR's do not. Although their form of co-ordination does change over time it evolves more slowly and they do not converge on the Matrix scheme by the end of the experiment. The marked difference in the distribution of description types between the experimental (CR) and control (baseline) groups thus supports hypothesis 1.

The second question raised in the introduction was whether there is a connection between the particular type or 'level' of CR and form of semantic co-ordination. The results reported here do not provide a clear answer to this question. Hypothesis 2 predicted that 'Whats' would cause more disruption than 'Frags', however no reliable difference was found in the distribution of description types in responses to the two CR types. This is not, however, because the two CR types failed to have any distinct effects. The response time data show participants took longer to formulate their responses to 'Whats' than 'Frags'. In addition, the 'disruption' data indicate that participants took longer to get the dialogue back on track after a 'What' than a 'Frag'.

Overall, participants were sensitive to the difference between the two classes of CR. As expected, the 'Whats' were more disruptive to the dialogue than the 'Frags'. However, while the evidence thus supports hypothesis 2, the results make it more difficult to explain the nature of the connection between CR's and forms of semantic co-ordination.

The third hypothesis considered above was that the local effect of CR's should be to prompt a shift from 'Abstract' (Matrix/Line) to 'Figurative' (Figural/Path) descriptions. The results provide some support for this. In the cases where participants do change description type in response to a CR, there is a greater preference for changing from 'Abstract' to 'Figurative' than vice versa. However, the more striking observation is that in 71% of cases participants do not change type. This local consistency in description type echoes Garrod's original findings (Garrod and Anderson, 1987; Garrod and Doherty, 1994). However it presents a puzzling contrast with the global effects of the CR's. Although the additional clarification questions have a significant impact on overall co-ordination –as indicated by choice of description type– it appears that these effects are not manifest in the immediate context in which the CR's occur.

Perhaps the simplest potential explanation for the apparent contrast between the local and global effects on semantic co-ordination is that the CR's undermine participants' confidence in the interaction as a whole. So, although they are locally consistent in their response to the CR, they subsequently become more generally conservative in their choice of description types. If we treat editing of the target turn as an index of confidence prior to the CR then there is some support for this in the data. Figurative responses are more likely after CR's to an 'edited' target. This is consistent with a view that the CR aggravates the lack of confidence. A 'confidence' explanation, however, still provides no mechanism that can explain the trend towards more abstract forms of semantic co-ordination.

In the introduction we noted some problems of the action 'ladder' approach as a way of analyzing differences in semantic co-ordination. The global character of the effects observed here suggests an additional problem. Rather than pointing to 'vertical' modifications to the ladder they indicate a need for more 'horizontal' co-ordination mechanisms that could operate over larger stretches of interaction.

There are two methodological issues which need to be resolved in future work. First, the experiment was designed to produce CR's with a frequency similar to everyday conversation. In practice this resulted in each participant being exposed to one CR approximately every 40 turns. If it is true that the effects of the CR's are global rather than local it is possible they interfered with each other. In particular it suggests that combining the 'What' and 'Frag' manipulations in a single within-subjects condition is problematic. Second, the main advantage of using the Maze task is that the taxonomy of description types provides an attested way of indexing semantic co-ordination. However, this is still a relatively crude measure. Within each category there is considerable variation in how the descriptions are constructed and used (see e.g., Garrod and Anderson 1987; Garrod and Doherty 1994 for discussion). As a result local ‘sub-description type’ changes in response to the CR’s would not be detected. The global vs. local contrast in the data could thus be an artifact of the measures of semantic co-ordination used.

Nonetheless, the results clearly show that participants reliably distinguish between CR types, and also show that the introduction of CR's into participants' dialogue has a strong effect on the kinds of description used in the maze game. Thus, future studies need to develop a more detailed analysis of the local impact of CR's on semantic co-ordination.

6 Conclusion

The data from the present experiment demonstrate a causal connection between the use of clarification questions and the development of semantic co-ordination in the maze task dialogues. Contemporary models of dialogue co-ordination need to be modified to accommodate these semantic effects. However, further empirical work is required to clarify the mechanisms involved.

References

Allwood, J. (1995). An activity based approach to pragmatics. *Gothenburg Papers in Theoretical Linguistics, 76* , Gothenburg, Sweden.

Clark, H. H. & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive Science, 13,* 259-294

Clark, H. H. & Marshall, C (1981) Definite reference and mutual knowledge. *Elements of discourse understanding,* Cambridge University Press.

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

Drew, P. (1997) ‘Open’ class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics* 28, pp.69-101.

Gabsdil, M. (2003). Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue,* Stanford, USA.

Garrod, S. C., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181-218

Garrod, S. C. & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition, 53,* 181-215.

Garrod, S. (1999) The Challenge of Dialogue for Theories of Language Processing. In S.Garrod and M.Pickering (Eds) Language Processing, Hove: Psychology Press. 389-416

Healey, P. G.T. (1997). Expertise or expert-ese? The emergence of task-oriented sub-languages. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society.* Stanford, California.

Healey, P.G.T., Purver, M., King, J., Ginzburg, J. and Mills, G. (2003) Experimenting with Clarification in Dialogue. In Alterman, R. and Kirsh, D. (eds) Proceedings of the 25th Annual Conference of the Cognitive Science Society. Mahwah, N.J.: LEA pp.539--544.

Healey, P.G.T. (2004). Dialogue in the degenerate case? (Peer commentary on Pickering & Garrod (2004), *Behavioural and Brain Sciences,* 27(2)

Healey, P.G.T. & Mills, G. (2006) Participation, Precedence and Co-ordination. (forthcoming) In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver. Canada.

Healey, P.G.T., Swoboda, N., Umata, I. and King, J, (in press) Graphical Language Games: Interactional constraints on representational form. *Cognitive Science.*

Purver, M. & Healey, H. & King, J. & Ginzburg J. & Mills, G. Experimenting with Clarification in Dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society (CogSci 2003)*, Boston, MA.

Purver, M., Ginzburg, J. and Healey, P.G.T (2003) On the Means for Clarification in Dialogue. In Smith, R. and van Kuppevelt, J. (eds) *Current and New Directions in Discourse and Dialgoue.* ACL, pp. 235-255.

Pickering, M., & Garrod, S. (2004). The interactive alignment model. *Behavioural and Brain Sciences, 27*(2), 169-189.

Rodriguez, K. J. & Schlangen, D. (2004). Form, Intonation and Function of Clarification Requests in German Task-oriented Spoken Dialogues. In *Proceedings of the Eighth Workshop on Formal Semantics and Dialogue.*

Schegloff, E. A. (1992) Repair after next turn: The last structurally provided defense of intersubjectivity in Conversation. In *American Journal of Sociology* (97)5, 1295-1345.

Schlangen, D. (2004). Causes and Strategies for Requesting Clarification in Dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*

Schwartz, D. L. (1995). The emergence of Abstract Representations in Dyad Problem Solving. *The Journal of the Learning Sciences,* 4(3), 321-354.

On “Uh” and “Uhm” and some of the Things They are Used to Do (invited presentation)

Emanuel Schegloff
Departmane of Sociology
University of California LA
`schegloff@soc.ucla.edu`

My invitation to participate in this occasion remarked that “it would be especially interesting and fruitful for us to initiate a dialogue between Conversation Analysis and fields more traditionally represented at this workshop (e.g., AI, computational linguistics, psychology), to increase our understanding of similarities (and perhaps differences) in approach and findings,” and I have chosen my topic accordingly. The empirical focus of my presentation will be “uh(m);” the “field more traditionally represented at this workshop” will be psycholinguistics/cognitive science; among the similarities and differences in approach to be taken up will be the study of naturalistic data and the use of *corpora*, as in *corpus (psycho-)linguistics*. The theme of the talk (most generally put) is that the natural home of language is in talking; that the natural home of talking is in interaction; that talking-in-interaction is the product of describable organizations of practice that we now know something about; that these organizations of practice engender *places* or *positions* in the talk; that virtually everything in conversation needs to be understood by reference to both position and composition; and, consequently, that a proper understanding of language and of its deployment and understanding in the natural world will require coming to terms with the practices of talking-in-interaction.

Goal-oriented Dialog as a Collaborative Subordinated Activity involving Collective Acceptance

Sylvie Saget
IRISA - Projet CORDIAL
6, rue de Kérampont - BP 80518
22305 Lannion
`Sylvie.Saget@enssat.fr`

Marc Guyomard
IRISA - Projet CORDIAL
6, rue de Kérampont - BP 80518
22305 Lannion
`Marc.Guyomard@enssat.fr`

Abstract

Modeling dialog as a collaborative activity consists notably in specifying the content of the Conversational Common Ground and the kind of social mental state involved. In previous work (Saget, 2006), we claim that Collective Acceptance is the proper social attitude for modeling Conversational Common Ground in the particular case of goal-oriented dialog. We provide a formalization of Collective Acceptance, besides elements in order to integrate this attitude in a rational model of dialog are provided; and finally, a model of referential acts as being part of a collaborative activity is provided. The particular case of reference has been chosen in order to exemplify our claims.

1 Introduction

Considering dialog as a collaborative activity is commonly admitted (Clark, 1996; Garrod and Pickering, 2004; Cohen and Levesque, 1991; Cohen and Levesque, 1994). Generally speaking, modeling a particular collaborative activity requires the specification of the collective intention helds by the agents concerned and requires the specification of the Common Ground linked to this activity. Common Ground refers to pertinent knowledge, beliefs and assumptions that are shared among team members (Clark, 1996). Thus, Common Ground is a collection of social mental attitudes.

The Common Ground linked to the dialogue itself (the Conversational Common Ground, CCG) ensures the mutual understanding of dialog partners. The CCG enables dialog partners to use abbreviated forms of communication and enables them to be confident that potentially ambiguous messages will be correctly understood (Klein et al., 2005). Dialogue partners become aligned at several linguistics aspects (Garrod and Pickering, 2004). There is an alignment, for example, of the situation model, of the lexical and the syntactic levels, even of clarity of articulation, of accent and of speech rate. Interactive alignment, of team members' situation model and of social representations, facilitates language processing during conversation and facilitates social interaction.

In the particular case of referent treatment, even for daily task, which use well-known objects with common known proper names to refer to, there is a wide range of possible manners to describe this object by words. To ensure mutual understanding, humans *”associate objects with expressions (and the perspectives they encode), or else from achieving conceptual pacts, or temporary, flexible agreements to view an object in a particular way”* (Brennan and Clark, 1996).

Thus, the Conversational Common Ground, since dialog is a mediated activity, contains all grounded elements linked to the way to communicate (as the necessary level of clarity of articulation or speech rate) as well as elements of dialog's history such as association between modes of presentation (linguistic objects) and mental representations: associations as conceptual pacts.

In previous work (Saget, 2006), we claim that Collective Acceptance is the proper social attitude for modeling Conversational Common Ground in the particular case of goal-oriented dialog. In the first part of this paper, we show

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 131–138.

that such a modelization fits better than stronger mental attitudes (such as shared beliefs or weaker epistemic states based on nested beliefs). We also show that this modelization may be considered as partly due to the subordinated nature of goal-oriented dialog. Then, in the last part of the paper, a formalization of Collective Acceptance and elements are given in order to integrate this attitude in a rational model of dialog. Finally a model of referential acts as being part of a collaborative activity is provided. The particular case of reference has been chosen in order to exemplify our claims.

2 Collective Acceptance: the proper social attitude for modeling CCG

2.1 General claims on reference

In order to model dialog as a collaboration, reference resolution has to be considered as the *"act identifying what the speaker intends to be picked out by a noun phrase"* (Cohen and Levesque, 1994). Moreover, the collaborative nature of reference have been brought to the forefront (Clark and Wilkes-Gibbs, 1986). More precisely, reference is not the simple sum of the individual acts of generating and understanding, but is a collaborative activity involving dialog partners. Thus, according to H.H. Clark et al. in (Clark and Bangerter, 2004), these individual acts are motivated by two interrelated goals:

- Identification: Speakers are trying to get their addressees to identify a particular referent under a particular description.
- Grounding: Speakers and their addresses are trying to establish that the addressees have identified the referent as well enough for current purpose.

How the identification goal is achieved ? First at all, when speaker has the intention to refer to a particular object, he has to choose a description of this object. Traditionally, this choice is viewed as depending on the beliefs of dialog participants and as depending on availability. In other words, speaker can refer with a definite description $\imath x.\phi(x)$ to an object o iff it is in the unique available object for which $\phi(o)$ holds. Moreover, H.H. Clark and C.R. Marshall (Clark and Marshall, 1981) claimed that mutual knowledge of $\phi(o)$ is necessary, if a description should refer successfully to an object o.

For example, let's imagine that two persons, Tom and Laura, who have been to the same school. Tom suggests to Laura: "Shall we meet in front of our ex-school's basketball court". The choice of the description of the intented place should be explained by the fact that Tom thinks that the following mutual belief is part of their common ground:

- $MBel_{Tom,Laura}(frontOf(l,h)$
  $\wedge basketballCourt(h)$
  $\wedge partOf(h,g)$
  $\wedge studentAt(Tom,g)$
  $\wedge studentAt(Laura,g))$,
  where:
  - $MB_{i,j}(\phi)^1$ stands for "ϕ is a shared belief between agents i and j, on i's point of view",
  - $frontOf(x,y)$ stands for "x is located in front of y",
  - $basketballCourt(x)$ stands for "x is a basketball court",
  - $partOf(x,y)$ stands for "x is part of y",
  - $studentAt(x,y)$ stands for "y goes or has been at school y".
- Tom's choice should also be explained by the following weaker belief state:
  $Bel_{Tom}(MBel_{Laura,Tom}(frontOf(l,h)$
  $\wedge basketballCourt(h)$
  $\wedge partOf(h,g)$
  $\wedge studentAt(Tom,g)$
  $\wedge studentAt(Laura,g))$
  where $B_i(p)$ stands for "i believes (that) p".

The main assumption behind this kind of approach is the rationality and the cooperativeness of dialogue participants. In addition, to infer from the fact that someone utters that p that she must also believe that p is commonly assumed as a general rule (Lee, 1997). Nonetheless, this assumption is difficult to handle in practice, as J.A. Taylor et al. have shown (Taylor et al., 1996), mainly because of the computational complexity involved. Furthermore, they proved that, in most cases, nested beliefs are not necessary beyond the second level of nesting (ie. what an agent thinks another agent thinks a third agent (possibly

---

[1] See mutual belief's definition in section 3.1

the first one) thinks), as long as deception is not involved. In the particular case of reference, deception may be involved, as the following situation exemplify, and then may require the handling of deeply nested belief.

Tom and Laura live both in Berlin. They lunched at a restaurant called ”Chez Dominique”. Following this meal, one may reasonably assume that:

- $Bel_{Laura}(name(l) =$” Chez Dominique ”),
- $BelTom(name(l) =$ ” Chez Dominique ”),
- And $MBel_{Tom,Laura}(name(l) =$ ” Chez Dominique ”.
  We only treat the particular case of definite reference, which counts as an indication to access a mental representation of the intended referent that is supposed to be uniquely identifiable for the hearer. So, it can be viewed as a result of a function.

Then, Laura left Berlin for two years. During this period, the restaurant changed name. Its new name is ”Restaurant la Petite Maison”. Tom knows it, but Laura does not know it. Thus, the following situation holds:

- $Bel_{Tom}(name(l) =$ ” Restaurant la Petite Maison ”),
- $Bel_{Laura}(name(l) =$ ” Chez Dominique ”).

The return-day Laura and Tom (who did not leave Berlin) must lunch together. They speak by phone in order to agree upon a time and a restaurant. Let’s consider the following exchange between them:

. . .

(U1) Laura: ” Will we lunch at the restaurant where we have been yet ? ”

(U2) Tom: ” Which one ? ”

(U3) Laura: ” *Chez Dominique.* ”

(U4) Tom: ” Ok. ”

. . .

At the end of this talk, a conceptual pact of conceptualizing the restaurant as ”the place called *Chez Dominique*” is established. If we consider that the Conversational Common Ground has to be modelled in terms of mutual belief, the following mutual belief has been formed, at least on Laura’s point of view: $MBel_{Laura,Tom}(name(l) =$ ”Chez Dominique”$)$. Tom’s choice of the referring expression can not be based on Tom’s point of view on the beliefs shared with Laura, because from $MBel_{Tom,Laura}(name(l) =$ ”Chez Dominique”$)$, one may infer, following mutual belief’s definition (ie. 3.1) that $Bel_{Tom}(name(l) =$ ”Chez Dominique ”$)$ which is incoherent with $Bel_{Tom}(name(l) =$ ”Restaurant la Petite Maison”$)$. In fact, Tom’s choice should be explained in terms of his nested belief: $Bel_{Tom}(MBel_{Laura,Tom}(name(l) =$ ”Chez Dominique”$))$ and this is a case of deception.

According to previous work (Saget, 2006), we claim that such a treatment of reference, depending on beliefs of dialogue participants at the first place, which may lead to computational representation and treatment with high complexity, are neither necessary, nor proper. The proper social attitude is Collective Acceptance.

2.2 Collective Acceptance, reference and subordinated activity

Modeling conceptual pacts in terms of belief states implies that the literal description has to be true, or, more precisely, consistent with dialog partners’ beliefs (at least with shared beliefs between dialog partners on addressee’s point of view), in order to ensure their rationality. But the goal of Tom and Laura, in our preceding examples, is to determine a place in such manner that each one identifies it correctly; then, they will be able to meet at the correct meeting-place. Their goal is not to establish the truth with respect to the place in question. Actually, the establishment of conceptual pacts is governed by the ” grounding criterion ” (Clark and Schaefer, 1989): *” The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for the current purpose.*” Thereby, one can establish a conceptual pact in conflict with ones own beliefs, if this pact enables each group member concerned to achieve the current common goal. In the first example, one can imagine that the basketball-court does not exists any more, and that both Tom and Laura know this fact. Tom’s utterance and Laura’s agreement are still realistic.

Collective acceptance of a group of agents, in contrast with belief, may be inconsistent with their beliefs (individual or shared beliefs). In fact, a description is accepted by the addressee if it allows him to identify the intented referent and if an inconsistent description is not an obstacle to the realization of current goals. For example, if Tom have to send a letter to Laura (having two postal addresses) and say: ”Must I send you the letter at 16 Collingham Road, London”. Even if Laura identifies the correct place, the address has to be correct to receive Tom’s letter. Finally, conceptual pact is a temporary and flexible concept, this property does not match with the ideal of integration or agglomeration of beliefs.

How explaining the use of Collective Acceptance ? Generally, this may be partly due to a particular aspect of goal-oriented dialog as a subordinated sub-activity. Goal-oriented dialogues are implied by two interdependent collaborative activities, as explained by A. Bangerter et al.: *”Dialogues, therefore, divide into two planes of activity (Clark, 1996). On one plane, people create dialogue* ***in service of*** *the basic joint activities they are engaged in-making dinner, dealing with the emergency, operating the ship. On a second plane, they manage the dialogue itself-deciding who speaks when, establishing that an utterance has been understood, etc. These two planes are not independent, for problems in the dialogue may have their source in the joint activity the dialogue is in service of, and vice versa. Still, in this view, basic joint activities are primary, and dialogue is created to manage them.”*[2] (Bangerter and Clark, 2003). One of team members’goals is to understand each other, in other words to reach a certain degree of intelligibility, sufficient for the current purpose.

One may distinguish between two kinds of sub-activities: sub-activities which are subparts of another activity (thus, which transcripts the compositionality of basic activities) and sub-activities *in service of* another activity, ie. subordinated (sub-)activities, such as planning, problem solving, interaction with other agents (goal-oriented dialog) and so on. On the logical point of view, the rationality of the involved agents is rendered by a coherent mental state and by the notion of rational action (Cohen and Levesque, 1990; Sadek, 1994). For example, the beliefs and intentions of an agent form a consistent set and agent’s actions are also consistent with his beliefs and intentions. At the first glance, the coherence of action with beliefs seems to be irrefutable. However, to the extend that the success of a subordinated activity is governed by the generalization of the sufficient criterion and on the basis of preceding arguments, one may reasonably assume that agents’ rationality does not strictly imply the coherence between the actions being parts of a subordinated activity and the beliefs states of the involved agents. For these reasons, conceptual pact match better with acceptance and modeling conceptual pacts by collective acceptance insure the rationality of team members. However, agent’s rationality is contingent on the motivational context and on the context of mental states of dialog partners.

2.3 The philosophical notion of Collective Acceptance

Studies on dialog modeling as a collaborative activity address the philosophical problem of determining the type of mental states which could be ascribed to team members. Based on the observation that sometimes one may encounter situations where one has to make judgements or has to produce utterances that are contrary to ones privately held beliefs, philosophers, such has (Cohen, 1992), have introduced the notion of (Collective) Acceptance, which is an intentional social mental attitude. (Collective) Acceptances have the following properties, in contrast with beliefs (Wray, 2001):

- They are voluntary (or intentional);
- They holds on utility or success (thus we can accept something we believe false);
- They does not required justifications;
- All or nothing: we decide to accept or not to accept.

In J.L. Cohen’s famous book, ”An essay on belief and acceptance” (Cohen, 1992), the author argue that the conversational implicature ”a person’s saying that p implies that this person believes p”

---

[2] This claim must be extended to other kind of basic activity such as cooperative or competitive activities.

is not the rule and that speech acts such as concessions, acknowledgements, agreements and admissions that p do not imply the existence of the corresponding belief. In such cases, ”I thereby accept that p” means that ”I take that proposition as a premiss for any relevant decision or argument” (Cohen, 1992). In previous work (Saget, 2006), we claim that an act of reference using a particular description $\imath x.descr(x)$ of an object o does not imply that the speaker believes that $\imath o.descr(o)$ holds, but implies that the speaker believes that this description enables the addressee to pick out the correct intented referent.

3 Formal part

3.1 The dialog model

Rational models, based on (Cohen and Levesque, 1990), can be considered as a logical reformulation of plan-based models. They integrate, in more, a precise formalization of dialog partners' mental states (their beliefs, choices (or desires) and intentions), of the rational balance which relates mental attitudes between them and relates mental attitudes with agents' acts. Moreover, dialogue acts' preconditions and effects are expressed in terms of dialog partners' mental states. Thus, this is hopeful to model precisely mental attitudes.

The chosen model is based on the rational model proposed by D. Sadek (Sadek, 1994), extending (Cohen and Levesque, 1990), which rests upon a set of principles (axiom schemas) of which dialog acts are branched off. A dialog system is considered as a cognitive agent which is rational and have a cooperative attitude towards other agents (as the dialog system's users) and this agent is able to communicate with other agents.

Mental states (beliefs, intentions,...) and actions are formalized in a first-order modal logic. In the following of the paper, the symbols $\neg, \wedge, \vee, \Rightarrow$ stand for the connectors of the classical logic (respectively negation, conjunction, disjunction and implication); $\forall, \exists$ stand for the universal and existential quantificators; p stands for a closed formula denoting a proposition; i, j denote agents and ϕ is a formula schemata. We only need to introduce here two mental attitudes, belief and intention:

$B_i(p)$ stands for ”i (implicitly) believes (that) p”,
$I_i(p)$ stands for ”i intends to bring about p”.

Action expressions can be formed with primitive acts: with $(a_1; a_2)$ which stands for sequential action (where a_1 and a_2 are action expressions) and with $(a_1|a_2)$ which stands for non-deterministic choice.

> $Done(a,p)$: ”a has just taken place, and p was true before that”
> $Done(a) = Done(a, true)$

The model of communicative acts is:

$$< i, TypeOfCommunicativeAct(j, \phi) >$$

> FP: ”Feasible Preconditions”: the conditions which must be satisfied in order to plan the act;
>
> PE: ”Perlocutionary Effect”: the reason for which the act is selected.

For example, the communicative model of ”i informing j that p” is:

$$< i, INFORM(j, \phi) >$$

> FP: $B_i(\phi) \wedge \neg B_i(B_j(p))$
>
> PE: $B_j(\phi)$

In this model, utterance generation and understanding, and thus referential acts are considered as individual acts. Furthermore, the perlocutionary effects are considered as achieved as soon as the communicative act has been performed.

So dialog and reference treatment are not considered as collaborative activities. In order to do so, notably, the set of mental attitudes has to be extended with notions such as collective intention and mutual belief.

There is no consensus on the definition of collaboration. We consider that a group of agents is engaged in a collaborative activity as soon as they share a collective intention.

$CollInt_{i,j}(\phi)$ stands for ”i and j collectively intends to bring about p, on i's point of view”.

$MB_{i,j}(\phi)$ stands for ”ϕ is a shared belief between agents i and j, on i's point of view” and mutual beliefs are formalized as:

$$MB_{i,j}(\phi) \equiv Bel_i(\phi \wedge MBel_{j,i}(\phi))$$

Furthermore, Collective Acceptance have to be included.

3.2 Collective Acceptance

We propose the following formalization of the philosophical notion of Collective Acceptance:

- $CollAcc_{ij}(\phi)$ stands for "ϕ is a collective acceptation between agents i and j, on i's point of view"

- Collective Acceptance is an intentional attitude, ie. it comes from individual acts of involved agents:
  $((\exists \alpha, \beta \in \{i, j\})$.
  $Done(Prop_{\alpha\beta}(\phi)) \wedge Done(Accept_{\beta\alpha}(\phi)))$
  $\Rightarrow CollAcc_{ij}(\phi)$
  where:
  - $Prop_{ij}(\phi)$ stands for "i proposes j to consider ϕ"
  - $Accept_{ji}(\phi)$ stands for "j accepts to consider ϕ (towards i)"
  - $Prop_{ij}(\phi)$ and $Accept_{ji}(\phi)$ are individual actions.

- A proposition involves a social obligation to react:
  $Done(Prop_{i,j}(\phi))$
  $\Rightarrow I_j(Done((Accept_{j,i}(\phi)$
  $|(Prop_{j,i}(\phi') \wedge (\phi' \neq \phi))$
  $|(request_{j,i}(Prop_{i,j}(\phi')) \wedge (\phi' \neq \phi)))))$

Following (Boella et al., 2000), we consider that social obligations as pro-attitudes are not required and that an anticipatory coordination takes place on the speaker's point of view. This phenomenon is govern by a social rule, acquired during preceding social interaction. This social rule is transcribed by repeated use through a reaction to the realization of a particular action (on the speaker's point of view) and through a reaction to the observation of an event which is the occurrence of a particular action (on the addressee's point of view). Since, reaction is a unintentional action, we have to extend the kind of action of the basic model. In fact, this model only considers what we name *intentional actions*. Intentional actions of an agent are those generated by a chain of intention, in our model they are generated by the activation of the rational axiom (Sadek, 1994):

$I_i(p) \Rightarrow I_i(Done(a_1 \vee \cdots \vee a_n))$
The intention of an agent, to achieve a given goal, generates the intention that one of the acts, which satisfies the following conditions, be performed:

1. $(\exists x) B_i(a_k = x) \equiv Bref_i(a_k)$: the agent i knows the action a_k,
2. $EP_{a_k} = p$ and
3. $\neg I_i(\neg Possible(Done(a_k)))$

Reactions have to be added. Reactions of an agent are defined as those generated by the activation of such axiom:

$\phi \Rightarrow I_i(Done(a_1 \vee \cdots \vee a_n))$
where ϕ is the result of the perception of an event or an action's occurrence.

4 Model of Reference as a collaborative activity

4.1 Model of Referential Act

In order to model dialog as a collaboration, reference treatment has to be considered at the speech act level (Cohen and Levesque, 1994), as it is done in A. Kronfeld's work (Kronfeld, 1990).[3]

In order to integrate Collective Acceptance in reference, we propose an extension of an existing model of referential acts based on A. Kronfeld's work in the rational model used (Bretier et al., 1995). The act of reference from an agent i to another agent j, using the conceptualization x (which corresponds to the semantics of the referential expression) to refer to the object y is formalized as:

$< i, REFER(j, x, o) >$

FP: $I_i(refer_{i,j}(o)) \wedge Bref_i(o)$;

EP: $B_j((\exists o) I_i(refer_{i,j}(o)))$
$\wedge I_j(Bref_j(o))$
$\wedge RepSameObj(o, o')$
$\wedge Done(Prop_{i,j}(referBy(x, o)))$
$\wedge B_j(Done(Prop_{i,j}(referBy(x, o))))$.
where:

- o et o' are object mental representations;
- $I_i(refer_{i,j}(o))$ stands for "a communicative intention of i to refer to o, the addressee is j";

---

[3] For a computational implementation is provided in (Jorgensen, 2000).

  - $RepSameObj(o, o')$ stands for "the mental representations o and o' represent the same object";
  - $referBy(D, R)$ stands for "the description D refers to the referent R".

Generating a referential expression is considered as the generation of an instance of such plan and the interpretation of a particular referential expression as the recognition of an instance of such plan. And the whole process is governed by two meta-goal, on the speaker's point of view [4]:

$$CollInt_{ij}(MB_{ij}(I_i(refer_{i,j}(o)))) \wedge$$
$$CollInt_{ij}((\exists D) CollAcc_{ij}(referBy(D, o)))$$

4.2 Return to the example

Let's consider the example shown in 2.1, the task level and the conversational level have to be separated [5]. In uttering (U1), Tom wants to make a necessary choice for the meeting task, such as :

$$I_{Tom}((\exists l) MBel_{Tom,Laura}(meetingPlace = l))$$

Tom makes his choice: his mental representation of the restaurant chosen is o. In order to realize his preceding intention, he has get through to Laura:

$$I_{Tom}(refer_{Tom,Laura}(l))$$

Remaining the goal of referential acts (2.1), the choice of the description of the intented place is guided by its capacity to enable Laura to pick out, *in her mental state,* the mental representation of the correct place. That is, the description enables Laura to isolate the correct mental representation from other possible ones, with sufficient evidence of mutuality. This is a pragmatic (ie. contextual) guideline, which corresponds to the Identification goal.

Thus, Tom produces a description of the intended place: *"the restaurant where we have been yet"*. He thinks that Laura is able to identify the correct place basing on the description, ie. he thinks that she is able to realize the following intention:

Identification task:

$$I_{Laura}(Bref_{Laura}(l') \wedge RepSameObj(o, o'))$$

But, Laura is not able to pick out a single place: there is other restaurants, where they have been together. Moreover, Laura has to answer to Tom's proposition:

$$B_{Laura}(Done(Prop_{Tom,Laura}(referBy(\imath x.\phi(x), l')))).$$

She is obliged to reply to his proposition by the social rule. Besides, the precondition of accepting a conceptual pact is to have realized the Identification goal; otherwise, the addressee has the choice between other possible reactions. As Laura failed to succeed, she chooses to ask for clarification in (U2):

$$request_{Laura,Tom}$$
$$(Prop_{Tom,Laura}(referBy(\imath x.\phi'(x), l'))).$$
$$\wedge(\phi' \neq \phi)$$

In order to achieve understanding, by a cooperative attitude, Tom realizes Laura's request in (U3). Laura is now able to pick out a single mental representation of the place. She likes it, so she agrees. The social goal obliges Laura to react to Tom's new proposition. As the precondition of accepting is fulfilled, with uttering (U4), Laura realizes the following intention:

$$Done(Accept_{Laura,Tom}(referBy(\imath x.\phi'(x), l'))).$$

Finally, following Collective Acceptance's definition, a conceptual pact is created:

$$CollAcc_{Laura,Tom}(referBy(\imath x.\phi'(x), l').$$

As well as, mutual understanding:

$$MB_{Laura,Tom}(I_{Tom}(refer_{Tom,Laura}(l')),$$

and the coordination on the task level:

$$MBel_{Laura,Tom}(meetingPlace = l')).$$

5 Conclusion

Modeling dialog as a collaborative activity consists notably in specifying the content of the Conversational Common Ground and the kind of social mental state involved. Even if mutual beliefs, or weaker forms of belief states, do not rise to inconsistencies, but, are still sufficiently strong for the participants to have successful cooperation or coordination of actions. Epistemic states involve computational treatments with high complexity.

We show that modeling the CCG by an epistemic state is neither necessary, nor proper.

---

[4] On the addressee point of view, it is govern by dual goals with existential quantifier.

[5] Further details may be found in(Saget, 2006).

Considering only genuine conceptual pacts limits the capacity of interaction and may leads to ”real” communicative errors.

We have proposed a formalization of Collective Acceptance, furthermore, elements haven been given in order to integrate this attitude in a rational model of dialog. Finally, a model of referential acts as being part of a collaborative activity has been provided.

Further studies will hold on the extension of the general principles proposed to the dialog itself. Moreover, collective acceptance is a particularly interesting attitude because it allows to model reference and dialog itself as situated activities in an elegant manner. Finally, this concept may provide symbolic elements in order to form the grounding criterion, which is a notion especially hard to make up, because this criterion is highly context dependant. Grounding criterion differs depending on the people involved, the domain concerned and so on.

Acknowledgment

This work is partially financed by the grant A3CB22 / 2004 96 70 of the regional council of Brittany.

References

A. Bangerter and H. H. Clark. 2003. Navigating joint projects with dialogue. *Cognitive Science*, (27):195–225.

G. Boella, R. Damiano, and L. Lesmo. 2000. Social goals in conversational cooperation. In *Proceedings of the First SIGdial Workshop on Discourse and Dialogue*, pages 84–93. Association for Computational Linguistics, Somerset, New Jersey.

S. E. Brennan and H. H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:482–1493.

P. Bretier, F. Panaget, and M.D. Sadek. 1995. Integrating linguistic capabilities into the formal model of rational agent : Application to cooperative spoken dialogue. In *AAAI-95, Fall Symposium of Rational Agency*, Stanford, MA.

H. H. Clark and A. Bangerter. 2004. Changing conceptions of reference. In I. Noveck & D. Sperber, editor, *Experimental pragmatics*, pages 25–49. Palgrave Macmillan, Basingstoke, England.

H. H. Clark and C. R. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of discourse understanding*, pages 10–63. Cambridge University Press, Cambridge.

H.H. Clark and E.F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

H.H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. In *Cognition*, volume 22, pages 1–39. The MIT Press.

H. H. Clark. 1996. *Using Language*. Cambridge: Cambridge University Press.

P.R. Cohen and H.J. Levesque. 1990. Rational interaction as the basis for communication. In *Intentions in Communication*, pages 221–256. MIT Press, Cambridge, MA.

P.R Cohen and H.J. Levesque. 1991. Confirmations and joint actions. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-91)*, pages 951–957.

P.R. Cohen and H.J. Levesque. 1994. Preliminaries to a collaborative model of dialogue. *Speech Communication*, 15:265–274.

J. Cohen. 1992. *An Essay on Belief and Acceptance*. Oxford University Press, Oxford.

S. Garrod and M.J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8:8–11.

S.W. Jorgensen. 2000. *Computational reference - An investigation, Development and implementation of Kronfeld’s Theory of reference*. Ph.D. thesis, Copenhagen Business School.

G. Klein, P.J. Feltovitch, and D.D. Woods. 2005. Common ground and coordination in joint activity. In *Organizational simulation*. Wiley, New York.

A. Kronfeld. 1990. *Reference and Computation : An Essay in Applied Philosophy of Language*. Cambridge University Press.

M. Lee. 1997. Rationality, cooperation and conversational implicature. In *Proceedings of the Ninth Irish Conference on Artificial Intelligence*.

M.D. Sadek. 1994. Communication theory = rationality principles + communicative act models. In *AAAI-94, Workshop on Planning for Interagent Communication*, Seattle, Washington, USA.

S. Saget. 2006. In favour of collective acceptance : Studies on goal-oriented dialogues. In *Proceedings of Collective Intentionality V (to appear)*, Helsinki, Finland.

J.A. Taylor, J. Carletta, and C. Mellish. 1996. Requirements for belief models in cooperative dialogue. *User Modeling and User-Adapted Interaction*, 6(1):23–68.

K.B. Wray. 2001. Collective belief and acceptance. *Synthese*, 129:319–333.

Scorekeeping in an Uncertain Language Game

David DeVault[1]
[1]Department of Computer Science
Rutgers University
Picataway, NJ 08845-8020
`David.DeVault@rutgers.edu`

Matthew Stone[1,2]
[2]Human Communication Research Centre
University of Edinburgh
Edinburgh EH8 9LW, UK
`Matthew.Stone@rutgers.edu`

Abstract

Received views of utterance context in pragmatic theory characterize the occurrent subjective states of interlocutors using notions like common knowledge or mutual belief. We argue that these views are not compatible with the uncertainty and robustness of context-dependence in human–human dialogue. We present an alternative characterization of utterance context as objective and normative. This view reconciles the need for uncertainty with received intuitions about coordination and meaning in context, and can directly inform computational approaches to dialogue.

1 Introduction

The question we address in this paper is how utterance context should by represented in implemented conversational systems. Strong intuitions about coordination in conversation (Clark and Marshall, 1981) have led many researchers, e.g. (Traum, 1994; Poesio and Traum, 1997; Rich et al., 2001; Blaylock, 2005), to aim to represent the common ground beliefs that seem to guarantee principled coordination between speaker and hearer on each new utterance (Lewis, 1969; Stalnaker, 1974). Other researchers, in pursuit of robust implementations for real-world dialogue, have opted to represent narrower aspects of interlocutor and conversational state using models that afford a straightforward treatment of uncertainty (Roy et al., 2000; Horvitz and Paek, 2001; Gruenstein et al., 2004).

These differences might seem to be a matter of emphasis rather than substance. In fact, however, the notion of uncertainty about the context is profoundly at odds with received views of context both in theories of presupposition in formal pragmatics (Stalnaker, 1974; Poesio and Traum, 1997) and theories of coordinated activity in AI (Cohen and Levesque, 1991; Grosz and Kraus, 1996; Blaylock, 2005). As we will argue, this tension originates in the central role these theories grant to various nested or higher-order beliefs that interacting agents may have about each other. In Section 2, we review both the rationale for defining utterance context in terms of these beliefs and the challenges that doing so poses to system builders. The contribution of this paper is a new characterization of utterance context which overcomes these challenges by narrowly circumscribing the aspects of interlocutor mental states that are necessary for utterance interpretation. We present this new characterization in Section 3. The discussion in Section 4 shows how this view of context reconciles the practical need for uncertainty with received intuitions about coordination and meaning in context.

2 Context and mutual belief

We use the term *utterance context* to label the body of information used in utterance interpretation, including both grammatical conditions required for the utterance to be meaningful and situational factors required to analyze the utterance as a contribution to interlocutors' ongoing joint activity. It is commonly assumed that this information must be mutually believed[1]; see, e.g., Stalnaker (1998). One of the first and most widely known definitions of mutual belief is due to Schiffer (1972). The definition records an infinite, hierarchical interrelation between the private beliefs of a speaker *S*

---

[1] or some analogous status of mutual knowledge, mutual supposition, etc.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 139–146.

and a hearer H about some proposition p:

$$\mathtt{MB}_{S,H}\ p =_{\mathrm{def}} \begin{array}{lll} & \mathtt{B}_S\ p & (a) \\ \wedge & \mathtt{B}_H\ p & (b) \\ \wedge & \mathtt{B}_S\mathtt{B}_H\ p & (c) \\ \wedge & \mathtt{B}_H\mathtt{B}_S\ p & (d) \\ \wedge & \mathtt{B}_S\mathtt{B}_H\mathtt{B}_S\ p & (e) \\ \wedge & \mathtt{B}_H\mathtt{B}_S\mathtt{B}_H\ p & (f) \\ \ldots & & \end{array} \qquad (1)$$

The modal operators $\mathtt{B}_S$ and $\mathtt{B}_H$ represent the beliefs of S and H, respectively.

The rationale for defining utterance context as the set of mutually believed propositions is theoretical. For example, we know that an agent that *has* mutual belief with its partner can avoid certain errors in solving coordination problems (Lewis, 1969), in interpreting definite references in conversation (Clark and Marshall, 1981), and in participating in multi-agent collaborations (Cohen and Levesque, 1991; Grosz and Kraus, 1996). In conversation more generally, it is argued that an agent that interprets utterances against the "common ground" of mutual belief can better avoid potential misunderstandings (Clark, 1996).

Yet the mutual belief view of context poses several serious challenges for system builders. First, it is unclear what implications arguments about the role of mutual belief in coordination have, if any, for *representation*. If agents sometimes need to *have* mutual attitudes, must their context representations therefore *describe* mutual attitudes? Of course not: agents might not be coordinating using only their context representations. Even if they are, there's still a huge gap between the *conditions* rational system behavior depends on (e.g., that a successful agent acts in accord with what is mutually believed) and the *meaning* of the underlying representations (e.g., that an agent's representations directly track what is mutually believed). See Dennett (1989). Nevertheless, it's common to assume that dialogue context representations *should* track the mutual beliefs of agents and their interlocutors—see, for example, (Traum, 1994; Poesio and Traum, 1997; Rich et al., 2001; Blaylock, 2005).

This leads immediately to the second problem, the methodological challenge of correctly identifying what is mutually believed, so that utterance context can be implemented correctly. It is relatively straightforward to implement a theoretically sound formalism for mutual belief in dialogue. However, it is another matter to enable such an implementation to model its conversations accurately. In the absence of any available "ground truth"—such as an utterance-by-utterance trace, for each conversation in a corpus, of empirically observed higher-order attitudes—we have no way to tell whether (1) does or should hold.

For this and other practical reasons, some researchers endorse a weaker notion than mutual belief for context representations in dialogue. For example, Taylor, Carletta and Mellish (1996) argue that we can eschew the indefinite nesting suggested by (1) in favor of a bounded depth of three. More commonly, researchers suggest that context representations should be relativized to a specific perspective (Ginzburg, 1996), so that they track nested information such as $\mathtt{B}_S\ \mathtt{MB}_{S,H}\ p$ or $\mathtt{B}_H\ \mathtt{MB}_{S,H}\ p$ or both. Note that such changes undermine one of the key virtues of the mutual belief view: its explanation of why reasoning about context helps interlocutors solve coordination problems. As long as there is any asymmetry across interlocutors, we can apply constructions like Clark and Marshall's "Roxy" scenario (1981) to show that coordination failure is possible.[2] In any case, even weakened versions of mutual belief still require higher-order beliefs like (1c,d) and (1e,f). And there is insufficient evidence for an analyst to make principled decisions *herself* about whether such beliefs obtain, much less automate these decisions.

A third challenge for treating context as mutual belief lies in cases where utterance interpretation felicitously exploits information one interlocutor lacks. For example, Kaplan (1989) argues that, regardless of interlocutors' information states, use of the noun phrase *that* refers to whatever the speaker designates with the accompanying demonstration. The correct interpretation, therefore, reflects what was actually designated, even when this differs from what the speaker believes was designated. Similarly, Gauker (1998) presents a hearer-independent explanation for the "informative presuppositions" of factive verbs like *regret*. For Gauker, *We regret that tonight's show is canceled* is felicitous because it requires for its meaningfulness only the *fact* that the show is canceled, not *mutual belief* between speaker and hearer that the show is canceled.

---

[2] Of course, coordination failure does sometimes occur in human dialogue, so this certainly does not rule out notions of context that differ from mutual belief.

A final challenge, which we will particularly emphasize in Section 4, comes in characterizing problematic communication on the mutual belief model. It has been common for discrepancies between the contexts believed to obtain by two interlocutors to be marginalized.[3] Yet in computational models of interpretation, some degree of uncertainty about what an utterance means is the *norm*, so discrepancies are unavoidable. When discrepancies do arise, interlocutors often seem to know they *lack* mutual belief, but manage to communicate with context-dependent language anyway. In the next two sections, we present a view of context that explains this capacity in terms of the interlocutors' uncertainty about a true context, and that answers the other challenges as well, while at the same time maintaining the intuitions about coordination that have historically made a higher-order attitude model of utterance context attractive.

3 Objective, normative context

The two basic principles in our characterization are that utterance context is *objective* and that it is *normative*. By *objective*, we mean that *there is a fact of the matter about what the context is at each time t in a conversation, and this context is not a function of the interlocutors' beliefs at time t.*[4] Thus, context is not a matter of one or the other interlocutor's perspective on the situation, and nor is it an interaction between their combined perspectives. Instead, the objective context is the *product of action* taken by the individual interlocutors at times $t' < t$. Agent mental states still play a role, but this role is limited to classifying actions as interlocutors intend them. Action-based characterizations of context have also been advanced on higher-order attitude views of context (Thomason, 1990; Poesio and Traum, 1997), but the presence of higher-order attitudes in these models creates all the challenges discussed in Section 2.

By *normative*, we mean that the job of interlocutors' context representations is to target the objective context as it really is. While systems might aim to *achieve* mutual belief to avoid misunderstanding, their context representations, we will argue, should not *mean* that propositions are mutually believed. They should mean simply that those propositions are *true* in the objective context.

More generally, we believe that *all* interactions where people coordinate by following conventional social rules give rise to an objective, normative context. A clear case is correspondence chess, where players send moves by email. Normally, we might expect each player to keep track of the game by moving pieces on a physical chessboard, keeping the board in sync with their moves. But actually two ambitious players could use only their emailed moves and their imaginations to play chess. In what follows, we adapt Lewis's (1979) scorekeeping metaphor to this case: we use such mental chess to develop a vocabulary for describing context as the abstract product of coordinated activity (Section 3.1), show how this vocabulary applies to dialogue (Section 3.2) with its much more complex and open-ended conventions and context, and use a case of misunderstanding to show how this vocabulary differs from models based on mutual belief (Section 3.3).

3.1 Context as a product of action

We can treat the state or *context* of a chess game as an abstract structure $c = \langle t, s_1, s_2, ..., s_{32}, h \rangle$ recording whose turn t it is to move next (one or the other of the players), the current status s_i of each of the 32 chess pieces (piece type and position—either some board position or "captured"), and limited historical information h (e.g. whether certain pieces have ever moved). Let us write c_t for the context at time t, and let the initial context c_{t_0} be the starting configuration for a game of chess.

In chess there is a set $\mathcal{A}$ of possible moves or *action types*, which we might formalize parametrically as $\mathcal{A} = \{\texttt{advancePawnOneStep}(P), \texttt{moveQueen}(Q, Pos), \texttt{castle}(R), ...\}$. Each move a is $\sigma(\alpha)$ where $\alpha \in \mathcal{A}$ and σ instantiates the free parameters of α. Doing a effects a *deterministic* transformation on the current context. We can formalize this by way of an update function:

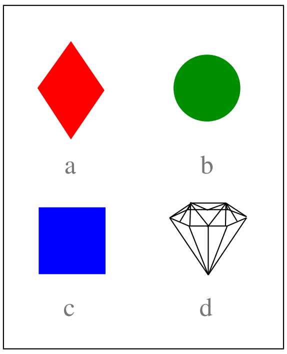
$$c_{t+1} = \texttt{update}(a, c_t) \qquad (2)$$

One goal of each participant in a mental chess game, then, is to track the evolving context c_t as a stream of chess moves $\langle a_1, a_2, ... \rangle$ plays out over email messages.

---

[3] This trend goes all the way back to the first formal model of context, that of Stalnaker (1978). Stalnaker calls each speaker's private context model *nondefective* if it coincides with that of his interlocutor, and suggests that this be treated as the normal case.

[4] When we say context is *objective*, we don't mean to suggest that context is visible, or easily definable in the language of physics, or even that it can be defined independently of human minds and purposes. The point is just that context is not determined by what the interlocutors are currently, privately thinking.



C_1: *the target is a diamond*

U_2: *you mean a diamond jewel?*

C_3: *yeah*

U_4: *ok*

Figure 1: User interaction with the COREF agent. The user (U) can see the four displayed objects, but not COREF's (C) private labels {`a`,`b`,`c`,`d`} for them. The target in this example is object `d`.

We maintain that the evolving context c_t in such a mental game of chess is objective, and that even though the current context is not physically realized (on a chess board, for example), it would be very *misleading* to define it in terms of the players' beliefs at time t. The context is objective because, at each time t, the context c_t is an abstract structure that is well-defined given the sequence of moves $\langle a_1, a_2, ..., a_{t-1} \rangle$ that have been exchanged by email. It would be misleading to define c_t by way of the players' beliefs about it because their individual beliefs may manifest any number of errors: one or even both players may have forgotten or misunderstood where one piece or another is, whether a knight has been captured, etc.[5] If we were to model the chess context by way of the beliefs (or mutual beliefs) of the players, our model would capture more of the players' *perspectives*, but it would obscure the objective status of the underlying game, and it would hide the normative role played by the true state as players improve their chess skills, recover from mistakes, and cope with their private uncertainties.

3.2 Utterance context and intended actions

We illustrate our approach to utterance context using COREF, an implemented dialogue system that collaboratively identifies visual objects with human users (Thomason et al., 2006). Figure 1 shows an excerpt of an interaction with COREF. COREF is designed to participate in *collaborative reference* (Clark and Wilkes-Gibbs, 1990), in which human interlocutors come to agree on a target object through an interactive, multi-utterance dialogue involving linguistic expressions of heterogeneous form and function.

We understand utterance context as an abstract, objective structure, analogous to the chess state, but now populated by the familiar attributes of dialogue state: sets of propositions contributed to the conversational record, plans that are underway, outstanding interlocutor obligations, linguistic forms of prior utterances, etc. The state depends on what interlocutors are doing. In COREF's domain, we have found that dialogue context takes the form $c^* = \langle R, P, T, C, U \rangle$, where R is a set of referents yet to be identified, P is a set of agreed propositions, T is a stack of tasks (where each task specifies what actions can occur next), C is a set of constraint networks (one for each target referent), and U is the universe of discourse (a set of properties and objects).

Such an utterance context evolves over the course of the dialogue through the domain-dependent set of action types, $\mathcal{A}$, that interlocutors take. The COREF action set $\mathcal{A}^*$ includes actions that select the referent sequence, initiate collaborative reference to a particular target referent, add a constraint C to the constraint network for a target (`addcr`(C)), mark a target as identified, initiate a clarification subtask, and inquire whether some action can be taken. Each action $a = \sigma(\alpha)$ for $\alpha \in \mathcal{A}$ has a *deterministic* effect on the current context, which we again capture by an update function as in (2). This way, we can implement an update mechanism (Larsson and Traum, 2000) that tracks the *objective* context—taking idealized representations of linguistic interpretation, as in (Asher and Lascarides, 2003), and using them for on-line processing, *without* analyzing updates in terms of higher-order attitudes as Poesio and Traum (1997) do.

A key feature of COREF is that the context engendered by these actions *is not mutually believed*. COREF dialogues, unlike chess, include *tacit actions* that allow one interlocutor unilaterally to update the context in ways the other does not know about. These moves allow COREF to handle grounding, clarification, task progress and problem-solving robustly with a model that applies symmetrically in understanding and generation and avoids exceptional pragmatic processes of accommodation or implicit repair. In understanding, when a speaker S utters a linguistic form l, we

---

[5] In case of a dispute, if the email logs were available, the actual chess state could be settled by examining the move history. This would show who was right.

view S as intending to signal both *what the current context is* and *how it should be updated.* If the last utterance resulted in context c_t, the next utterance should be interpreted in a new context c_{t+n} that incorporates the results of some tacit action sequence $A_l = \langle a_1, ..., a_n \rangle$:

$$c_{t+n} = \texttt{update}(a_n, \texttt{update}(..., \texttt{update}(a_1, c_t)...))$$

The utterance then signals a_{n+1} and creates context c_{t+n+1}. For example, in interpreting the user's utterance *you mean a diamond jewel?*, U_2 in Figure 1, COREF interprets the user as signaling the following sequence:

1. initiate a clarification subtask,
2. start collaborative reference targeting COREF's intended property `P`, (3)
3. inquire whether to take action `addcr(equals(P,diamondJewel))`

Only the last of these is directly associated with the semantics of the utterance. After interpretation, COREF updates its model of the context to reflect these actions.

3.3 Representing the ground truth context

It is easiest to highlight where our characterization of context differs from mutual belief with a case of misunderstanding. Consider the COREF dialogue excerpt D_1–M_2–D_3 presented in Figure 2. The figure tracks the evolution of the context, under both objective and mutual belief characterizations, in a case of misunderstanding. D begins with the red rhombus, i.e. object `a` at the top left of Figure 1, as the value of a target variable `t`. Within this domain, *diamond* can mean either `rhombus` (as in card games) or `diamondJewel` (as in jewelry stores). D utters D_1, *the target is a diamond.* While D intends action `addcr(rhombus(t))`, as it happens, M interprets D as doing `addcr(diamondJewel(t))`.

What happens, we argue, is that after D_1, the intended action `addcr(rhombus(t))` takes its objective effect. D knows what his intended action was, so D updates his model of the context *correctly*. M however comes to believe *erroneously* that `diamondJewel(t)` is in the context. By contrast, if context is mutual belief (or any higher-order attitude), the misunderstanding keeps both `rhombus(t)` and `diamondJewel(t)` out of the context. So *both* D and M are mistaken: D believes it mutually believed that `rhombus(t)`, as D intended, while M believes it mutually believed that `diamondJewel(t)`, as M interpreted.

These then are the basic facts about ground truth and the agents' representations thereof on the two views of utterance context. In the next section, we will use this example to assess the merits of the objective view for system building.

4 Discussion

In designing a representation of context, system builders should be able to explicate their agents' representations and inference in terms of the events in the dialogue on the one hand and the meanings of the representations on the other. Section 2 posed four challenges that make this difficult when context is construed as mutual belief. Our characterization of context allows system builders to meet each of them. For ease of presentation, we take them up in reverse order.

4.1 Miscommunication and uncertainty

The challenge of representing context in the presence of miscommunication and uncertainty is well-illustrated by the example of Figure 2. When D says D_3, *the diamond is red*, M will detect a problem, because while the context appears to M to describe the target as a `red diamondJewel`, there is no such object. Upon detecting the problem, M can reinterpret D_1 and thus correct his private model of the objective context: M had at first thought the context was $[c_2]$ `diamondJewel(t)` whereas M now recognizes that the true, objective context was $[c_2]$ `rhombus(t)`. This allows D_3 to be interpreted as meaning that the target `rhombus` is `red`, as intended. Because context is normative, utterances can be seen as contextually acceptable iff they are interpretable in the true context. This allows interlocutors, like M here, to reason "backwards" from a presumably acceptable intended interpretation to what the true context must be.

Compare the mutual belief model, where the true context before D_3 does not include `rhombus(t)`, because that isn't mutually believed before D_3. On this model, although M did have an erroneous representation of the context before D_3, fixing *that* error does not help to interpret D's utterance. When M discovers what is mutually believed, it's that *nothing* is mutually believed. This correction neither remedies the misunderstanding of D_1 nor makes D_3 interpretable. Thus the ground truth about mutual belief cannot play the simple

	D_1: *the target is a diamond*	
D intended:	`addcr(rhombus(t))`	
M interpreted:	`addcr(diamondJewel(t))`	
	objective context	**mutual belief**
update:	$c_2 = \texttt{update}\big(\texttt{addcr(rhombus(t))}, c_1\big)$	(mental events)
ground truth:	$[c_2]$ `rhombus(t)`	$\mathrm{MB}_{D,M}$ (nothing about `t`)
D private:	$\mathrm{B}_D[c_2]$ `rhombus(t)`	$\mathrm{B}_D\mathrm{MB}_{D,M}$ `rhombus(t)`
M private:	$\mathrm{B}_M[c_2]$ `diamondJewel(t)`	$\mathrm{B}_M\mathrm{MB}_{D,M}$ `diamondJewel(t)`
	M_2: *ok*	
	(no change from c_2 to c_3)	(no change)
	D_3: *the diamond is red*	
	objective context	**mutual belief**
update:	$c_4 = \texttt{update}\big(\texttt{addcr(red(t))}, c_3\big)$	(mental events)
ground truth:	$[c_4]$ $\texttt{rhombus(t)} \wedge \texttt{red(t)}$	$\mathrm{MB}_{D,M}$ `red(t)`
D private:	$\mathrm{B}_D[c_4]$ $\texttt{rhombus(t)} \wedge \texttt{red(t)}$	$\mathrm{B}_D\mathrm{MB}_{D,M}$ $\texttt{rhombus(t)} \wedge \texttt{red(t)}$
M private(?):	$\mathrm{B}_M[c_4]$ $\texttt{diamondJewel(t)} \wedge \texttt{red(t)}$	$\mathrm{B}_M\mathrm{MB}_{D,M}$ $\texttt{diamondJewel(t)} \wedge \texttt{red(t)}$

Figure 2: A misunderstanding in COREF's domain. D is the director (the initiator of reference) and M is the matcher. The visual display is as in Figure 1. We write $[c_t]$ p to mean p is part of context c_t.

D_1:	*the target is a diamond*
D intended:	`addcr(rhombus(t))`
M interpreted:	
$p = 0.6$	`addcr(diamondJewel(t))`
$p = 0.4$	`addcr(rhombus(t))`
M_2 :	*ok*

Figure 3: A probabilistic misunderstanding.

normative role that the objective context does.[6]

The normative role of context also allows an agent to employ straightforward statistical reasoning to cope with its uncertainty. Concretely, suppose M assigns probabilities to alternative interpretations, as illustrated in Figure 3. In this example, M is sufficiently certain of D's intention to proceed with *ok* in M_2. On our model, there is no impediment to treating M's private model of the context after M_2 as:

$$\begin{aligned} P\big([c]\ \texttt{diamondJewel(t)}\big) &= 0.6 \\ P\big([c]\ \texttt{rhombus(t)}\big) &= 0.4 \end{aligned} \tag{4}$$

The probabilities capture M's uncertainty about how D's intended action in D_1 changed the context. And M can exploit evidence across multiple utterances, such as when M realizes his misunderstanding after D_3, to reduce uncertainty about the true context. It's just Bayesian inference.

On the mutual belief approach, however, there seems to be no practical route to a useful internal model of context analogous to (4). Let's assume, for exposition, that beliefs and higher-order beliefs are all bivalent. Then upon hearing D_1 in Figure 3, M must choose what to believe. Suppose M chooses to believe `diamondJewel(t)`, and further to believe `diamondJewel(t)` is mutually believed. Then M must assign $P(\mathrm{MB}_{D,M}\ \texttt{rhombus(t)}) = 0$: *$M$'s own lack of belief rules it out!* M then ends up with this private model:

$$\begin{aligned} P\big(\mathrm{MB}_{D,M}\ \texttt{diamondJewel(t)}\big) &= 0.6 \\ P\big(\mathrm{MB}_{D,M}\ \texttt{rhombus(t)}\big) &= 0.0 \\ P\big(\mathrm{MB}_{D,M}\ \texttt{(nothing about t)}\big) &= 0.4 \end{aligned}$$

The model frustrates future Bayesian inference: D's intended context is ruled out, while the seemingly irrelevant "no mutual belief" scenario remains. In fact, no matter how we take uncertainty into account,[7] M's uncertainty after D_1 is not well summarized as uncertainty about occurrent mutual beliefs with D; M's problem, as Figure 3 suggests, lies in M's *own* belief state—to which M has complete introspective access. Reducing uncertainty about mutual beliefs does not solve this problem; reducing uncertainty about objective context does.

---

[6] To respect the mutual belief semantics, we must construe M's misunderstanding recovery process at best as one of constructing counterfactual sets of mutual beliefs, sets which could have been actual if certain private mental events had occurred that did not. For example, McRoy and Hirst (1995) can make repairs only by explicitly abducing specially-axiomatized *events* of misunderstanding.

[7] E.g., even if M somehow overcame the hopeless task of assigning meaningful probabilities to *all* the beliefs in (1).

4.2 Felicitous use of non-mutual information

The interaction in Figure 2 also illustrates the ubiquity of utterances that seem perfectly acceptable, yet exploit for their interpretation facts that are not mutually believed. Utterance D_3, *the diamond is red*, is such a case here. Objective context captures such utterances directly. Utterance D_3 is contextually acceptable because its contextual requirement `rhombus(t)` is in fact satisfied in the objective context, despite *M*'s misrepresentation of that context. On the mutual belief view, however, D_3 looks like a case of *presupposition failure* given the ground truth context, and a special explanation is required for why the utterance is felicitous and how it changes its context.

4.3 Identifying the true context

The mutual belief model poses the challenge of identifying in practice what the correct set of mutual beliefs is at any given time. In comparison, our model defines the objective context directly in terms of the interlocutors' prior communicative intentions. As illustrated in (3), modeling communicative intentions within an application domain requires connecting words to desired domain entities like ongoing subtasks, intended referents, and domain actions. Of course, we need such a model anyway—for example, even to accurately characterize the potential for a misunderstanding like that in Figure 2. Fortunately, an external observer can construct such a model by examining the utterances that interlocutors use as they perform real-world tasks, without access to their higher-order attitudes. Thus, our approach to context exploits representations that are independently necessary and situates the facts about context much closer to empirical observations than are the facts about higher-order attitudes.

4.4 Coordination and context

Perhaps the hardest challenge in representing context is understanding how a representation should fit into a more abstract characterization of collaboration. While representing mutual beliefs directly seems to preclude certain errors in collaboration, there may of course be other representations that allow an agent to collaborate equally successfully, or at any rate, effectively enough. From this perspective, we can consider agents that try to represent the *objective* context in two cases: ideal communication, and cases of miscommunication and/or uncertainty. In ideal communication, every utterance is actually understood exactly as intended, and both speaker and hearer are perfectly certain that this is so. In this case, not only does each interlocutor privately track the objective context correctly, but each is certain that the other does as well, and further that the other is certain that *they* do, and so on. Provided the speaker and hearer are non-deceptive and trust each other, they will achieve mutual belief.[8]

In cases of miscommunication or uncertainty, their private representations of objective context will differ, and mutual belief will not generally obtain. However, each interlocutor will have a clearly interpretable, practical, uncertain representation of what their prior communicative intentions have been. This means the interpretations they assign to utterances in context will be defensible in terms of these prior intentions. In our view, this highlights interlocutors' ability to target the utterance context implicitly established by their prior conversational activity and to work to make contextual information mutually believed. Thus we can see mutual belief as a desirable but contingent *outcome* of the interlocutors' interaction, rather than as a precondition for it, or as the moment-to-moment target of their representations (Thomason et al., 2006).

Compare this perspective with recent work by researchers pursuing robust human–machine dialogue, who have found it practical to simply identify "context" with the *user*'s state (Roy et al., 2000; Horvitz and Paek, 2001). While this enables coherent probabilistic reasoning, it abandons the role of context as a grammatical resource linking meaning to interpretation and as a mechanism for coordinating dialogue. Our view shows how to keep intuition and implementation aligned.

5 Conclusion

The view of utterance context we have proposed yields simpler representations and reasoning than does the mutual belief model of context. At the same time, it enables straightforward statistical reasoning about context, and offers clearer guidance about what context representations a practical system should have, and how to develop them.

In the end, of course, an interlocutor's uncertainty is pervasive: it affects not only the interpretation of individual words, but also the games

---

[8] or mutual supposition, etc.

(like "collaborative reference") that other interlocutors play, including the contextual actions those games contain. Fortunately, by connecting utterance interpretation to the objective effects these games and actions have on the context, a language speaker can exploit linguistic experience to reduce uncertainty about them. Interlocutors try, in concert with their other goals, to minimize uncertainty and avoid misunderstandings. When they succeed, mutual belief may be achieved. But by adopting an objective view of context, we can understand how interlocutors proceed on sound footing in any case, and can more transparently design systems that will do the same.

Acknowledgments

This work was supported by the Leverhulme Trust, Rutgers University and NSF HLC 0308121. We thank our anonymous reviewers, Alex Lascarides, Chung-chieh Shan, Mark Steedman, and Rich Thomason.

References

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge.

N. J. Blaylock. 2005. *Towards Tractable Agent-based Dialogue*. Ph.D. thesis, Rochester.

H. H. Clark and C. R. Marshall. 1981. Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge.

H.H. Clark and D. Wilkes-Gibbs. 1990. Referring as a collaborative process. In P. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 463–493. MIT.

H. H. Clark. 1996. *Using Language*. Cambridge.

P. R. Cohen and H. J. Levesque. 1991. Teamwork. *Nous*, 25:11–24.

D. Dennett. 1989. *The Intentional Stance*. MIT.

C. Gauker. 1998. What is a context of utterance. *Philosophical Studies*, 91:149–172.

J. Ginzburg. 1996. Interrogatives: Questions, facts and dialogue. In S. Lappin, editor, *Handbook of Contemporary Semantic Theory*. Blackwell.

B. J. Grosz and S. Kraus. 1996. Collaborative plans for complex group action. *AI*, 86(2):269–357.

A. Gruenstein, L. Cavedon, J. Niekrasz, D. Widdows, and S. Peters. 2004. Managing uncertainty in dialogue information state for real time understanding of multi-human meeting dialogues. In *Proceedings of Catalog*.

E. Horvitz and T. Paek. 2001. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In *User Modeling Conference*, pages 3–13.

D. Kaplan. 1989. Demonstratives. In J. Almog, J. Perry, and H. Wettstein, editors, *Themes from Kaplan*, pages 481–563. Oxford.

S. Larsson and D. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *NL Engineering*, 6:323–340.

D. Lewis. 1969. *Convention: A Philosophical Study*. Harvard.

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

S. W. McRoy and G. Hirst. 1995. The repair of speech act misunderstandings by abductive inference. *Computational Linguistics*, 21(4):435–478.

M. Poesio and D. R. Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

C. Rich, C. L. Sidner, and N. Lesh. 2001. Collagen: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, 22(4):15–25.

N. Roy, J. Pineau, and S. Thrun. 2000. Spoken dialogue management using probabilistic reasoning. In *Proc. of ACL*, pages 93–100, Hong Kong.

S. Schiffer. 1972. *Meaning*. Oxford.

R. Stalnaker. 1974. Pragmatic presuppositions. In *Context and Content*, pages 47–62. Oxford.

R. Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics 9*. Academic Press.

R. Stalnaker. 1998. On the representation of context. *Journal of Logic, Language, and Information*, 7(1):3–19.

J. Taylor, J. Carletta, and C. Mellish. 1996. Requirements for belief models in cooperative dialogue. *User Modeling and User-Adapted Interaction*, 6(1):23–68.

R. H. Thomason, M. Stone, and D. DeVault. 2006. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. To appear in Byron, D., Roberts, C., and Schwenter, S., eds, Presupposition Accommodation.

R. H. Thomason. 1990. Accommodation, meaning, and implicature. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 325–363. MIT.

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

Modeling anaphora in informal mathematical dialogue

Magdalena Wolska Ivana Kruijff-Korbayová

Fachrichtung Computerlinguistik
Universität des Saarlandes, Postfach 15 11 50
66041 Saarbrücken, Germany
{magda,korbay}@coli.uni-sb.de

Abstract

We analyze anaphoric phenomena in the context of building an input understanding component for a conversational system for tutoring mathematics. In this paper, we report the results of data analysis of two sets of corpora of dialogs on mathematical theorem proving. We exemplify anaphoric phenomena, identify factors relevant to anaphora resolution in our domain and extensions to the input interpretation component to support it.

1 Introduction

Our goal is to develop a discourse understanding module for a dialog-based system for tutoring mathematics. A number of computational anaphora resolution approaches have been proposed (Mitkov, 2002), including solutions specific to modeling reference to entities other than nominals (Byron, 2004), as well as approaches specific to dialogue (Eckert and Strube, 1999; Jain et al., 2004). We can partly draw on those solutions, however, our domain differs from the domains these approaches address in that it involves formalized mathematical notation. While parsing and interpretation techniques for mixed natural and symbolic language do exist (Baur, 1999; Zinn, 2003; Wolska and Kruijff-Korbayová, 2004), referentiality phenomena have not, to our knowledge, been thoroughly studied. An additional challenge is posed by formal errors and sloppiness in students' proofs that may introduce referential ambiguity.

(Wolska et al., 2004; Wolska and Korbayová, 2006) presented two corpora of tutorial dialogs on mathematical theorem proving collected in a Wizard-of-Oz setup. We conducted an analysis of this data in order to guide the development of an anaphora resolution algorithm suitable for dialogs in the domain of mathematics,. Our goal is to (i) systematically investigate reference phenomena specific to mathematical dialog, (ii) based on empirical findings, propose a co-reference resolution method for our domain.

In this paper, we report the first results of data investigation. We concentrate on the peculiarities of the genre at hand: notably, references to mathematical concepts and expressions. With this focus in mind, we present and exemplify anaphoric phenomena observed in the two corpora. Second, we discuss our observations on implications for anaphora resolution and the functionality of the input interpretation component necessary to support it.

The paper is organized as follows: In Section 2, we briefly present the corpora we study. In Section 3, we show corpus examples of reference phenomena. In Section 4, we present our observations related to modeling anaphora in our domain, and extensions to an input interpretation module needed to support anaphora resolution.

2 Corpus

Our analysis is based on two tutorial dialog corpora[1] collected in Wizard-of-Oz experiments: Corpus-I (Benzmüller et al., 2003) and Corpus-II (Wolska and Korbayová, 2006). In both experiments, the subjects were told that they were interacting with a conversational tutoring system. They were using natural lan-

---

[1] The corpora are available online.

brandial 2006: *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 147–154.

Let R be a relation on a set M. Prove: $R = (R^{-1})^{-1}$

A relation is defined as a set of pairs. The above equation expresses an equality between sets. Set equality can be proven by The Principle of Extensionality, where one shows that every element of one set is also an element of the other set. Let (a, b) be a pair on $M \times M$. We have to show that $(a, b) \in R$ if and only if $(a, b) \in (R^{-1})^{-1}$. $(a, b) \in (R^{-1})^{-1}$ holds by definition of the inverse relation if and only if $(b, a) \in R^{-1}$ and this again holds by the definition of the inverse relation if and only if $(a, b) \in R$, which was to be proven.

Figure 1: Example proof from Corpus-II.

guage (German) typed on the keyboard as well as mathematical symbols. Both the subjects and the tutors were unconstrained in the way they formulated their turns. Corpus-I contains 66 dialogs (775 turns) on proofs in the domain of naive set theory, and Corpus-II 37 dialogs (1615 turns) on binary relations.

Analysis of the corpora reveals various phenomena that present challenges for modeling anaphora and anaphora resolution. The prominent phenomenon is reference to (parts of) the formal mathematical notation. This raises questions about introducing discourse entities for mathematical expression parts as well as requires extensions to the standard functionality of input processing subcomponents. We discuss the extensions in Section 4.3, but first, illustrate the phenomena with examples from the corpora.

3 Phenomena

To indicate the overall complexity of the anaphora resolution task in our setting, we present an overview of common reference phenomena. First, we give a brief characterization of the language of informal mathematical discourse, and then present anaphoric phenomena specific to the domain: reference to (parts of) mathematical expressions and mathematical propositions.

3.1 Language of informal mathematical discourse

Informal mathematical discourse can be characterized as a mixture of natural language interleaved with conventionalized formal expressions. Formal mathematical language consists of a vocabulary of symbols and operators, and technical terminology specific to a sub-field. Mathematical expressions include terms (denoting abstract mathematical objects) and statements (formulas) built from the vocabulary, both of arbitrary structural complexity. An informal proof consists of a sequence of assertions derived by application of inference rules. Figure 1 shows an example proof from Corpus-II presented to a subject at the end of a tutoring session.[2] In the course of the proof exposition, symbols that denote domain-objects (here: e.g. relations, pairs, sets) are mentioned and anaphoric devices are used to refer to abstract entities they denote or their specific (symbolic) instantiations in the discourse.

Below, we illustrate examples of references in informal mathematical dialogue from the point of view of the type of entity referred to. The phenomena themselves are not new, but the formal domain adds complexity to them, in particular from the point of view of referential ambiguity and functionality needed for anaphor resolution in general. The dialog excerpts to which we refer here are included in the Appendix.

3.2 Referring to (parts of) symbolic notation

Using pronouns and pronominal adverbs In (1), a pronoun, "it", is used to refer to a term in a formula, <u>a set variable</u> "B", whose syntactic/semantic function in the formula can be viewed as that of a subject/agent, parallel to the semantic function of the anaphor. In (2), a pronoun is referring to <u>a variable naming a member of a set.</u> In (3) the same name, "x", was introduced with the intention of denoting two different

---

[2] We present only an English translation here for space reasons.

entities. The entities are moreover of different types (in one case, a pair is a variable in a set abstract[3] and "x" is refers to an element of this pair, in the other case, "x" refers to a set-member variable of a simple form). Ambiguous designation is invalid in a mathematical proof and the tutor issues a clarification sub-dialog, in which, in turn, the pronominal reference in **S19** has an ambiguous denotation.

In (4), a pronominal adverb "davon" (en. *of it*) is used to refer to a complex term, "R $\circ$ S", on the left-hand side of the definition. In principle, the reference is ambiguous: a competing antecedent for "davon" is the definiens part of the definition.

Using noun phrases In our analysis, we include bridging references. We have found certain types of bridging references to systematically recur in our corpora. For example, noun phrases such as "the inner bracket" and "the left side" refer to a formula's structural part: a term in a formula. Both need a metonymic re-interpretation: "the left side" refers to the term to the left of the top-node operator in a formula (rather than a topological area), while "the inner bracket" refers to a bracketed subterm of a bracketed term in a formula (rather than to a bracket itself).

There are two ways of interpreting the definite noun phrase "the powerset" in the student utterance "I have problems with the powerset, I don't know how to compute it…" On the one hand, it may be referring to a term headed by the powerset operator (rather than the powerset operator itself) in the previous student turn that contains the following expression: "$P((A \cup C) \cap (B \cup C)) = P(C \cup (A \cap B))$". In this case, it needs a metonymic extension. Under this interpretation, the reference is ambiguous as there are two powersets in the expression. On the other hand, it is more plausible to interpret the reference generically; the student has a general problem in understanding the concept of a powerset.

Analogously, the quantified noun phrase, "beide Komplemente" in (5) needs a metonymic re-interpretation. Moreover, the reference is truly ambiguous in that there are five complement-headed terms in the preceding formula. A resolution algorithm must, therefore, not only decide on distributive vs. collective reading of the plural, but also identify plausible scopes for antecedent search.

In (7), the definite noun phrase "diese Menge" (en. *this set*) in **S35** is again a bridging reference to the set denoted by a term in **S34** (where the type of the result of the top-node operator is set).

Using demonstratives and discourse deixis In (6), the deictic reference "der obere Ausdruck" (en. *the above expression*) refers to the entire formula in the preceding turn, while the demonstrative pronoun "dies" (en. *this*) in (7) refers to a term in the previous formula.

3.3 Referring to propositions

Pronouns, **demonstratives** and **adverbial pronouns** may be used to refer to propositions as well as partial proofs constructed in the course of a dialog. In (8) the adverbial pronoun "damit" (en. *with this*) in **S7**, refers to the proposition stated in the first clause of the utterance. The pronominal adverb "somit" (en. *with that*) in **S8** in the same excerpt may refer to the conjunction or implication of the assertions in **S7** or only the last assertion (marked with j in the example). In (9), the pronoun "es" (en. *it*) is referring to the proposition in the tutor's turn **T19**.

3.4 Referring to domain-concepts

Both **definite and bare noun phrases** are used generically to refer to concepts in the domain, e.g. "the union" in: "The union of sets R and S contains all elements from R and all elements from S.". In "Powerset contains all subsets therefore also $(A \cap B)$", "powerset" is a generic reference, whereas "$(A \cap B)$" is a specific reference to a subset of a specific instance of a power set introduced earlier. Moreover, named theorems and lemmata may be referred to by their **proper names**, for example, "deMorgan rule 2".

---

[3] A set abstract is a set-denoting expression of the form $\{v : \phi\}$, where v is a variable and ϕ a formula

	Corpus-I	Corpus-II
math. expr. part	26	13
proof step	35	81
formula	19	46
mixed	16	35
Total	61	94

Table 1: References to domain objects

	PRP	prp. or loc. adv.	defn.art./dem.	def.NP
simple term	2	0	0	2
complex term	3	2	2	28
Total	5	2	2	30
ante. in S-turn	5	2	2	18
0 (same turn)	3	1	0	4
1	0	1	2	5
2 or earlier	2	0	0	9
ante. in T-turn	0	0	0	12
1 (prev. turn)	0	0	0	4
2 or earlier	0	0	0	8
of that in task def.	0	0	0	10

Table 2: References to parts of mathematical expressions

To summarize, the first and most obvious observation based on the corpus is that anaphoric references are used to refer to the formal notation of mathematical expressions. References may address entire formal expressions or their parts, and antecedents may lie in either own or the other party's turns. In spite of a seemingly high potential for ambiguity, only in one case was an explicit clarification dialog initiated by the tutor to clarify an ambiguous reference. Below, we present details of our corpus analysis and observations relevant for modeling anaphora.

4 Modeling anaphora in tutorial dialogues on proofs

We looked at all occurrences of references to *domain objects* in both corpora. For the purpose of this paper, by domain objects we mean (i) symbolic mathematical expressions and their parts, (ii) domain relevant propositions (mathematical assertions); e.g. proof steps proposed by the student expressed either formally or in words.[4] Below, we present a quantitative result of our analysis, summarize the observations concerning referentiality phenomena with in our context, and present extensions to the input understanding module we have implemented to support anaphor resolution in our domain.

4.1 Quantitative corpus analysis

Overall, of the 1269 student turns in both corpora, 140 turns contained references to some domain object: 46 out of 332 (14%) in Corpus-I and 94 out of 709 (13%) in Corpus-II. The details of the analysis are presented in three tables which we discuss below.

Table 1 presents an overview of references to domain objects: parts of mathematical formulas and propositions (proof steps). There were overall 155 anaphoric references. The relatively large number of references to proof steps in the second corpus, we think, is related to the style in which proofs were conducted. Most students built their proofs by re-writing preceding terms, and referring to the previous step either with discourse markers, such as "hence" or "therefore" or with pronominal adverbs (e.g. "somit", en. *with that*).

Table 2 shows an overview of references to (parts of) mathematical expressions. Here we include references to simple terms (i.e. symbolic identifiers such as variables A, B, x, etc.) and complex terms (terms containing at least one operator symbol). Of the 27 references, the largest proportion are nominal bridging references to formula parts (such as "left side" or "inner brackets" exemplified in Section 3.2). The antecedent tends to be found either in the student's own turn or in the task definition (the goal formula to be proven).

Table 3 presents a summary of references to propositions. There are 116 instances of such references, the majority of which are realized with German pronominal or locative

---

[4] We do not include in this analysis proper name named entity references referring to domain concepts, theorems, lemmata, etc., such as "The Second De Morgan Law", "The Distributivity Law".

	PRP	prp. adv. or loc. adv.	demonstr. or def.art.	def. NP
formula	0	28	27	10
mixed nl+formula	2	31	13	5
Total	2	59	40	15
ante. in S-turn	1	59	38	12
0 (same turn)	0	21	22	4
1	1	38	16	6
2 or earlier	0	0	0	2
ante. in T-turn	1	0	2	3
1 (previous turn)	1	0	2	0
2 or earlier	0	0	0	3
of that in task def.	0	0	0	3

Table 3: References to propositions

adverbs (59) and demonstrative pronouns or definite articles (40). A large proportion of these were found in Corpus-II. References to propositions tend to be local: most of the time, the antecedent is found in the student's own turn, in the same turn as the anaphor or preceding turn with respect to the anaphor.

4.2 Factors in anaphor resolution

Our corpus analysis of anaphoric reference to domain objects, yields the following observations relevant to anaphora resolution:

Sources of information There appear to be three major information sources to which an anaphor resolution module in our domain needs access:

(i) The semantic interpretation of the utterance and the utterance's function;
In order to provide information on the semantic content, in particular, with respect to proof contribution, the utterances in the student turn must be parsed and interpreted in the context of the given domain. In the further discussion, we assume the approach to interpretation as the one presented in (Wolska and Kruijff-Korbayová, 2004) and discuss required extensions. Of particular importance for anaphor resolution is whether according to the assigned interpretation, the given utterance is intended to convey a proof step (domain contribution) or not.

(ii) The correctness status of the last student's proof step;
For example, in re-writing style of proofs, students tend to make references to the last correct proof step (or partial proof) to indicate that it justifies the current step. We will return to this when we discuss salience of propositions below.

(iii) The semantic content of the last tutor move;
The tutor dialog moves include, among others, proof step evaluations (e.g. "That is not correct.") and hints (e.g. "How about starting the proof like this: ..."). If the last tutor's turn contains a hint that gives away the correct step expected at the time, the student is likely to refer to that step. Moreover, the first tutor's dialog contribution defining the exercise (the goal proposition) is also often referred to.

Antecedent candidates in references to (parts of) formulas Anaphoric references to mathematical expression parts appear to have local scope. In most cases, the referent occurred in the same or immediately preceding turn with respect to the anaphor, as exemplified in (1). In all cases of "it"-references (neuter personal pronouns) the anaphor was the entity on the left side of the candidate mathematical expression of type formula. This can be explained by the fact that in the verbalized form of such expressions, the entity on the left side plays the role of the subject or agent of the predicate.

Moreover, the structure of mathematical expressions is a strong indicator in identifying the search space for antecedents. This holds both in case of noun phrase references to topographical structure (e.g. "inner bracket" or "left side") as well as in case of quantified phrases referring to sub-structure. In the latter case, the topographical structure may help in guiding the search (e.g. in (5)).

In order to support resolution of references to (parts of) mathematical expressions, an input interpretation module must include a mathematical expression tagger and a deep parser for mathematical expressions, in particular, the parser must be capable of identifying all the relevant sub-structures of mathematical expressions. On the domain modeling side, it needs procedures for dealing with metonymic references to formula sub-structures.

Salience of propositions As the student develops the proof, the cognitive salience of the propositions that are part of the proof (proof steps) changes. At the beginning of the dialog, the most salient proposition is the goal formula (the exercise definition). According to our observations, as the proof progresses, the most salient proposition at a given time is the last correct proof step. If the student made several incorrect steps, no correct steps, and the tutor has not given away any steps, the goal formula in the exercise definition remains the most salient proposition even after several turns.

4.3 Extensions to input understanding module

To resolve references to (parts of) mathematical expressions, two issues must be taken into account: first, as mentioned above, we need a comprehensive analysis of mathematical expressions, and second, we need to include the entities specific to mathematical expression analysis in the domain-specific knowledge representation. Below, we summarize our implementation of domain modeling extensions required for reference resolution in the corpora we analyzed.

Mathematical expression parsing The mathematical expression parser uses simple indicators to identify mathematical expressions within sentence- and word-tokenized text. They include single character tokens, designated strings for mathematical symbols, and new-line characters.

The parser converts the infix notation used in the input into an expression tree whose nodes are marked as to whether they denote operators or variables; the expression type is marked on the root-node operator (e.g. FORMULA, TERM, etc.). Moreover, the parser has access to domain-knowledge on the type of result of mathematical operations (e.g. the subset relation takes two sets and the type of the result is a truth-value). The expression tree is an input structure to subroutines relevant for reference resolution.

Considering the complexity of the mathematical expressions, we take a pragmatic approach in modeling reference to mathematical expression sub-parts, in that at the time of parsing we only create a discourse referent for the entire expression[5], but not for every sub-structure entity relevant for anaphor resolution. Instead, the mathematical expression parser includes subroutines that on-demand recover (i) specific parts of mathematical expressions in specific PART-OF relations to the original expression, (ii) their types.

The choice of identified sub-structures is motivated by systematic reference in natural language to those parts (see Section 3.2) and includes: (i) topological features (such as “sides” of terms and formula); (ii) linear orders (e.g. “first”, “second” argument); (iii) structural groupings (bracketed sub-expressions) with information on their embedding. Execution of those subroutines is triggered by lexical semantic interpretation of the utterances (e.g. the meaning of “side” together with its modifier “left” in the representation of the noun phrase “the left side”).

Domain modeling Objects associated with types of mathematical expressions (e.g. FORMULA, TERM) as well as substructure delimiters (e.g. bracket, vertical bar of a set abstract) are represented in an ontological representation of domain objects.

Motivated by the systematicity in metonymic references to mathematical expression sub-parts, as part of the domain-model we encode “metonymy rules” that allow to reinterpret utterances with certain sortal restriction conflicts. Currently, the choice of rules is guided by phenomena found in our two cor-

---

[5] See (Wolska and Kruijff-Korbayová, 2004) for parsing the mixed language.

pora and includes the following:[6]

1. SIDE : TERM;
2. BRACKET : TERM;
3. OBJECT : RESULT;
4. RESULT : OPERATOR;
5. OPERATOR : SUB-TREE.

For example, the noun phrase “this set” referring to the expression $(S \cup R) \circ S^{-1}$ in (7), can be then interpreted by applying rule **3** first and then rules **4** and **5**.

Discourse modeling Our preliminary implementation of the discourse model, includes a data structure storing a dialog history. Aside from the interpretation of student input utterances, a dialog history stores information on the semantic content of tutor moves, in particular, information about the correctness of the proof steps proposed by the student, as well as symbolic representation of proof steps that were disclosed to the student during tutoring.

5 Conclusions

Based on experimentally collected data, we presented examples of anaphoric phenomena in tutorial dialogs on mathematical proofs and a quantitative analysis of two corpora from the point of view of reference to entities specific to the genre: mathematical expressions and propositions expressing proof steps. We discussed corpus observations relevant to building an anaphor resolution algorithm for the domain. These observations yield constraints and preference criteria for forming sets of candidate antecedents and for antecedent search, that we can directly incorporate into a first implementation of an anaphor resolution algorithm which we are presently developing. We also presented our extensions to an input interpretation component necessary to support the anaphor resolution algorithm.

References

Judith Baur. 1999. *Syntax und Semantik mathematischer Texte*. Diplomarbeit, Fachrichtung Computerlinguistik, Universität des Saarlandes, Saarbrücken, Germany.

Christoph Benzmüller, Armin Fiedler, Malte Gabsdil, Helmut Horacek, Ivana Kruijff-Korbayová, Manfred Pinkal, Jörg Siekmann, Dimitra Tsovaltzi, Bao Quoc Vo, and Magdalena Wolska. 2003. A Wizard-of-Oz experiment for tutorial dialogues in mathematics. In *AIED2003 Supplementary Proceedings*, volume VIII: Advanced Technologies for Mathematics Education, pages 471–481, Sydney, Australia.

Donna K. Byron. 2004. Resolving pronominal reference to abstract entities. Technical Report 815, University of Rochester Computer Science Dept.

M. Eckert and M. Strube. 1999. Resolving discourse deictic anaphora in dialogues. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–44, Bergen, Norway.

Prateek Jain, Manav Ratan Mital, Sumit Kumar, Amitabha Mukerjee, and Achla M. Raina. 2004. Anaphora resolution in multi-person dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL-04*, pages 47–50, Ann Arbor, MI.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman.

Magdalena Wolska and Ivana Kruijff Korbayová. 2006. Factors influencing input styles in tutoring systems: the case of the study-material presentation format. In *Proceedings of the ECAI-06 Workshop on Language-enabled Educational Technology*, Riva del Garda, Italy. To Appear.

Magdalena Wolska and Ivana Kruijff-Korbayová. 2004. Analysis of Mixed Natural and Symbolic Language Input in Mathematical Dialogs. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-04)*, pages 25–32, Barcelona, Spain.

Magdalena Wolska, Bao Quoc Vo, Dimitra Tsovaltzi, Ivana Kruijff-Korbayová, Elena Karagjosova, Helmut Horacek, Malte Gabsdil, Armin Fiedler, and Chris Benzmüller. 2004. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *Proceedings of International Conference on Language Resources and Evaluation (LREC-04)*, pages 1007–1010, Lisbon, Potugal. ELDA.

Claus Zinn. 2003. A Computational Framework For Understanding Mathematical Discourse. *Logic Journal of the IGPL*, 11(4):457–484.

---

[6] The presentation included here is only schematic; for example, SIDE : TERM means that the concept “side” (left or right) may be alternatively interpreted as “term” in a formula to the left or right accordingly.

Appendix – Dialog excerpts

(1) **S6:**… Da, wenn $\mathrm{A}_j \subseteq \mathrm{K}(\mathrm{B}_i)$ sein soll, A_j Element von $\mathrm{K}(\mathrm{B}_i)$ sein muss. Und wenn $\mathbf{B}_i \subseteq \mathrm{K}(\mathrm{A}_j)$ sein soll, muss **es**$_i$ auch Element von $\mathrm{K}(\mathrm{A}_j)$ sein.

Because if it should be that $A_i \subseteq K(B_j)$, A_i must be an element of $K(B_j)$. And if it should be that $B_k \subseteq K(A_l)$, it must be an element of $K(A_l)$ as well.

(2) **S1:** Wie ist R ∘ S definiert?

How is R ∘ S defined?

T1: $\mathrm{R{\circ}S} := \{(x,y) \mid \exists z(z \in M \wedge (x,z) \in R \wedge (z,y) \in S\}$

…

S4: ist z$_i$ nur fuer die Definition eingefuehrt oder hat **es**$_i$ einen anderen Sinn?

is z introduced only for the definition or does it have a different meaning?

(3) **S18:** … Daraus folgt $(R \cup S) \circ T = \{(x_?,y) \mid \exists z(z \in M \wedge (x,z) \in \{x_? \mid x_? \in R \vee x_? \in S\} \wedge (z,y) \in T)\}$

… From that follows $(R \cup S) \circ T = \{(x_?,y) \mid \exists z(z \in M \wedge (x,z) \in \{x_? \mid x_? \in R \vee x_? \in S\} \wedge (z,y) \in T)\}$

T19: Was bedeutet **die Variable**$_i$ $\mathbf{x}_i$ bei Ihnen?

What is the meaning of your variable x?

S19: $\mathbf{x}_i$ hat zwei Bedeutungen es kommt in zwei verschiedenen Mengen vor

x has two meanings it is contained/comes in two different sets

T20: Benutzen Sie bitte fuer die zwei verschiedenen Bedeutungen von x zwei verschiedene Bezeichnungen.

Please use two different designations for the two different meanings of x.

(4) **S: R ∘ S**$_i$ $:= \{(x,y) \mid \exists z(z \in M \wedge (x,z) \in R \wedge (z,y) \in S)\}$

T: Das ist richtig!

This is correct!

S: Nun will ich das Inverse **davon**$_i$

Now I want the inverse of it

(5) **T1:** Bitte zeigen Sie: $K((A \cup B) \cap (C \cup D)) = (K(A) \cap K(B)) \cup (K(C) \cap K(D))$!

Please show: …

S2: de morgan regel 2 auf **beide komplemente** angewendet

de morgan rule 2 applied to both complements

(6) **T1:** Bitte zeigen Sie: $[K((A \cup B) \cap (C \cup D)) = (K(A) \cap K(B)) \cup (K(C) \cap K(D))]_i$!
S1: laut De-Morgan-Regel-2 gilt : $(K(A) \cap (B)) = K(A) \cup K(B)$, damit kann ich **den oberen Ausdruck**$_i$ wie folgt schreiben: …

By De-Morgan-Law-2 holds: $(K(A) \cap (B)) = K(A) \cup K(B)$, given that I can write the above expressions as follows: …

(7) **S33:** Nach Aufgabe W ist $(S \circ (S \cup R)^{-1})^{-1} = [((S \cup R)^{-1})^{-1} \circ S^{-1}]_i$

By Exercise W holds: …

T34: Das ist richtig!

S34: Dies$_i$ ist nach Theorem 1 gleich $[(S \cup R) \circ S^{-1}]_j$

This is by Theorem 1 equal to …

T35: Das ist auch richtig!

S35: Ein Element (a,b) ist genau dann in **dieser Menge**$_j$, wenn es ein $z \in M$ gibt mit $(a,z) \in S \cup R$ und $(z,b) \in S^{-1}$

An element (a,b) is in this set if and only if there is a $x \in M$ such that $(a,z) \in S \cup R$ and $(z,b) \in S^{-1}$

(8) **S7:** Also $[\text{ist } (z,x) \in S \text{ und } (y,z) \in R]_i$ und **damit**$_i$ auch $[(y,x) \in R \circ S]_j$

Therefore holds $[(z,x) \in S$ and $(y,z) \in R]_i$ and by that $_i$ also $[(y,x) \in R \circ S]_j$

T7: Sie haben recht.

You are correct.

S8: Somit$_j$ ist $(x,y) \in (R \circ S)^{-1}$

Given that it holds that $(x,y) \in (R \circ S)^{-1}$

(9) **S12:** $\exists\ z \in M$, so dass $(x, z) \in S^{-1}$ und $(z, y) \in R^{-1}$

$\exists\ z \in M$ such that $(x, z) \in S^{-1}$ and $(z, y) \in R^{-1}$

T18: Richtig. Wissen Sie, ob ein solches z existiert?

Correct. Do you know whether such z exists?

S13: Nein

No

T19: Erinnern Sie sich daran, $[$ dass es ein z gibt mit $(x, z) \in S^{-1}$ und $(z, y) \in R^{-1}$. $]_i$

Do you remember that there is a z such that $(x, z) \in S^{-1}$ and $(z, y) \in R^{-1}$.

S14: Ja, ich habe **es**$_i$ vorausgesetzt

Yes, I made such assumption

Dialogue pressures and Syntactic Change

Ruth Kempson
Philosophy Department
King's College London
London, WC2R 2LS
`ruth.kempson@kcl.ac.uk`

Ronnie Cann
Linguistics and English LanguageDepartment
University of Edinburgh
Edinburgh EH8 9LL
`ronnie@ling.ed.ac.uk`

Abstract

On the basis of the Dynamic Syntax framework, this paper argues that the production pressures in dialogue determining alignment effects and given versus new informational effects also drive the shift from case-rich free word order systems without clitic pronouns into systems with clitic pronouns with rigid relative ordering. The paper introduces assumptions of Dynamic Syntax, in particular the building up of interpretation through structural underspecification and update, sketches the attendant account of production with close coordination of parsing and production strategies, and shows how what was at the Latin stage a purely pragmatic, production-driven decision about linear ordering becomes encoded in the clitics in the Medieval Spanish system which then through successive steps of routinization yield the modern systems with immediately pre-verbal fixed clitic templates.

1 Introduction

This paper argues that production pressures in dialogue that determine given versus new informational effects drive the progressive grammaticalization of pronouns into ever weaker forms, potentially leading ultimately to the formation of clitics with fixed positions in the clause. Grammaticalization of this sort is argued to take place through progressive psycholinguistic routinization of general linguistic procedures that utilize context dependence (i.e. through the use of anaphoric devices) to ameliorate problems on language production with regard to lexical searches for appropriate word forms.

The case study used to support this hypothesis is the shift from Latin, with its free word order and rich case morphology, through Medieval and Renaissance forms of Spanish, with their atrophied case systems, to the modern language in which case is only expressed in the clitic pronoun system, with fixed immediate preverbal position in finite clauses. The formal framework within which this account is set out is Dynamic Syntax (Kempson *et al*, 2001; Cann *et al*, 2005).

The starting point is the DS account of dialogue and its analysis of the widespread use of ellipsis, pronouns and alignment effects involving repeating words, interpretation, and syntactic structures (Cann *et al* 2005, Purver *et al* 2006):

(1)
A: What should Michael give Ruth for Christmas?
B: A pianola.
C: Unless he's giving her a harpsichord.
D He could give her a spinet, if you prefer.

The account analyses all such effects as the minimization of the production task of searching in the lexicon. It is this which we argue is the driving force behind the emergent syntactic properties of clitics in the shift from Latin to Medieval and Renaissance Spanish, with successive steps of routinization involving the storage of previous syntactic information from the context within the

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 155–162.

lexicon as macros of actions associated with the parse of specific forms (in other words, as a process of grammaticalization involving the lexicalization of syntactic processes). Each new stage of routinization then develops this process as a means of reducing production costs.

2 Towards a Dynamic Syntax of Latin

DS is a parsing-directed grammar formalism, in which a decorated tree structure representing a semantic interpretation for a string is incrementally projected following the left-right sequence of the words, from a starting point with just a rootnode and a requirement for some propositional value, to an endpoint which is a fully decorated binary branching tree structure encoding functor-argument structure of a familiar sort:[1]

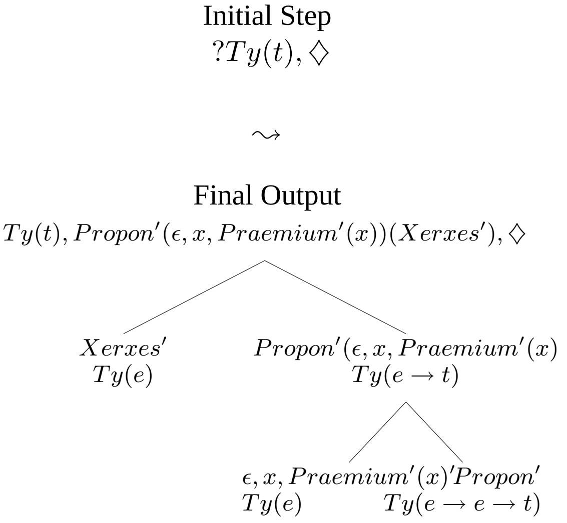

Initial Step
$?Ty(t), \diamondsuit$
$\rightsquigarrow$
Final Output
$Ty(t), Propon'(\epsilon, x, Praemium'(x))(Xerxes'), \diamondsuit$
$Xerxes'$
$Ty(e)$
$Propon'(\epsilon, x, Praemium'(x)$
$Ty(e \to t)$
$\epsilon, x, Praemium'(x)$
$Ty(e)$
$Propon'$
$Ty(e \to e \to t)$

Figure 1: Parsing *Xerxes praemium proposuit*

The process of tree-growth is the basis of syntactic explanation: a sentence is defined to be well-formed just in case there is at least one possible route through that process. Central to this is the concept of requirement $?X$ for any decoration X, representing a type, formula or treenode address. For example, decorations on nodes such as $?Ty(t)$, $?Ty(e)$, $?Ty(e \to t)$ etc. express requirements to construct formulae of the appropriate type on the nodes so decorated , and these drive the subsequent tree-construction process.[2] These steps are determined either by general computational actions, such as anticipating a subject-predicate structure, or lexical actions triggered by parsing lexical items in the order in which they are presented in some string of words.[3] Crosslinguistic variation is expressed in terms of the actions invoked in parsing particular classes of words. In particular, variations in word order are determined, at least in part, by how much of the argument structure of a predicate is constructed by such actions. For example, SVO order in English is accounted for by a condition on parsing a (main) verb that a subject has already been constructed and only internal argument(s) of the verb are then projected as part of the tree growth process. In Latin, however, with its freer word order and possibility of pro-drop, parsing verbs induces a whole propositional structure whose argument nodes are decorated with **metavariables**: placeholders that stand for some real value to be assigned from the context, capturing the effect of null pronouns without the assumption that such things are real parts of a *linguistic* string:[4]

---

[1] Fo is a predicate that takes a logical formula as value, Ty a predicate that takes logical types as values, Tn a predicate that takes tree-node addresses as values, eg $Tn(0)$ being the rootnode.

[2] The formal system underpinning the partial trees that are constructed is a logic of finite trees (LOFT). There are two basic modalities, $\langle\downarrow\rangle$ and $\langle\uparrow\rangle$, such that $\langle\downarrow\rangle\alpha$ holds at a node if α holds at its daughter, and its inverse, $\langle\uparrow\rangle\alpha$, holds at a node if α holds at its mother. Function and argument relations are distinguished by defining two types of daughter relation, $\langle\downarrow_0\rangle$ for argument daughters, $\langle\downarrow_1\rangle$ for functor daughters (with their inverses $\langle\uparrow_0\rangle, \langle\uparrow_1\rangle$).

[3] Quantification is expressed in terms of variable-binding term operators, so that quantifying NPs like all other NPs are of type e. The underlying logic is the epsilon calculus, whose internal contains an epsilon binder,ϵ, a variable, and a restrictor: eg $\epsilon, x, Man'(x)$. Since in Latin, nouns project full specification of terms, the structure defined to be projected by *praemium* would be a subtree of which the quantifying term is the topnode, dominating a subtree decorated with binder, variable, and restrictor specification. We leave all details on one side.

[4] According to this characterization, Latin is object drop. One way to capture canonical verb object orderings within a full pro-drop system is to define the pointer to be at the object node on the tree following the parse of a verb, characterizing ordering of the object after the verb as the least marked of available options. We ignore details of tense specification throughout this paper.

(2)

```
IF      $?Ty(t)$
THEN    $\mathtt{put}(Tns(PAST));$
        $\mathtt{make}(\langle\downarrow_0\rangle) : \mathtt{go}(\langle\downarrow_0\rangle);$
        $\mathtt{put}(Ty(e), Fo(\mathbf{U}), ?\exists\mathbf{x}.Fo(\mathbf{x})); \mathtt{go}(\langle\uparrow_0\rangle)$
        $\mathtt{make}(\langle\downarrow_1\rangle); \mathtt{go}(\langle\downarrow_1\rangle); \mathtt{put}(?Ty(e \rightarrow t));$
        $\mathtt{make}(\langle\downarrow_1\rangle); \mathtt{go}(\langle\downarrow_1\rangle);$
        $\mathtt{put}(Fo(\text{Propon'}), Ty(e \rightarrow e \rightarrow t), [\downarrow]\bot)$
        $\mathtt{go}(\langle\uparrow_1\rangle); \mathtt{make}(\langle\downarrow_0\rangle); \mathtt{go}(\langle\downarrow_0\rangle);$
        $\mathtt{put}(Fo(\mathbf{V}), Ty(e), ?\exists x.Fo(x))$
ELSE    Abort
```

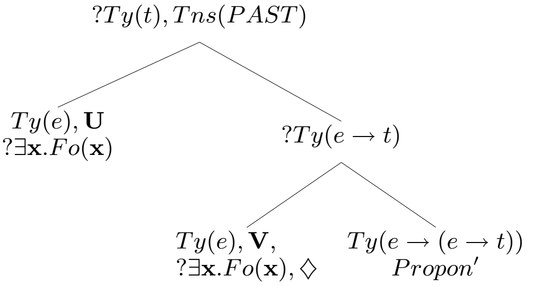


Figure 2: Result of lexical actions of *proposuit*

There is in DS also the concept of structural underspecification, with the construction of only weakly specified tree relations, which licenses the introduction of a node in some newly initiated logical structure, characterized only as $\langle\uparrow_*\rangle Tn(0)$ (“this node is dominated by the rootnode”).)[5] In case-rich languages such as Latin, this strategy is manipulated in conjunction with case-specifications which are used to update an unfixed node to a fixed relation (subject, direct object, indirect object). By this strategy, a string such as (3) can be parsed using case specifications to update each weak ‘dominate’ tree-relation before the parsing of the verb:[6]

(3) Praemium Xerxes proposuit
‘Xerxes offered a reward.’

Once any one relation is fixed, another unfixed node can be introduced, following through on the same sequence of actions. The verb then follows, filling out the remainder of the propositional structure to yield the appropriate output tree with $Fo(Xerxes')$ as subject argument $Fo(\epsilon, x, Praemium'(x))$ as object argument.[7] This allows ‘free’ word order effects without any necessary interpretational difference.

This specification of verbs as inducing full propositional structure equally applies in cases where its associated metavariable argument annotations are provided from context. Such a case occurs in the building of paired, ‘linked’ trees, which are subject to a restriction that they are anaphorically linked, a process used for relative clauses, clausal adverbials, and also external topic constructions. Such secondary structures have an attendant requirement that the newly introduced proposition-requiring tree have somewhere within it a copy of that term (specified as $?\langle\downarrow_*\rangle Fo(\alpha)$):[8]

$$\begin{tikzcd}\langle L\rangle Tn(0), Fo(\alpha), Ty(e) \arrow[r, bend left] & Tn(0), ?Ty(t), ?\langle\downarrow_*\rangle Fo(\alpha)\end{tikzcd}$$

Figure 3: Building Link transitions

For example, such a structure is invoked in the parse of *My new boss, she’s insane* in which the initial term is recapitulated in context by the pronoun. Link structures of this sort provide one instance of the dependence of the parsing process in DS on contextual information, but such context dependence is invoked throughout the system to account for anaphoric and other underspecified expressions whose values may be determined from within the current tree, from some linked tree or from some tree provided by the discourse context.

---

[5] $\langle\uparrow_*\rangle Tn(0)$ is the regular formal characterization of dominate: see footnote 2. The provided annotation then indicates that the rootnode dominates the current node.

[6] A formal restriction imposed by the system is that there be only one unfixed structural relation of a type at a time, any duplication leading to immediate collapse of the two nodes into one.

[7] Unlike two case-distinguished unfixed nodes, either subject or object nodes induced by actions of the verb harmlessly collapse with those introduced as unfixed and updated through constructive use of case (Nordlinger 1998), as annotations provided by the verb are compatible with those provided by computational actions used in parsing the NPs.

[8] The process of inducing such pairs of semantic trees is permitted by defining an additional modal operator in the tree logic, $\langle L\rangle$, and its inverse $\langle L^{-1}\rangle$; and a rule is defined to yield a transition from an arbitrary node in one tree across a LINK relation to the top node of a new propositional tree.

With the options of building an unfixed node within an individual tree, and building transitions from one tree to another to yield pairs of 'linked' trees, there are several strategies at the outset of building structure for any single string-interpretation pair; but this is unproblematic as the parsing-directed grammar formalism makes available alternative strategies for specifying the fine structure of HOW interpretation is built up.

2.1 Production

In production, the same rules used in parsing apply: the difference is that while the parser may not know in advance the interpretation to be constructed, the producer in contrast must do so, at least in part. So in generation, the same computational actions initiate the development of some tree but each update step licensed by the parsing mechanism has to meet the restriction of being a sequence of progressive enrichments towards completing a 'goal tree' representing the interpretation to be conveyed.[9] For example, in producing (3), *Praemium Xerxes proposuit*, the first action in initiating a sequence of steps to yield the goal tree is to start with a step that introduces a node decorated with the requirement $?Ty(t)$, just as in parsing; and one possible follow-up to this step is to introduce an unfixed node (as in Figure 4). Transparently, both the initial tree and this development subsume the goal tree in the sense that there is a licensed progression from these to the richer goal tree.

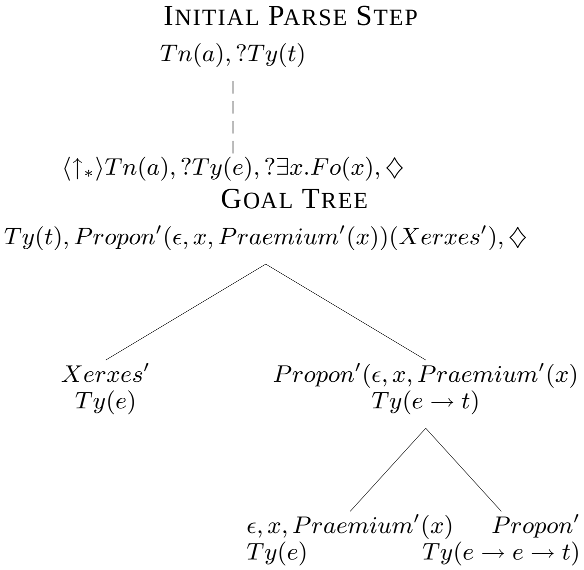

INITIAL PARSE STEP
$Tn(a), ?Ty(t)$
$\langle\uparrow_*\rangle Tn(a), ?Ty(e), ?\exists x.Fo(x), \Diamond$
GOAL TREE
$Ty(t), Propon'(\epsilon, x, Praemium'(x))(Xerxes'), \Diamond$
$Xerxes'$
$Ty(e)$
$Propon'(\epsilon, x, Praemium'(x)$
$Ty(e \rightarrow t)$
$\epsilon, x, Praemium'(x)$
$Ty(e)$
$Propon'$
$Ty(e \rightarrow e \rightarrow t)$

Figure 4: First production steps for *Praemium Xerxes proposuit*

From this step on, there is the problem of searching in the lexicon for words to express the given conceptual array. With this weak an update in structure, a very large number of options are available; and in principle the entire lexicon needs to be scanned. Appropriate continued lexical scanning may select *praemium* as providing a licensed update, a sequence of computational actions plus lexical search which is repeated all over again in producing *Xerxes*. Given the incrementality of parsing, carried over to production, this task is computationally expensive, threatening to be cognitively non-viable, all the more so in free word order languages as there are so many parsing options. However, we assume that production is just as context-dependent as parsing, re-using structure or formula values, even actions used to construct trees, wherever possible. Any element in context that can be identified as adding appropriately to the tree may not require words to be uttered, as long as the effect of adding it as a tree update matches the subsumption condition. For example: consider the mechanisms for producing an utterance of (5) in the context of having processed (4):

(4) *Xerxes* *iussit* *milites*
Xerxes$_{NOM}$ ordered soldiers$_{ACC}$
castra *captare*
camp$_{ACC}$ capture$_{INFIN}$
'Xerxes ordered the soldiers to capture the camp.'

(5) *Praemium* *proposuit*
Reward offered
'He offered a reward.'

In uttering (5), the subject argument node provided by the verb's actions is identified from

---

[9] Formally a subsumption relation is required to hold between the parse tree and the goal tree. For an early development of this view, see Purver and Otsuka 2003.

context. And it is here that using the very same process as in parsing reaps its rewards. As long as the minimal context contains a suitable term, matching the subsumption constraint, that term can be substituted as the value of the metavariable without more ado, so there need be no explicit morphologically presented subject: the simple verb form is sufficient.

This minimization on cognitive costs in production extends beyond merely using elements in context wherever possible. It also applies to choice of words, structure, and actions. Once a word or sequence of actions has been used in processing a string – parsing it or producing it - these actions can be re-used, this being the basis for the very considerable alignment effects:

(6) *Te, dea Te fugiunt*
you goddess you flee
venti. Te nubila
the winds$_{NOM}$ you clouds$_{NOM}$
coeli
of-heaven
'You goddess, the winds flee from you, the clouds of heaven (flee from you).'

Minimizing on production costs also affects word order, even without alignment. Though in Latin, there may be no need of a pronoun, anaphoric expressions serve a purpose in the linearization task as they enable argument terms to be identified independently of processing the verb. This consideration, in conjunction with the parallelism of parsing and production and general cognitive constraints such as relevance, helps to explain their preferred early positioning. In relying on context, both speaker and hearer need the search for a substituend to be as small as possible (by general relevance considerations). Accordingly, unless there is reason to the contrary, the position of an anaphoric expression will be as early as possible in the setting out of any propositional structure since this ensures that the search in the context for the value to be assigned to this expression will thereby be as small as possible. In order to minimize the search space effectively, there is pressure not to introduce words expressing new information into the string before contextually determined ones. This is of course no more than a pragmatic relevance-based explanation of the very wellknown given-before-new ordering that is regularly reported in free-constituent-order situations.

However, pronouns in Latin may be used to provide some initial term which constitutes a point of departure for what follows, or to provide a contrast, an update to what follows, in both such cases being set out initially in order to be identifiably separate from the structure to be constructed from what follows:[10]

(7) *"Tibi ego dem?" "Mihi*
you$_{Dat}$ I$_{NOM}$ give$_{1st.ps.sg}$ me$_{Dat}$
hercle uero"
by Hercules in truth
'Am I to give it to YOU?' 'Yes, by god, to ME'
[Plautus, Pseudolus 626 (Adams (2))].

Such uses of so-called strong pronouns are analysed as involving the projection by the pronoun of a term decorating a node at the left edge of a propositional boundary, i.e. as a separate linked structure, or an unfixed node, (7). In such uses, these provide the means of identifying boundaries to propositional domains, either in the projection of a separate tree, a linked structure, or to identify the initiation of a new propositional structure within which the term that they serve to introduce will provide an update.

There are in addition so-called weak uses of pronouns, which serve only as anaphoric devices. Being by definition complementary to the strong use of pronouns, this remainder of the set of pronouns will not be associated with those very structural devices which serve to identify some initiation of an emergent propositional structure. Nevertheless, like their "strong" counterparts, the positioning of these pronouns under this use will be driven by relevance considerations. That is, once an emergent propositional structure is identified by some *other* expression, we can expect weak pronouns to occur as closely following as possible.[11] With all pronouns, that

---

[10] The pronouns noted in (7) are taken by Adams 1994 to be illustrative of an emphatic use "often marked by placement of the pronoun at the head of its clause"(p.104).

[11] Following Sperber and Wilson 1995, if there are spe-

is, the search within the context has to be minimized by placing the pronoun as close to the context within which its value is to be identified as is commensurate with its function in that context.

3 Towards diachrony

We now have everything in place to explain why clitic pronouns cluster at some early position in a string. The weak pronouns of Latin occur as close to the left-edge of a clause as possible, but not quite at the edge. Rather, as noted above, they follow those devices which define an emergent propositional boundary, immediately following focussed elements, expressions containing a negative element, complementizers, relative pronouns, subordinate temporal adverbials, and verbs, these having in common their identification of some emergent edge of a new propositional domain:

(8) *quae* *tibi* *nulla* *debetur*
which$_{neut,pl}$ you$_{dat}$ no$_{neut,pl}$ is owed
‘nothing of which is owed to you.’
[relative-pronoun+pronoun]

(9) *Nihil* *me* *aliud*
nothing me$_{acc}$ other$_{neut,sg,nom}$
consolatur
it consoles
Nothing else gives me comfort.
[negative-quantifier+pronoun]

(10) *Magno* *me* *metu*
great$_{neut,abl}$ me$_{acc}$ fear$_{abl}$
liberaveris
you will have freed
‘You will have released me from great fear.’
[split part+pronoun]

(11) *rogo* *ut* *mi* *mittas* *dalabram*
I ask that me$_{dat?}$ you send mattock
‘I ask you to send to me a mattock.’
[complementiser+pronoun]

(12) *et* *non* *eum* *uendedi*
and not him$_{acc}$/it I sold
‘and I did not sell him’
[negation+pronoun]

(13) *delectarunt* *me* *tuae* *litterae*
delighted me your letter
‘I was delighted with your letter.’
FAM.IX.16.1
[verb+pronoun]

In the subsequent Medieval Spanish system the clitic pronouns share this distribution:

(14) *Esto* *es* *el* *pan* *de* *Dios* *que* *vos*
this is the bread of God that CL
da *a* *comer*
he-gives to eat
‘This is the bread of God that he gives you to eat.’Granberg, 1988: 35
[rel-pro+pronoun]

(15) *E* *non* *los* *hi* *fallo.* *and*
them there found.3sg
not

And he did not find them there.(XIII)
[negation+pronoun]

(16) *Dixo* *la* *mugier:* *Quien* *te*
said.3sg the woman: who you
fizo *rey?*
made.3sg king
‘The woman said: Who made you king?’ (XIII)
[WH+pronoun]

(17) *e* *dizie* *que* *lo* *tenie*
and he-said that CL-DO he-had
del *prior* *de* *Sancti* *Johannis*
of-the prior of Saint Johan
‘and he said that he got it from the prior of Saint John.’ [XIII; Granberg 1988]
[complementiser+pronoun]

(18) *e* *todo* *lo* *metieron* *a* *espada*
and all CL-DO they-put to sword
que....
that...

---

cific inferential effects to justify commensurate enlargement of the context to be searched, this would explain the lack of tightness of fit that Adams 1994 notes of weak pronoun positioning in Latin, even assuming that the effects are clause by clause (or “colon” by “colon”).

‘and he said that he got it from the prior of Saint John.’ [XII Granberg 1988]
[quantifier+pronoun]

(19) *Connociola* *Jacob.*
recognised.3sg-her Jacob
‘Jacob recognised her.’ (XIII)
[verb+pronoun]

Such left-peripheral items may however be a sequence of NPs (Devine and Stephens 2006):

(20) *caseum* *per* *cribrum*
cheese through sieve
facito *transeat*
make$_{2nd.sg.imp}$ go-through$_{3rd.sg.subjunct.}$
in *mortarium*
in bowl
‘Make the cheese go through the sieve into the bowl.’ Cato 76.3
[scrambled NP pair]

And this pattern recurs in medieval Spanish, at that later point in time associated specifically with clitic pronouns:

(21) *Et* *los* *dioses* *me* *quisieron* *mal*
And the gods CL want$_{3pl}$ harm
e *me* *lo* *quieren*
and CL CL want$_{3pl}$
‘and the gods wanted to harm me and they still want to.’
(XIII; cited by Granberg 1988: 235-236)

Thus the proclisis and enclisis effects in finite clauses for the weak pronouns of Latin and the clitic pronouns of medieval Spanish, can be described by a single generalization as a minimizing of context search, given the new introduction of an appropriate-sized domain.

4 Alignment, routinization and Change

Without an explanation of the change, this is not yet the full diachronic account; but dialogue effects go further than mere use of anaphoric devices and alignment. Dialogue participants, having having set up a parse sequence of actions may, over a very short time set up routines for retrieval of a stored sequence of actions encompassing more than one word (Garrod and Doherty 1994), yet another saving on cognitive costs since it involves retrieval from the lexicon of only one sequence of actions for a multiple string. Production, storage, and language change can now be seen as going hand in hand in the shift from Latin in the development of Spanish. One form of pronoun gets progressively phonologically reduced in virtue of predictability and recoverability from context. Given increasing phonological dissimilarity, separate clitic forms get encoded, what at that later stage has become an unstressable clitic being defined to follow the set of triggers previously established through pragmatically induced production constraints.[12] This process constitutes a form of routinization, listing, as triggers, the environments within which weak pronouns were construed as dependent for a value on some immediately preceding context.

The first observable step of encoding this heterogeneous set of triggers is a step of economy that combines computational and lexical actions as one lexical macro of actions. But this involves a disjunction of triggers, such as a negation feature, a subordinate marker inducing a new proposition-requiring node (for subordinating complementisers) the transition from a linked structure onto a decorated unfixed node (for a relative pronoun), a WH term decorating an unfixed node (for *wh* questions), and so on. This is not only clumsy, but hard to learn. So once the clitic is stored as a discretely encoded form, its macro of actions is a natural candidate for further routinization effects. In all such cases, much the commonest expression to immediately follow the clitic(s) is the verb (Adams 1994 amongst others); and a natural subsequent step of routinization, given the DS form of analysis, is to call up the actions associated with the verb together with those of the clitic, again as a further economy measure in reducing processing effort. We achieve the effect of re-bracketing,

---

[12] The strong pronouns subsequently come in Modern Spanish to be restricted to decorating linked structures, necessitating clitic doubling (see Cann et al 2005):
(i) le hablaron a ella
her$_{DAT}$ spoke$_{3pl.}$ to her [mod.Spanish]
‘They spoke to her.’

often observed. With such routinization, restrictions on proclisis collapse, since the heterogeneous set of triggers defining the environment licensing construal of a clitic is not a property that appropriately subclassifies the verbs with which the clitics are stored; and we get the intermediate stage of Renaissance Spanish, when all constraints on pre-verbal positioning of the clitics drop (see Bouzouita 2002, Bouzouita and Kempson forthcoming, Bouzouita in preparation).

From this point in time, the Romance languages, with the disappearing free constructive use of case, face the problem of confronting a ban on more than one unfixed node at a time on its NP construal. A variety of divergent routinizations emerge to side-step the problem. Some clitics directly induce the construction of the requisite fixed structural relation (eg French *le*). Others induce the building of a locally underspecified tree relation, hence underspecified with respect to the two discrete object construals (eg. French *me, te,* Castilian Spanish *le*). And in some cases a phonologically distinct composite clitic form is introduced that induces a single unfixed relation from which are constructed two argument nodes (eg Italian *glielo*, Spanish *se lo*). It is notable that each of these possibilities corresponds to actions independently justified, albeit at this point in time stored as a lexical sequence of actions, the last alternative corresponding to the sequence of actions earlier freely available in licensing examples such as (20). The framework thus can explain the idiosyncratic, highly restricted templatic sequencing of clitics, without introducing separate morphology-specific vocabulary. Overall, the full range of idiosyncratic variation is expressible through the simple assumption of building locally unfixed nodes, with various ways in which routinized conflation of macros can take place in the wake of internalised morphological changes imposing concomitant pressures for change.

Acknowledgement

The detailed work on medieval Spanish, and the diachronic study of Spanish clitics could not have taken the form it has without the input of Miriam Bouzouita. See Bouzouita 2002, Bouzouita and Kempson forthcoming.

References

John Adams 1994. Wackernagel’s law and the position of unstressed personal pronouns in Classical Latin. *Proceedings of the Philological Society*, 92: 103–78.

Miriam Bouzouita 2002. Clitic Placement in Old and Modern Spanish. MSc dissertation. King’s College London..

Miriam Bouzouita and Ruth Kempson. forthcoming. Clitic placement in Old and Modern Spanish: a dynamic account Nedergaard Thomsen O. (ed.) *Current Trends in the Theory of Linguistic Change*. John Benjamin, Amsterdam.

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

James Devine and Laurence Stephens. 2006. *Latin Word Order: Structured Meaning and Information*. Oxford University Press, Oxford.

Simon Garrod and Doherty G. 1994. *Cognition*, 53: 181–215. Cambridge University Press, Cambridge, UK.

R. Granberg. 1988. *Object Pronoun Position in Medieval and Early Modern Spanish*. University Microfilms International, Ann Arbor Michigan.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax*. Blackwell, Oxford.

Rachel Nordlinger 1998. *Constructive Case*. CSLI, Stanford.

Matthew Purver and Masayuki Otsuka. 2003. Incremental generation for dialogue. ACL workshop proceedings.

Matthew Purver, Ruth Kempson, and Ronnie Cann, 2006. Grammars as parsers: meeting the dialogue challenge. *Research on Language and Computation*, 4.

Martin Pickering and Simon Garrod. 2004. Towards a mechanistic account of dialogue. *Behavioral and Brain Sciences*,27: 169–226.

Dan Sperber and Deirdre Wilson. 1995 *Relevance: Communication and Cognition*. Blackwell, Oxford.

Route instruction dialogues with a robotic wheelchair

Thora Tenbrink
I1-[OntoSpace], SFB/TR 8
Spatial Cognition
U Bremen, Germany
tenbrink@tzi.de

Shi Hui
I3-[SharC], SFB/TR 8
Spatial Cognition
U Bremen, Germany
shi@tzi.de

Kerstin Fischer
I1-[OntoSpace]
SFB/TR 8 Spatial Cognition
U Bremen, Germany
kerstinf@uni-bremen.de

Our aim is to enable natural and intuitive spoken dialogue between users and the Bremen robotic wheelchair "Rolland" (Röfer & Lankenau 1998), in order to instruct the robot to move to other places autonomously. The purpose of our interdisciplinary work is to develop specific functionalities matching potential users' intuitive expectations. Our current focus is on a detailed qualitative analysis of the discourse flow between human and robot, using a realistic interaction scenario with uninformed users that is tailored to the actual technological requirements. This approach is useful to establish and improve the relationship between implemented functionalities and humans' intuitive reactions at being confronted with an autonomous transportation device.

In earlier work (Shi & Tenbrink 2005) we identified a range of potential problems and devised a dialogue model to address them. Our dialogue modelling approach (see Ross et al. 2005) is based primarily on the COnversational Roles model (Sitter & Stein 1992) combined with the information state based approach (Traum & Larsson 2003). Figure 1 shows a depiction of a clarification subdialogue initiated by the robot, a part of the dialogue model that we start from in the present study. Following an utterance by the user, the robot can request something or inform the user; or it can make a suggestion, which can be rejected or accepted by the user. Following such a rejection or a request by the robot, the user instructs the robot, reaching the final state of this specific subdialogue.

Based on this model, we carried out a second study, this time in a Wizard-of-Oz scenario, to test the communicative success of a number of systematic robot reactions. Here we present the results of this study, outlining the range of problems that could successfully be handled by the robotic reactions, and pointing to a number of novel problems that arise precisely because of the robotic output. Our results show that high-level or generic robotic reactions will lead to increased confusion, while specific and aligned clarification questions enable smooth and efficient dialogue between humans and robots. This leads to an improved dialogue model.

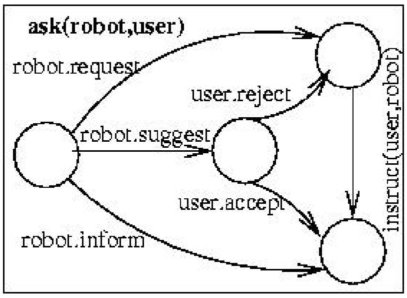


Figure 1: Clarification subdialogue

Our scenario resembles a situation in which new wheelchair owners need to acquaint themselves with their new device, since they are confronted with a robotic wheelchair without being informed in detail about its functionalities. Our experimental participants (17 German and 11 English native speakers) were told that the robot is able to augment its internal map by using the verbal information given by the users while moving around. They were first asked to familiarize the wheelchair with an environment (here: a university hallway with offices). After that, they instructed the robot to drive to one of the rooms they just encountered. In this way, we elicited route instructions related to a specific, relatively simple indoor setting, directed at a robot currently under development in our research group (the SFB/TR 8, funded by the DFG). In our study, the robot did not move autonomously, and the robotic utterances were triggered by a human "wizard" (unseen by the participant) according to a specifically devised schema. Using a range of preformulated utterances, the wizard thus produced a reasonably natural dialogue with the user without necessitating natural language generation while still sounding "automatic" (like a robot).

Our analysis shows that our proposed dialogue model is successful in encouraging the user to provide missing information and to use a suitable level of granularity. In fact, some of the dialogues turned out to be entirely unproblematic, they appeared to be completely natural and did not exhibit any communication problems whatsoever.

However, we also found that even slight confusions and temporal misplacements of the robot's utterances can lead to severe communication problems and distortions of the user's

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 163–164.

spatiotemporal representation. For example, getting back to the mental position in the route description is a problem if the robot's utterances are conceived as slightly incoherent. Therefore, clarification questions from the robot need to be formulated and placed with specific care. The clarification attempts by the robot work best for the discourse flow when they can be integrated into the user’s current mental representation of the spatial as well as the discourse situation. Thus, it is essential for the robot to align with the human’s utterances to a high degree. This may be even more important than in the interaction between humans: Since humans are specifically unsure about the robot’s capabilities, they sometimes turn to solutions that would be unnatural in a human-human interaction situation (Fischer 2006). For instance, in our data, some speakers returned to the very beginning of the spatial description in reaction to a mere clarification question. Such a discourse behaviour is very difficult to model even in very sophisticated models of clarification requests such as Schlangen (2004).

A further source for confusion is when the robot asks for clarification in an area where the user has good reasons to expect that it should have sufficient knowledge. Such cases can easily arise if knowledge already conveyed by the user could not be integrated properly by the system. Therefore, it is important that the robot informs the user about its current state of knowledge in as much detail as possible, and suggests a solution concerning how to proceed further. This will be specifically helpful in the case of spatiotemporal sequencing confusions. Also, it is important that the robot acknowledges what it has understood so far, to let the user know where exactly there is an information gap that needs to be filled in.

As a result, the dialogue model can be suitably extended. The precise discourse history is important since specific requests providing information about successfully integrated knowledge are more useful than generic clarification questions. In our improved model, we substitute the three simple dialogue acts, robot.request, robot.inform and robot.suggest (Figure 1) by subdialogues. Each subdialogue uses the current information state consisting of the discourse history and the internal map representation (denoted as [H,M]). Figure 2 represents the 'request' subdialogue as an example. First, the robot acknowledges the part of the instruction that it has understood, based on [H,M]. The user can react by rejecting this account and providing a further instruction which is integrated in the robot's internal model, in which case the robot does not formulate the request in the intended way. However, if the user does not react or reacts by accepting the robot's description, the robot continues by requesting information about entities, boundaries, orientations, or segments, depending on the current requirements, in a way that is aligned to the users' descriptions as much as possible (using the dialogue history). The dialogue will then continue with the user providing the missing information.

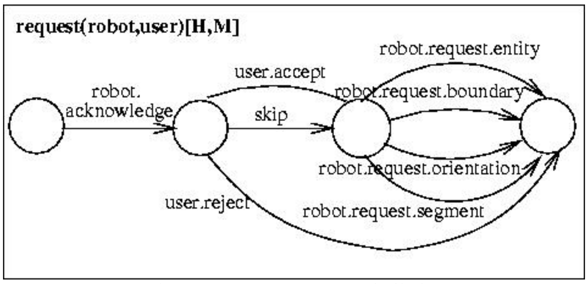


Figure 2: 'Request' subdialogue

References

Fischer, K. 2006. *What Computer Talk Is and Isn’t. Human-Computer Conversation as Intercultural Communication*. Saarbrücken: AQ.

Röfer, T. and A. Lankenau. 1998. Architecture and Applications of the Bremen Autonomous Wheelchair. In P. P. Wang (Hrsg.), *Proc. of the 4th Joint Conference on Information Systems, 1*, 365–368.

Ross, R.J., Bateman, J. and Shi, H. 2005. Using Generalized Dialogue Models to Constrain Information State Based Dialogue Systems. In *Proc. of the Symposium on Dialogue Modelling and Generation.* http://lubitsch.lili.uni-bielefeld.de/ DMG/Proceedings/proc.html.

Schlangen, D. 2004. Causes and Strategies for Requesting Clarification in Dialogue. *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Boston, USA, April.

Shi, H. and T. Tenbrink. 2005. Telling Rolland where to go: HRI dialogues on route navigation. Proc. *Workshop on Spatial Language and Dialogue (5th Workshop on Language and Space)*, October 23-25, 2005, Delmenhorst, Germany.

Sitter, S. and A. Stein. 1992. Modelling the Illocutionary Aspects of Information-Seeking Dialogues. *Information Processing and Management 28.* pp. 124-135.

Traum, D. and Larsson, S. 2003. The Information State based Approach to Dialogue Management. In *Current and New Directions in Discourse and Dialogue*. Kluwer Academic Publishers. pp. 325-353.

Non-verbal Information in Communication: A Study of Interaction in a Tourist-information Setting

Ichiro Umata*† **Sadanori Ito**$^{*\sharp}$ **Shoichiro Iwasawa***† **Noriko Suzuki***†
Tomoji Toriyama† **Naomi Inoue***† **Kiyoshi Kogure**† **Kenji Mase**†‡
*National Institute of Information and Communications Technology/ †ATR;
2-2-2, Hikaridai, Seikacho, Soraku, Kyoto, Japan
$^{\sharp}$Tokyo University of Agriculture and Technology; 2-24-16 Nakacho, Koganei, Tokyo, Japan
‡Nagoya University; Furocho, Chigusaku, Nagoya, Aichi, Japan

A preliminary analysis is done to characterize the overall tendency of communication by the occurrences of non-verbal behaviors throughout an entire interaction event. Based on the previous findings in cross-modal interaction(Argyle et al. (1976), Kendon (1967), Clark (1996), etc), we analyze the occurrences of non-verbal behaviors such as gaze, pointing, nodding, and body-posture in a tourist-information setting. Analyses show that an interaction event can be categorized by the occurrence pattern of non-verbal behaviors of the participants.

1 Tourist-information Experiment

An experiment was conducted in a tourist-information setting. Subjects (22 university students) who played the "customer" role were asked to obtain information on sightseeing spots from a professional information clerk (female, 30 years old). The information was given through the communication between a customer and a clerk in front of 7 information display panels with short descriptions and pictures of the spots. Subjects were able to walk around freely within the area. Nineteen successfully recorded sessions were analyzed in this paper.

Body motions and locations were measured by a set of Vicon Motion Capture System. The directions of the subjects' gazes were measured by EMR-8B head-mount eye-trackers of Nac Image Technology Inc.. Each subject wore an eye-tracker, a close proximity microphone, and markers for the motion capture device.

The number and total duration of the following situations were recorded for each subject: a) utterances, b) gazes at the parter or gazes at a panel, c) followed gaze in which a gaze at a panel was overlapped or followed within 0.5 seconds by the partner's gaze.

Also the number of the following situations were recorded for each subject: a) eye movements within each panel, b) nods, c) finger-pointings, d) occurrences of states when the distance between the head and a panel was less than 1000 mm.

2 Analysis1: Correlation between Non-verbal Behaviors

The correlation of the non-verbal behaviors of the customers and the clerk is analyzed from the viewpoint of mutual effects in joint activities.

Strong correlations are found between customers' gaze at the clerk and the clerk's gaze at the customers, both in number ($\rho = .664$, $p < .01$) and total duration($\rho = .637$, $p < .01$).

On the other hand, strong negative correlations were found between the total duration of joint gaze at display panels and gaze at the partner, again both in number (customer's gaze: $\rho = -.732$, $p < .01$; clerk's gaze: $\rho = -.562$, $p < 01$) and total duration (customer's gaze: $\rho = -.746$, $p < .01$; clerk's gaze: $\rho = -.635$, $p < 01$). Although joint gaze and gaze at a partner are both considered to play important roles in establishing common ground in communication, these activities are imcompatible, and so the participants must choose one of them in each occasion according to their communication styles and information aquisition strategies in this task setting.

Strong correlations were observed between the number of gazes at the partner and the number of the gazer's nods (customer's gaze: $\rho = .631$, $p < .01$; clerk's gaze: $\rho = .678$, $p < .01$). This is due to the tendency of people to nod while looking at their partner.

No strong correlations were found between the customers' utterances and the clerk's, either in number or total duration. This may be a result of the asymmetry in the amount of information between the customers and the clerk produced by the task setting.

A strong negative correlation was observed between the number and the total duration of the

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 165–166.

clerk's utterances($\rho = . -.546$, $p < .01$), whereas a strong positive correlation was observed between the number and the total duration of the customers' utterances. ($\rho = .955$, $p < .01$) This may be due to the fact that the customers' utterances are mainly spontaneous ones like questions or answers, while the clerk's utterances are mainly well-planned ones following the script. A large number of clerk utterances likely indicates that she felt some difficulty in communication and her speech was cut into short utterances; otherwise, it would have been long and fluent.

The number of the clerk's gazes at the customers as well as that of the clerk's nods also shows a strong negative correlation with the number of the clerk's utterances(gaze: $\rho = -.645$, $p < .01$; nods: $\rho = -.612$, $p < .01$) , and these results also support the possibility of difficult communication.

3 Analysis 2: Factor Analysis

In this section, we conduct a factor analysis of the occurences of their non-verbal behaviors based on the results. The number of occurrences of the behaviors that showed significant correlations were identified and standardized by the time of interaction. Factors were extracted by the principal factor method, and promax rotation was adopted. The factors with loading value of more than 0.5. were subjected to interpretation, and four factors were extracted by giving consideration to the decay of the eigenvalues. These factors were named as follows.

Customer-led: The Customer-led Factor is characterized by high loading of the customers' positive interaction activities such as utterances(.567), gaze at the clerk(.530), nods(.742), and closing up(.492). High loading on the Clerk's pointing(0.871) means active information exchange involving obvious non-verbal cues. The number of customer gaze movements also shows high loading(.838), and this indicates the customers' active attitude in interaction.

Cooperative: The Cooperative Factor is characterized by strong negative loading on the number of the clerk's utterances(-.932). The analysis in the previous section suggests that the small number of the clerk's utterances indicates fluent interaction between the clerk and the customer. The numbers of customers' gazes at the clerk(.627) and the clerk's gazes at the customers(.677) also show high loading. The number of the clerk's nods shows high loading(.579) only in this factor, and this could be regarded as a sign of smooth and cooperative interaction between the clerk and the customers.

Non-interactive: The Non-interactive Factor can be characterized by high loading on the customers' gaze at a display panel(.898). The numbers of followed gazes also show high loading in both directions, but especially high in the customer-first case(customer-first: 0.893; clerk-first: .559). This shows the customers' tendency to acquire information at their own pace. On the other hand, the number of customers' pointings shows strong negative loading(-.610). This indicates that the customers are not active in interacting with the clerk.

Clerk-led: The Clerk-led Factor can be characterized by high loading on the numbers of clerk's gazes at a panel(.855), at a customer(.601), the clerk's moving her face close to a panel(.763), and the clerk's gaze movements(.693). This shows the clerk's positive attitude in appealing to a customer.

4 Summary

We have analyzed the nonverbal behaviors in a tourist-information setting. A factor analysis of non-verbal behaviors revealed four factors of communication style from the viewpoint of interaction. These results shed some light on how to characterize communication based on the activity level and the initiative-taking pattern of interaction by analyzing the non-verbal cues of the participants.

Acknowledgments

This research was supported by The National Institute of Information and Communications Technology (NICT) of Japan. We would like to thank Takugo Fukaya, Kenji Susami and Roberto Lopez for their intellectual support of this work.

References

Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

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

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32*, 1–25.

Compliments in virtual dialogs

Lilia Mironovschi, Humboldt-Universität zu Berlin, lilia.mironovschi@hu-berlin.de

Abstract

This paper studies the realization of the compliments by native Russian and German speakers in electronic communication (Internet forums). Whereas the native German speakers spelled their compliments in a standard way and used normative vocabulary, the native Russian speakers mostly used the colloquial and jargon expressions and spelling.

1 Introduction

Beginning from the Manes und Wolfson's (1981) study of compliments in American English, numerous studies of compliments have been conducted with the aim to characterize their use in direct communication. Based on the recordings of natural talk, the studies of German compliment sequences (Golato, 2005) proved that compliments can not be analyzed irrespectively of the communicative situation in which they occur. In certain cultures, realization of compliments in public and private communicative situations differs significantly (Fukushima, 1990).

Whereas the direct interaction can be either public or private, the forms of communication occurring in Internet contain features of both public and private communication. The aim of the present study is to find out which pragmatic norms dominate in virtual compliments.

2 Methodology

A random selection of about 80 Russian and German forums (education-, popular-science-, art-, and literature-related) was searched for positive assessments of certain persons. In both Russian and German forums related to the photographic art compliments were found more often than in the others; correspondingly, 20 topliner photo forums resulted from the searches with Google and Yandex were selected. On the corresponding web-sites voting was organized to select the top-rated pictures, which should remain in the gallery. Viewer's ratings could be accompanied with comments. 128 Russian and 123 German compliments were found. Judging by the nicknames, the most of the forum participants were males in both cultures.

3 Results

Table 1 contains the expressions which occurred more than once in at least one of the cultures.

No	expression	Ru	De
1	очень	25	
very	sehr		25
2	!!!(+)	23	14
3	красив- (о, ый)	14	
nice(-ly)	schön		19
4 such	так (-ой), как (-ой)	21	
(what a)	so (-lch, ein, was…für ein)		10
5	класс	4	
fine	Klasse		24
6	супер (-ский)	14	
super	super		14
7	нравится, понравилось, нра	15	
I like it	gefällt		11
8 good	хорош- (о, ий)	13	
(well)	gut		12
9	здорово	8	
great	toll		15
10	чудо	1	
wonder	wunder (-)		9
11	интересн- (о, ый)	6	
interesting	interessant		2
12	вау	3	
wow	wow		5
13	впечатляет	3	
impressive	überzeugend, beeindruckend, eindrucksvoll		5
14	абсолютно, совершенно	0	
absolutely	absolut		8
15	исключительн- (о, ый)	3	
unique	einzigartig		4
16	отличн- (о, ый)	5	
excellent	ausgezeichnet		2
17	действительно	1	
really	echt		6
18	гениальн- (о, ый)	0	
brilliant	genial		7
19	круто, сильно	0	
cool	stark		6
20	безупречн- (о, ый)	0	
perfect	perfekt		5
21	симпатичн- (о, ый)	3	
pleasant	sympathisch		0
22	великолепн- (о, ый)	0	
imposing	großartig		3
23	с ума сойти, обалдеть	0	
crazy	Wahnsinn		3

Table 1. Numbers of occurrence of different expressions in Russian and German compliments

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 167–168.

The numbers shown in the Ru and De columns of the Table 1 were compared using the chi-square test. Certain equivalent expressions occurred in both languages with similar frequencies (No 1-3, 6-9 in Table 1). Some words or expressions occurred with significantly different frequencies (No 4 and 5) or were repeatedly used in only one language (in the latter case the chi-square test could not be performed).

As follows from the data presented above, 159 expressions of assessment were repeatedly found in the Russian compliments and 213 in the German ones (the totals of the corresponding columns of Table 1, single occurrences not counted). This suggests that the cliché constructions are more frequently used by German authors of compliments than by Russian ones (the difference is very significant). The conventionality of assessment expressions could also be traced at the syntactical level: though most of the sentences used in assessments were incomplete, there was a significant difference between 23 Russian and 37 German assessments realized only by means of full sentences.

The average length of Russian and German positive assessments was 6.4 and 12.5 words, correspondingly. That can only partially be explained by the absence of the definite and indefinite articles in Russian language. Syntactic compression of assessments was significantly higher in Russian than in German: 30 assessments consisting of a single word were found among Russian compliments, and only 7 – among German ones.

In Russian, the compression means typical for the Internet jargon were utilized as well, e.g. the single word *очень (very)* without an adjective or adverb used as an expression of assessment. Among the Russian compliments, 68 were made using the colloquial or Internet jargon words and/or spelling, i.e. *фЫлософски* (jargon spelling) ~ *philosophically*. All of the German compliments were realized using the standard spelling, only 4 of them contained colloquial expressions; all of the 3 neologisms found were placed in inverted commas. Thus, the frequencies of the slang use in Russian and in German Internet compliments were significantly different.

The most of the compliments were paid on the quality of the photo or author's skills; nevertheless, there were found 38 Russian and only 2 German positive assessments of not the author's work but of the persons, objects, or places pictured, the difference that should be considered as very significant. Moreover, only those assessments have been responded.

4 Discussion

Both Russian and German compliments have three primary functions:

a) contacting a person;
b) flirting with a person;
c) appraising the achievements of a person.

The third function (appraising the achievements) was dominating in the compliments found in the forums related to photographic art. However, the positive assessments of the photographed persons, objects, or places (which occurred mostly in Russian forums) were definitely made to contact other forum participants. That function was successfully realized in most cases. The less frequent use of cliché compliment expressions suggests that the function of contacting is more important for Russian compliments than for German ones, since the use of uncommon expressions when making new acquaintances is typical for Russian private communication. Probably, the frequent use of colloquial and jargon words and spelling in Russian compliments was also aimed to make them look more original and to induce verbal reaction to them.

5 Conclusion

Realized in more conventional way, the German compliments could be considered as public-oriented communicative tactics; the Russian compliments were both public- and personally-oriented.

Acknowledgement

The financial support of this work from Konrad Adenauer Foundation is gratefully acknowledged.

References

Andrea Golato. 2005. *Compliments and Compliment Responses. Grammatical structure and sequential organization*. John Benjamins, Studies in Discourse and Grammar 15.

Joan Manes, Nessa Wolfson. 1981. The compliment formula. In: Coulmas, F. (Ed.), *Conversational Routine*, pp. 115-132. Mouton, The Hague.

Norikazu J. Fukushima. 1990. *A Study of Japanese Communication: Compliment-Rejection Production and Second Language Instruction*. PhD. Dissertation, University of Southern California.

DISCUS: a dialogue simulation and context update system

Roser Morante and Simon Keizer
Department of Language and Information Science
Faculty of Arts
Tilburg University
`{r.morante,s.keizer}@uvt.nl`

In this abstract we present DISCUS (Dialogue Simulation and Context Update System), a research tool for simulating dialogues between a user and a system in terms of context update of the system's information state. DISCUS has been developed to test an algorithm for context updating, that builds on Dynamic Interpretation Theory (DIT), (Bunt, 2000). In DIT, meanings of dialogue utterances are viewed as intended context–changing effects that are determined by the dialogue act(s) being performed with the utterance. Dialogue acts in DIT are organized in a multidimensional dialogue act taxonomy (Bunt, 2006). DIT establishes four levels of understanding, that reflect the extent to which an utterance has been processed successfully by the Addressee: *perception, interpretation, evaluation,* and *execution*. The processing levels are also reflected in the dialogue act types in the auto- and allo-feedback dimensions.

The starting point for the model for context update are the preconditions of the dialogue acts, which represent the motivation and ability for an agent to perform a dialogue act. The preconditions are specified in terms of properties of the information state of the speaker. The model makes explicit how every dialogue act contributes to changing the information state, it defines the types of effects that an utterance provokes in dialogue participants, and it establishes the operations that cause the change of state in the context (Keizer and Morante, 2006). Additionally, the model can determine when information has been grounded.

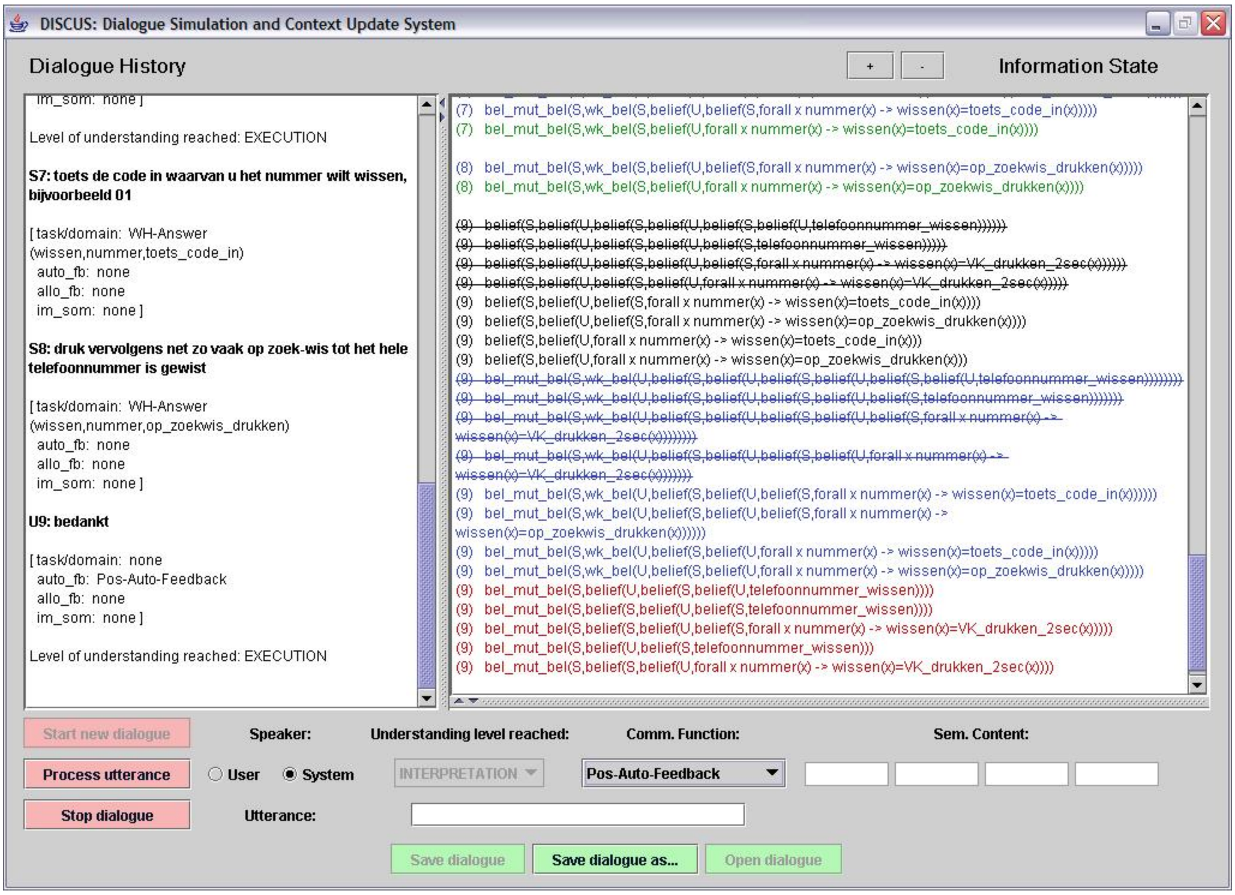
The primary use of the tool is to simulate the update of the context model of a dialogue system participating in a dialogue with a user. As the update algorithm itself assumes dialogue acts as input, the tool abstracts away from the processes of natural language understanding and generation. That is, both system and user utterances are represented in terms of dialogue acts, except for user utterances for which the system's level of processing reached is too low: processing problems on the levels of interpretation or perception will prevent the system from being able to associate the utterance with a dialogue act.

The functionality of the tool consists in: 1) allowing the researcher to simulate dialogues between a dialogue system (S) and a user (U), and 2) automatically updating the system's context model by applying the algorithm, and presenting it on the screen.

The interface of the tool (see Figure 1 for a screenshot) allows the researcher to specify system utterances in terms of dialogue acts and user utterances in terms of the level of understanding reached by the system and, provided that level is interpretation or execution, a dialogue act.

The components in the bottom part of the GUI can be used to specify the speaker of the utterance simulated, the system's understanding level reached (in case of a user utterance), a literal text representation of the utterance, and the communicative function (CF) and semantic content (SC) of the dialogue act performed in the utterance. At this moment, the SC is specified in a rather simplistic way by means of at most four slots, in which the parameters for the SC can be specified, depending on the CF. For example, the SC of a dialogue act with CF YN-QUESTION is specified with one parameter, representing the proposition the question is about: "will it rain tomorrow?" is represented as YN-QUESTION(`rain_tomorrow`); a dialogue act with CF WH-QUESTION requires two parameters, representing the property the

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 169–170.



Dialogue History

im_som: none]

Level of understanding reached: EXECUTION

S7: toets de code in waarvan u het nummer wilt wissen, bijvoorbeeld 01

[task/domain: WH-Answer
(wissen,nummer,toets_code_in)
auto_fb: none
allo_fb: none
im_som: none]

S8: druk vervolgens net zo vaak op zoek-wis tot het hele telefoonnummer is gewist

[task/domain: WH-Answer
(wissen,nummer,op_zoekwis_drukken)
auto_fb: none
allo_fb: none
im_som: none]

U9: bedankt

[task/domain: none
auto_fb: Pos-Auto-Feedback
allo_fb: none
im_som: none]

Level of understanding reached: EXECUTION

Information State

(7) bel_mut_bel(S,wk_bel(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=toets_code_in(x)))))
(7) bel_mut_bel(S,wk_bel(S,belief(U,forall x nummer(x) -> wissen(x)=toets_code_in(x))))

(8) bel_mut_bel(S,wk_bel(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=op_zoekwis_drukken(x)))))
(8) bel_mut_bel(S,wk_bel(S,belief(U,forall x nummer(x) -> wissen(x)=op_zoekwis_drukken(x))))

~~(9) belief(S,belief(U,belief(S,belief(U,belief(S,belief(U,telefoonnummer_wissen))))))~~
~~(9) belief(S,belief(U,belief(S,belief(U,belief(S,telefoonnummer_wissen)))))~~
~~(9) belief(S,belief(U,belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=VK_drukken_2sec(x))))))~~
~~(9) belief(S,belief(U,belief(S,belief(U,forall x nummer(x) -> wissen(x)=VK_drukken_2sec(x)))))~~
(9) belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=toets_code_in(x))))
(9) belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=op_zoekwis_drukken(x))))
(9) belief(S,belief(U,forall x nummer(x) -> wissen(x)=toets_code_in(x)))
(9) belief(S,belief(U,forall x nummer(x) -> wissen(x)=op_zoekwis_drukken(x)))
~~(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,belief(S,belief(U,belief(S,belief(U,telefoonnummer_wissen))))))))~~
~~(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,belief(S,belief(U,belief(S,telefoonnummer_wissen)))))))~~
~~(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=VK_drukken_2sec(x))))))))~~
~~(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,belief(S,belief(U,forall x nummer(x) -> wissen(x)=VK_drukken_2sec(x)))))))~~
(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=toets_code_in(x))))))
(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=op_zoekwis_drukken(x))))))
(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,forall x nummer(x) -> wissen(x)=toets_code_in(x)))))
(9) bel_mut_bel(S,wk_bel(U,belief(S,belief(U,forall x nummer(x) -> wissen(x)=op_zoekwis_drukken(x)))))
(9) bel_mut_bel(S,belief(U,belief(S,belief(U,telefoonnummer_wissen))))
(9) bel_mut_bel(S,belief(S,belief(U,belief(S,telefoonnummer_wissen))))
(9) bel_mut_bel(S,belief(S,belief(U,belief(S,forall x nummer(x) -> wissen(x)=VK_drukken_2sec(x)))))
(9) bel_mut_bel(S,belief(U,belief(S,telefoonnummer_wissen)))
(9) bel_mut_bel(S,belief(S,belief(U,forall x nummer(x) -> wissen(x)=VK_drukken_2sec(x))))

Speaker: ○ User ● System
Understanding level reached: INTERPRETATION
Comm. Function: Pos-Auto-Feedback
Sem. Content:
Utterance:

Figure 1: DISCUS Graphical User Interface.

value of which is asked for and the entity to which the property applies: “When does the train to Berlin leave?” is represented as WH-QUESTION(`depart_time,train_Berlin`).

The text panel in the top left part of the GUI displays the simulated dialogue (Dialogue History). For each utterance in the dialogue the literal utterance and speaker (S or U), the dialogue act information along four dimensions, and in case of user utterances, the processing level reached by the system, are indicated.

The text panel in the top right part of the GUI displays the beliefs and goals in the context model (Information State). The various kinds of beliefs (beliefs about understanding, adopted beliefs, beliefs about mutual beliefs about understanding and adoption, and strengthened beliefs) are displayed in different colors; cancelled beliefs get a ’strikethrough’ font. The information state panel can also be split in two, allowing to show any pending beliefs.

Besides using the GUI components to simulate dialogues and monitor the context model, the entered simulations can also be saved to file in an XML-format. Upon opening existing simulations, the context model is regenerated. In this way, a fixed set of simulations, covering a wide range of dialogue act patterns, can be used to efficiently test different context update models.

References

H. Bunt. 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue*, Studies in Computational Pragmatics, pages 81–150. John Benjamins.

H. Bunt. 2006. Dimensions in dialogue act annotation. In *Proceedings Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy. To appear.

S. Keizer and R. Morante. 2006. Dialogue acts as context operators constraining the generation of dialogical discourse. In *Proceedings of the Workshop on Constraints in Discourse*, pages 117–124, Maynooth, Ireland.

VISA - Corpus Annotation with OWL [*]

Stephanie Becker and **Thomas Kleinbauer** and **Stephan Lesch**
DFKI Gmbh
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
`<firstname.lastname>@dfki.de`

Abstract

We present VISA, a graphical annotation tool for OWL-based annotation schemes with a focus on generality and usability.

1 Introduction

The W3C standard OWL was originally designed as an ontology language for the semantic web, but it is progressively finding its way into various other fields of application. Annotated (linguistic) corpora, on the other hand, still often rely on their own specific data storage formats, although newer developments show a trend towards the use of XML (Carletta et al., 2005).

We believe that OWL is a suitable format for future corpora and annotations thereof, as it provides a semantically potent language based on a simple and open format. The main advantage is that further processing of corpus data can make use of automatic inference mechanisms, working only on one underlying formalism for all annotations. Existing annotation schemes can easily be expressed in OWL; annotation then becomes a process of assigning instances of ontology classes to corpus segments.

A number of tools specialized for different kind of annotations exist, as well as programs for working with OWL data. However, the number of tools for annotating OWL ontologies is rather small. One way to build such tools is to combine existing software for annotation and for OWL – a procedure taken for instance by (Bontcheva et al., 2004) or (Lauer et al., 2005) which both integrate the Protégé [1] editor for OWL into their own annotation framework.

But this approach suffers from the fact that Protégé was not originally designed for annotation work. Ontology instances, for example, are displayed as a flat list which makes it difficult for the annotator to discern which corpus segment was annotated with which instances. Relations between instances are displayed in a similar fashion. Furthermore, we found that Protégé reactivity decreases notably with increasing ontology size.

Hence, although a tool that combines existing programs is commendable in principal, practical application may prove very difficult under certain circumstances in which the user might prefer a tool tailored specifically to annotation with OWL. Furthermore, these observations illustrate the importance of good usability for annotation tools.

2 The VISA Annotation Tool

Based on the analysis of deficiencies of existing annotation tools we derived a first requirements specification for a new tool which was followed by the development of a prototype. The further development process has been accompanied by further theoretical considerations with respect to the possible extension of the requirements specification. Moreover we have conducted practical evaluations in form of repeated testing and the prototype has continuously been adapted according to the extended requirements specification.

The following screenshot displays the VISA tool. On the left hand side the classes of the ontology are displayed with their hierarchical relationships, on the right hand side the relation hierarchy of the ontology is shown. In the middle of the window an annotation panel and the text segments that are to be annotated are displayed.

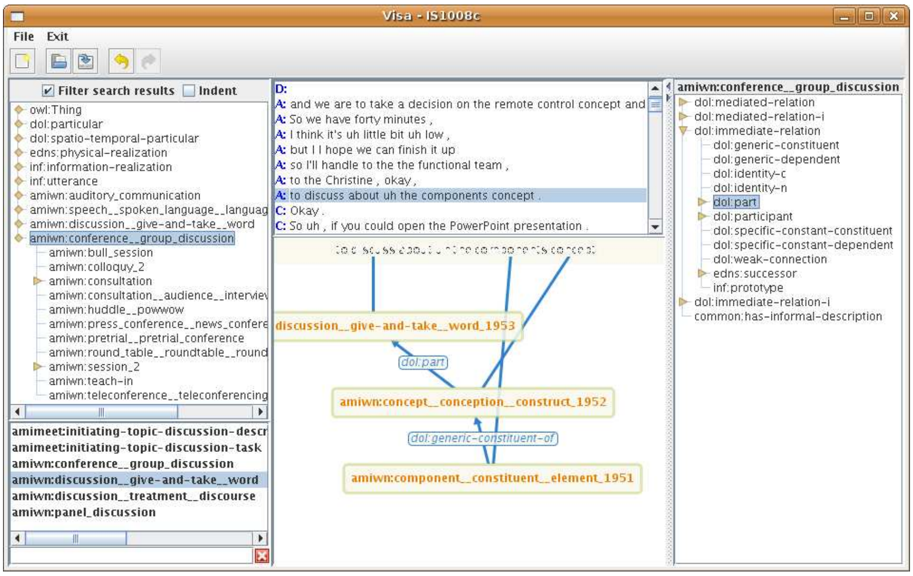
To create a class instance during the annotation process, the corresponding class is selected in the

---

[*] The research presented here is funded by the EU under the grant FP6-506811 (AMI).

[1] http://protege.stanford.edu

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 171–172.



Visa - IS1008c
File
Exit
owl:Thing
dol:particular
dol:spatio-temporal-particular
edns:physical-realization
inf:information-realization
inf:utterance
amiwn:auditory_communication
amiwn:speech__spoken_language__languag
amiwn:discussion__give-and-take__word
amiwn:conference__group_discussion
amiwn:bull_session
amiwn:colloquy_2
amiwn:consultation
amiwn:consultation__audience__interviev
amiwn:huddle__powwow
amiwn:press_conference__news_confere
amiwn:pretrial__pretrial_conference
amiwn:round_table__roundtable__round
amiwn:session_2
amiwn:teach-in
amiwn:teleconference__teleconferencing
amimeet:initiating-topic-discussion-descr
amimeet:initiating-topic-discussion-task
amiwn:conference__group_discussion
amiwn:discussion__give-and-take__word
amiwn:discussion__treatment__discourse
amiwn:panel_discussion
discussion__give-and-take__word_1953
dol:part
amiwn:concept__conception__construct_1952
dol:generic-constituent-of
amiwn:component__constituent__element_1951
amiwn:conference__group_discussion
dol:mediated-relation
dol:mediated-relation-i
dol:immediate-relation
dol:generic-constituent
dol:generic-dependent
dol:identity-c
dol:identity-n
dol:part
dol:participant
dol:specific-constant-constituent
dol:specific-constant-dependent
dol:weak-connection
edns:successor
inf:prototype
dol:immediate-relation-i
common:has-informal-description

☑ Filter search results ☐ Indent

D:
A: and we are to take a decision on the remote control concept and
A: So we have forty minutes ,
A: I think it's uh little bit uh low ,
A: but I I hope we can finish it up
A: so I'll handle to the the functional team ,
A: to the Christine , okay ,
A: to discuss about uh the components concept .
C: Okay .
C: So uh , if you could open the PowerPoint presentation .

hierarchy. An instance of the selected class is then created on the annotation panel by drag and drop.

Class instances can be connected with one or several words of the current text segment by dragging from the instances to the words. Relations between instances can be annotated by selecting a relation from the relation hierarchy and dragging from the instance of the corresponding domain class to the instance of the range class.

The graphical instances are arranged automatically on the annotation panel, thus the annotator does not need to take care of the graphical layout of the annotation. To facilitate navigation in the ontology, keyword search functions are available.

Visa is capable of dealing with large-sized ontologies without slowing down the annotation process. One of the ontologies we tested Visa with , e. g., contains more than 60.000 concepts.

Visa is based on NXT (Carletta et al., 2003) which supports the development of corpus tools through the provision of an open source Java API. However, through its modular architecture, Visa allows the integration of other data formats as well.

3 Conclusion and Future Work

We developed a tool for the annotation of text segments with OWL-based ontologies, focussing on a rich feature set an good usability. Visa can deal with large-sized ontologies without slowing down the annotation process.

Visa requires that the text to be annotated is pre-segmented. Furthermore an already existing ontology is required. As our primary concern is to provide an appropriate tool for annotation, Visa does not provide functions for creating or editing ontologies, nor for segmenting or editing of the corpus.

Currently, Visa should still be considered as a prototype. Several features are planned to be added, particularly with regard to the further facilitation of the annotation process, but also features like a reasoning function in order to prohibit inconsistent annotations.

References

K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. 2004. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10(3/4):349–373.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voormann. 2003. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*, 35(3):353–363.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: A pre-annoncement. In *Proceedings of MLMI 2005*.

Christoph Lauer, Jochen Frey, Benjamin Lang, Jan Alexandersson, Tilman Becker, Thomas Kleinbauer, and Harald Lochert. 2005. Amigram–a general-purpose tool for multimodal corpus annotation. In *Proceedings of MLMI 2005*, Royal College of Physicians, Edinburgh, UK, 11-13 July.

Browsing Meetings: Automatic Understanding, Presentation and Feedback for Multi-Party Conversations*

Patrick Ehlen, Stéphane Laidebeure, John Niekrasz,
Matthew Purver, John Dowding and **Stanley Peters**
Center for the Study of Language and Information
Stanford University
Stanford, CA 94305, USA
{ehlen, laidebeu, niekrasz, mpurver, jdowding, peters}
@csli.stanford.edu

Abstract

We present a system for extracting useful information from multi-party meetings and presenting the results to users via a browser. Users can view automatically extracted discussion topics and action items, initially seeing high-level descriptions, but with the ability to click through to meeting audio and video. Users can also add value: new topics can be defined and searched for, and action items can be edited or corrected, deleted or confirmed. These feedback actions are used as implicit supervision by the understanding agents, retraining classifier models for improved or user-tailored performance.

1 Introduction

Research on multi-party dialogue in meetings has yielded many *meeting browser* tools geared toward providing visual summaries of multimodal data collected from meetings (Tucker and Whittaker, 2005). Why create another? Existing tools focus on facilitating manual annotation and analysis of abstracted knowledge, or on assisting the meeting process by allowing users to conveniently (but manually) add relevant information online.

Because our aim in the CALO Meeting Assistant project is to automatically extract useful information such as the topics and action items discussed during meetings, our meeting browser has a different goal. Not only do we need an end-user-focused interface for users to browse the audio, video, notes, transcripts, and artefacts of meetings, we also need a browser that presents automatically extracted information from our algorithms in a convenient and intuitive manner. And that browser should allow – even compel – users to modify or correct information when automated recognition falls short of the mark.

2 Automatic Understanding

User studies (Banerjee et al., 2005) show that amongst the most requested pieces of information from a meeting are the *topics* discussed and *action items* established.

Action Item Identification. Our understanding suite therefore includes an agent for action item identification – see (Purver et al., 2006). We exploit a shallow notion of discourse structure, by using a hierarchical combination of supervised classifiers. Each sub-classifier is trained to detect a class of utterance which makes a particular discourse contribution to establishing an action item: proposal or description of the related *task*; discussion of the *timeframe* involved; assignment of the responsible party or *owner*; and *agreement* by the relevant people. An overall decision is then made based on local clusters of multiple discourse contributions, and the properties of the hypothesized action item are taken from contributing utterances (the surface strings, semantic content or speaker/addressee identity). Multiple alternative hypotheses about action items and their properties are provided and scored using the individual sub-classifier confidences.

Topic Identification. Another agent splits meetings into topically coherent segments, providing models of the associated topics using vector space

---

*This work was supported by DARPA grant NBCH-D-03-0010. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 173–174.

models. Topics are extracted as probability distributions over words, learnt over multiple meetings and stored in a central topic pool; they can then be used for audio/video browsing (labelled via the top most distinctive words) or to interpret a user keyword or sentence search query (by finding the weighted mixture of learnt topics which best match the words of the query).

3 User Interface

Agents that generate multiple hypotheses fare better with feedback from users about which hypotheses sound reasonable, but getting that feedback isn't always easy. A meeting browser is the ideal place to solicit feedback from end-users about what happened during a meeting. Our browser interface exploits the *transparency of uncertainty* principle, which counts on people's tendency to feel compelled to correct errors when those errors are (a) glaringly evident, and (b) correctable in a facile and obvious way.

A user can view action items detected from the meeting in the browser and drag them to a bin that adds the items to the user's to-do list. For the properties of action items – such as their descriptions, owners, and timeframes – the background colors of hypotheses are tied to their sub-classifier confidence scores, so less certain hypotheses are more conspicuous. These hypotheses respond to mouse-overs by popping up the most likely alternate hypotheses, and those hypotheses replace erroneous ones with a simple click. If an entire action item is rubbish, one click will delete it and provide negative feedback to our models. A user who just wants to make a reasonable action item disappear can click an *ignore this* box, which will still provide positive feedback to our model.

Topics appear as word vectors (ordered lists of words) for direct browsing or to help with user-defined topic queries. Given a user search term, the most likely associated topics are displayed, together with sliders that allow the user to rate the relevance of each list of words to the actually desired topic. As the user rates each topic and its words are re-weighted, a new list of the most relevant words appears, so the user can fine-tune the topic before the browser retrieves the relevant meeting segments.

4 Learning from Feedback

Action Item Feedback. The supervised action item classifiers can be retrained given utterance data annotated as positive or negative instances for each of the utterance classes (task description, timeframe, owner and agreement). User confirmation of a hypothesized action item allows us to take the utterances used to provide its properties as positive instances; conversely, deletion allows us to mark them as negative instances. Switching from one hypothesis to another for an individual property allows us to mark the utterances corresponding to the accepted hypothesis as positive, and the others as negative. Creation of a new action item, or manual editing or insertion of a property value requires us to search for likely utterances to treat as corresponding positive evidence; this can be done by using the relevant sub-classifier to score candidate utterances, and/or by string/synonym comparison, depending on the property concerned. Feedback therefore provides implicit supervision, allowing re-training models for higher accuracy or user-specificity.

Topic Feedback. The topic extraction and segmentation methods are essentially unsupervised and therefore do not need to use feedback to the same degree. Yet even here we can get some benefit: as users define new topics during the search process (by moving sliders to define a new weighted topic mixture), these new topics can be added to the topic pool. They can then be presented to the user (as a likely topic of interest, given their past use) and used in future searches.

References

S. Banerjee, C. Rosé, and A. Rudnicky. 2005. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human-Computer Interaction*.

M. Purver, P. Ehlen, and J. Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proceedings of the 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

S. Tucker and S. Whittaker. 2005. Accessing multi-modal meeting data: Systems, problems and possibilities. In S. Bengio and H. Bourlard (Eds.), *Machine Learning for Multimodal Interaction: First International Workshop, 2004*, v. 3361 of *Lecture Notes in Computer Science*, 1–11. Springer-Verlag.

Scene-Sentence Integration:
Incremental Effects of Mismatch and Scene Complexity

Pia Knoeferle
Dept. of Computational Linguistics
Saarland University, Germany
knoeferle@coli.uni-sb.de

Monica Rodriguez
Dept. of Computational Linguistics
Saarland University, Germany
monic@coli.uni-sb.de

Abstract

We monitored eye movements in a scene during spoken sentence comprehension to investigate the effects of different types of scene-sentence mismatch (action vs. role relations) and of scene complexity on comprehension. Gaze analyses revealed rapid effects of both role relations mismatch and scene complexity, while effects of action mismatch were slightly delayed.

1 Introduction

Verification-task studies have reported longer response latencies (e.g., Just & Carpenter, 1971) and gaze durations (Underwood, Jebbett, & Roberts, 2004) for resolution of a sentence-picture mismatch compared with a match, suggesting a mismatch is more complex to process than a match. We extended the mismatch approach by investigating how different types of scene-sentence mismatch (action versus role relations mismatch, Experiment 1), as well as scene complexity (Experiment 2) affect incremental thematic interpretation. To obtain further insights into the time-course of scene-sentence integration, we monitored participants' eye movements in a scene during comprehension of a related utterance.

2 Experiment 1

2.1 Method

Twenty-four German native speakers with normal vision received each five euro for experiment participation. There were 24 items. Presenting the sentence in Table 1 with the four images in Fig. 1 (A to D) created four conditions (see Table 1).

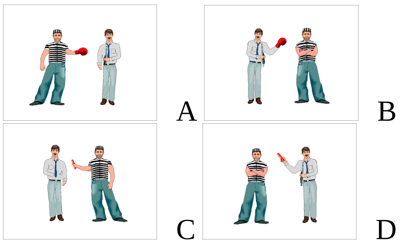


Figure 1: Example Item Images

Sentence & Fig.	Role	Action
1A Der Sträfling boxt gerade den Flötisten 'The convict (S) punches currently the flautist (O)'	Match	Match
1B Der Sträfling boxt gerade den Flötisten 'The convict (S) punches currently the flautist (O)'	Mism.	Match
1C Der Sträfling boxt gerade den Flötisten 'The convict (S) punches currently the flautist (O)'	Match.	Mism.
1D Der Sträfling boxt gerade den Flötisten 'The convict (S) punches currently the flautist (O)'	Mism.	Mism.

Table 1: Example Item Sentences

For counter-balancing reasons, one item had two sentences and four images, resulting in eight experimental lists. Items were rotated across lists such that no participant saw more than one version of each item, and such that each condition appeared equally often in each list. Consecutive experiment trials were separated by at least one of 48 filler trials. An SMI EyeLink I head-mounted tracker monitored participants' gaze in the scene during spoken comprehension. There was no verification task. Rather, participants were instructed to try to understand both sentences and depicted scenes. For half of the 48 filler trials, a written yes/no question about the sentence ensured that people performed a comprehension task. We report analyses of gaze durations that started in the ADV (from adverb onset to the onset of the second noun phrase), and NP2 regions. During these time regions the available scene and utterance information should permit resolution of both the action and role mismatch. If these two types of mismatch rapidly affect thematic interpretation, then their effects should be reflected in the inspection

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 175–176.

durations on the target characters (the scene agent, ‘the convict’, and patient, ‘the flautist’) during the analyses regions.

2.2 Results and Discussion

The key finding is the rapid effect of the role relations mismatch on thematic interpretation as evidenced by an interaction between target character (agent, patient) and role mismatch in the ADV region ($ps < 0.01$, see Fig. 2). People inspected the patient longer than the agent for a role match (C1 & C2, Fig. 2), while there was no such difference for a role mismatch. In contrast, there was no reliable effect of action mismatch in the ADV region. For the NP2 region, there were no reliable effects of the mismatch regarding gaze durations on the target characters.

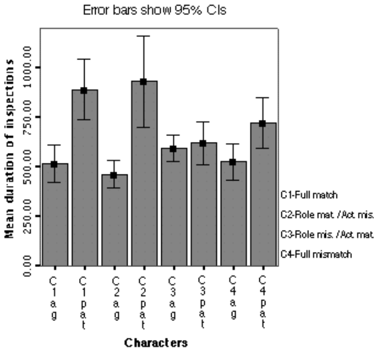


Figure 2: Mean inspection durations to the target characters for the ADV region in Experiment 1

3 Experiment 2

Experiment 2 reused the materials from Experiment 1 but retained only the action mismatch to verify its effects independent of the role relations mismatch. We further examined the influence of scene-complexity (simple vs. complex) on scene-sentence integration. Simple scenes contained the two target characters (agent, patient) of Experiment 1 and four distractor objects. Complex scenes showed an additional three characters.

3.1 Method

Thirty-two further participants from the same population as in Experiment 1 were each paid five euro. Procedure, task, and the analyses regions were the same as in Experiment 1. In addition, we examined early effects of scene complexity by analyzing the duration of inspections that started after NP1 and before verb onset.

3.2 Results and Discussion

There was a main effect of scene complexity for NP1 ($ps < 0.01$), with longer inspection durations on target characters (agent, patient) for simple than complex images. During the ADV region we found no effects of either action mismatch or scene complexity. For NP2, there was an interaction of mismatch and target character ($ps < 0.001$): people fixated the patient longer than the agent for the action-match conditions (C1 & C3). For action-mismatch conditions (C2 & C4), in contrast, inspection duration on the agent and patient did not differ (Fig. 3).

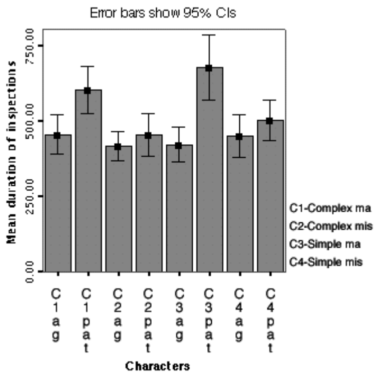


Figure 3: Mean inspection durations to target characters for the NP2 region in Experiment 2

4 Conclusions

Taken together, our findings support the view that scene-sentence integration takes place incrementally. There were, however, differences in the time course of processing actions and role relations mismatch: While the role relations mismatch influenced thematic interpretation post-verbally, effects of the action mismatch only affected thematic interpretation later, during the NP2 region. Scene complexity did not interact with action mismatch, but influenced the inspection duration of the target characters during NP1.

References

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with qualification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244–253.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, *56*, 165–182.

Perspective guides interpretation of questions, declarative questions and statements in unscripted conversation

Sarah Brown-Schmidt*
Department of Brain and Cognitive Sciences
University of Rochester
brownsch@uiuc.edu

Christine Gunlogson
Department of Linguistics
University of Rochester
gunlog@ling.rochester.edu

Duane G. Watson*
Michael K. Tanenhaus
Department of Brain and Cognitive Sciences
University of Rochester

Abstract

This paper describes research investigating the on-line production and interpretation of questions, declarative questions, statements and their replies. Specifically, we examine the role of shared and private knowledge in the processing of these constructions in unscripted conversation. Questions provide a critical test case for the use of perspective in language processing because their felicitous use requires speakers to distinguish common from private knowledge. Analyses of speech and gaze demonstrate that interlocutors distinguish shared from private information and that attention is directed toward different types of entities depending on utterance form. We argue for a central role of perspective in language processing. Discrepancies in experimental findings regarding use of perspective are discussed in terms of relevance of perspective to the task and the utterances of interest.

Cooperative speakers ask questions when they don't know the answer, but believe their addressee might. They assert things they know but believe their addressee might not know. Since Stalnaker's pioneering work on mutual knowledge (Stalnaker, 1978), formal theories of discourse in computational linguistics and within pragmatics and semantics have assumed that keeping track of shared and private commitments and knowledge is central to conversation (Clark, 1992).

While the presuppositions tied to use of different constructions suggest that the distinction between private and shared knowledge is basic to language processing, addressees often fail to distinguish shared from private information (Keysar, Lin and Barr, 2003), and when they do, the egocentric perspective can interfere with reference interpretation (Hanna, Tanenhaus & Trueswell, 2003). However, this and other online work on perspective used imperatives, which may encourage egocentrism due to authority-induced suspension of skepticism and the addressee's aim not to appear confused. Additionally, in order to have control over the interaction and generate specific experimental utterances, these experiments typically employ confederate speakers who are practiced and knowledgeable about the task. However, there is reason to believe that participants interact with confederates differently than they interact with another naïve participant (see Lockridge & Brennan, 2001).

In the experiment described in this paper, we used a goal-directed interactive conversation to examine five semantic-syntactic forms (a-e, see Table 1) that differ in discourse function (requesting/ imparting/ confirming information). Using interactive conversation between naïve participants assures that the constructions are appropriate for the linguistic context and for the knowledge states of the two participants. Thus, speakers will only ask questions when they really don't know the answer, and only make statements when they do. Examining utterance forms which presuppose a distinction between speaker and hearer knowledge (e.g. questions and replies) should provide insights into whether and when this information is used as language is processed on-line.

---

*Sarah Brown-Schmidt and Duane Watson are now at the Department of Psychology, University of Illinois at Urbana-Champaign.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 177–178.

a	Wh-Question	What's next to the pig with the hat?
b	Statement	There's a cow with shoes next to the pig with the hat.
c	Declarative question	It's a cow with shoes?
d	Question response	(*What's next to the pig with the hat?*)..A cow with shoes.
e	Acknowledgment	(*There's a cow with shoes.*)… A cow with shoes.

We examined the on-line interpretation of wh-questions, declarative questions and statements, and the on-line production of question responses and acknowledgments. Wh-questions and statements were selected to have parallel syntactic structures; each asked about or mentioned the location of one entity (target) with respect to another previously mentioned entity (anchor). If the distinction between shared and private perspectives can be used on-line, we would expect that addressees would direct attention toward private information as they interpret wh-questions, and towards shared or speaker-private information for statements.

Declarative questions, or rising declaratives (Gunlogson, 2001) were used because they have the syntactic form of a declarative, but have question-like intonation and distinct discourse functions. In this task, participants typically used declarative questions to request confirmation or to express skepticism (e.g. *That's a cow with shoes?*). We expected the interpretation pattern for declarative questions to share similarities with both wh-questions and statements.

The question responses and acknowledgments shared a similar syntactic structure (typically a bare noun phrase), however we expected that speakers would direct more attention to private entities when preparing question responses and to shared entities when preparing acknowledgments.

Our results demonstrate that the distinction between shared and private game-pieces is reflected in referent-type differences across utterance forms, and on-line production and interpretation of utterances with different discourse functions.

Wh-questions primarily inquired about addressee-private game-pieces, whereas statements were about shared or speaker-private game-pieces. The pattern of referent-types for declarative questions was half-way between that for wh-questions and statements, with declarative questions primarily inquiring about addressee-private game-pieces and sometimes about shared or speaker-private game-pieces.

When we analyzed the fixations that addressees made as they interpreted these expressions, we saw evidence for a distinct interpretation pattern for wh-questions: Fixations to addressee-private and shared game-pieces were initially equivalent, but following reference to the anchor, addressee-private fixations rose and shared fixations dropped. In contrast, for statements, most fixations were directed to shared game-pieces, suggesting that addressees distinguish shared and private information during on-line interpretation, and direct attention to information relevant for the type of utterance being interpreted.

The relationship between referent type and utterance form confirms our assumptions about the felicity conditions associated with questions and statements. More importantly, using goal-directed conversation and naïve participants, we demonstrated that interlocutors take into account each other's perspective when producing and comprehending utterances for which perspective is relevant. Differences in experimental findings regarding the use of perspective in on-line language processing may be best understood by considering whether perspective was relevant to the task and relevant for interpreting the critical utterances. Continued work using a variety of syntactic structures and communicative situations is needed to understand more precisely when perspective is and is not used in language processing.

References

Clark, H. H. (1992). *Arenas of Language Use*. Chicago: University of Chicago Press.

Gunlogson, C. (2001). *True to Form: Rising and Falling Declaratives as Questions in English.* Unpublished Doctoral Dissertation, University of University of California, Santa Cruz, Santa Cruz, CA.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49, 43-61.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25-41.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin and Review*, 9, 550-557.

Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), Syntax and semantics: *Pragmatics* (Vol. 9, pp. 315-332). New York, NY: Academic Press.

Prosodic marking of contrasts in information structure

Markus Guhe, Mark Steedman,
Ellen Gurman Bard
Human Communication Research Centre
University of Edinburgh
{m.guhe; m.steedman; e.bard}
@ed.ac.uk

Max M. Louwerse
Department of Psychology /
Institute for Intelligent Systems
University of Memphis
mlouwers@memphis.edu

1 Prosodic marking of contrast

Successful dialogue requires cultivation of common ground (Clark, 1996), shared information, which changes as the conversation proceeds. Dialogue partners can maintain common ground by using different modalities like eye gaze, facial expressions, gesture, content information or intonation. Here, we focus on intonation and investigate how contrast in information structure is prosodically marked in spontaneous speech.

Combinatory Categorial Grammar (CCG, Steedman 2000) distinguishes *theme* and *rheme* as elements of information structure. In some cases they can be distinguished by the pitch accent with which the corresponding words are realised. We experimentally evoke instances of contrasting themes and rhemes to establish the circumstances under which the pitch accents occur in unrestricted spoken dialogue. ‘Contrast’ means ‘alternatives are available’, not ‘contrastive accent’. It is difficult to manipulate context or outcome in quasi-natural engaging situations. Even if contrasting themes and rhemes are available, speakers choose from among a wider set of contrastable elements when framing utterances. Their choice may be difficult to predict: contrasts not apparently critical to the local context may be as important to speakers as ones usually thought to define the situation under discussion.

Unscripted dialogue with pressing communicative motivation is difficult to control for genre, topic, and goals. We use a modified map task (Anderson et al. 1991), a restricted-domain route-communication task, which establishes what each participant knows at any time. Without sight of each other’s maps, an Instruction Giver (IG) and Follower (IF) collaborate to reproduce on IF’s map a route printed on IG’s. The route can be adequately described by route-critical landmarks. As Fig. 1 illustrates, map pairs differ in the features of landmarks and in ‘ink damage’ that obscures the colours of some landmarks on IF’s map. Participants know that maps can differ but must learn where and how.

The discrepancies between maps do not fully define the alternatives sets speakers may wish to contrast. Instead, speakers define that alternatives set by their intonation. Provided that it is consistent with the context, the hearer will accommodate that set. Take:

(1) IF: *Do you see the two brown trees and the and the four black trees?*
IG: *You mean* THREE *black trees right?* (1:1–2:T:700.7; 1–1)

By deaccenting ‘black’ and ‘trees’ IG presupposes that the alternatives are confined to sets of black trees; specifically to IG’s set of three and IF’s set of four. Both can then adjust common ground incrementally.

As there is intense debate about whether the involved pitch accents (L+H* and H*) are actually categorically distinct (Ladd & Schepman 2003, Calhoun 2004), we simply seek to establish that contrasts in the information structure are indeed marked overtly by some form of prominence. We therefore use an undifferentiated notion of perceptual prominence to determine whether contrasts are marked by phonetic means.

Our prediction is the following: Only words whose denotation contributes to distinguishing the entity referred to from the other entities in the alternatives set are marked by prominence.

2 Experiment

Key-objects (here: trees) provide the route-critical landmarks for a map. They differ among a single map’s landmarks by colour and by one other feature (here: number). We report findings for two dialogues for the maps in Fig. 1 in order to identify episodes containing the predicted contrasts. (We superficially looked at others, which corroborated our findings.) The results are consistent within and between participant dyads. Landmarks differ in colour of tree groups; group size (1 to 5), presence of the group on IG’s /IF’s map, whether ink obscures the colour on IF’s.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 179–180.

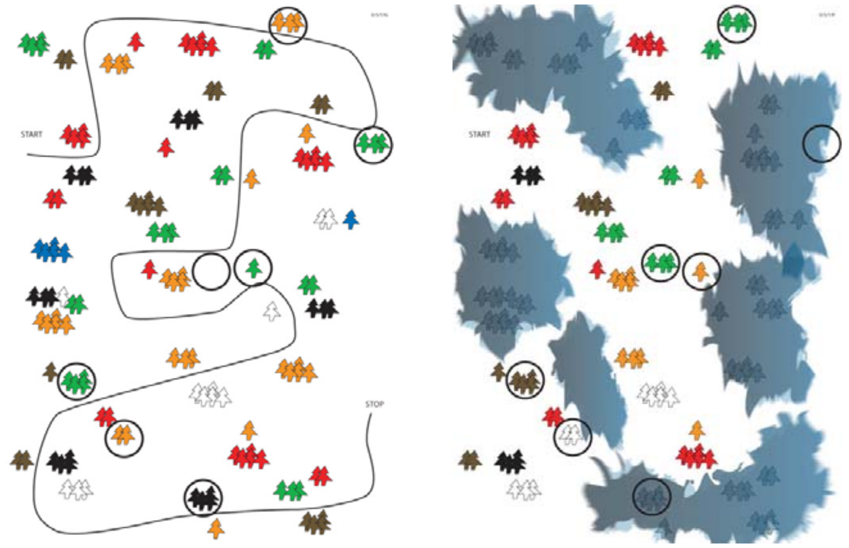


Figure 1: Maps for the analysed dialogues; IG's map (left) contains a route and a START and STOP mark; IF's map contains 'ink blots' that obscure the colour of some objects; circles (added here for expository purposes) indicate the differences between the maps

We assessed perceptually whether the mentioned items are prominent. For landmarks differing between maps (except those inked out) we also established the most prominent item of the intonation phrase – the contrasted element.

The material contains 146 intonational phrases that mention one or two landmarks in the form [number] [colour] ['tree'/'one'] and where at least one of [number] or [colour] is present. There are 334 mentions of features (e.g. 'red', 'two') in these phrases. In only 6 mentions is the feature term *non*-prominent, but not all prominences are realised by pitch movement. Seven differences between the maps are unrelated to ink-blots: 4 colour differences, 1 number difference, 1 landmark present only on one map, respectively. They are the prime place for eliciting contrasting intonation that correct the dialogue partner's knowledge representation, cf (1). Of the 146 phrases, 9 refer to differences between maps.

The phrases include 210 mentions of landmarks, of which 124 mention both features. There is no clear preference for assigning prominence to features (86 use equal prominence; 21 make the number term more prominent, 17 the colour term). Number mentions predominate in single-feature mentions (65 number vs 21 colour). This appears to be a response to the fact that number is the more reliable feature. 137 phrases describe landmarks on a single map, of which 131 instances mention landmarks within the 'magic circle', an imaginary circle around the current position that contains the landmarks identifying the next leg. Of the other 6, 4 are close to the circle and 2 are only in the discourse history.

The two dialogues mention 9 of the 14 possible differences between maps; in 8 cases a pitch accent marks the contrast. In 2 instances the participants are off-route. So, the speakers could have chosen to mention 12 differences between the maps. The ratio of 9(8)/12 is very satisfactory.

3 Discussion

In this exploratory evaluation we looked at places in the maps that are prone to prompt intonation patterns marking a contrast in the information structure. Differences *within* one map do not seem to elicit prosodic structures that mark contrasts between landmarks. These mentions are only informing or describing. Differences *between* maps require to correct the dialogue partner's knowledge representation and to introduce new information into the common ground. These contrasting items receive the most prominent pitch accent. With the exception of Ito et al (2004) we are not aware of experimental settings that can elicit 9 of 12 possible contrasts in unrestricted dialogue. In contrast to reading sentence lists this will provide deeper insight into actual dialogue.

Acknowledgements

This research was supported by the National Science Foundation (IIS-0416128)**.**

References

Anderson, A., et al (1991). The HCRC Map Task Corpus. *Language and Speech*, 34: 351–366.

Calhoun, S. (2004). Phonetic dimensions of intonational categories: The case of L+H* and H*. In: *Proceedings of Prosody 2004*, Nara, Japan.

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

Ito, K., Speer, S. R., & Beckman, M. E. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English. In: *Proceedings of Speech Prosody 2004*, 279–282.

Ladd, D. R. & Schepman, A. (2003). 'Sagging Transitions' between high pitch accents in English. *Journal of Phonetics*, 31(1): 81–112.

Steedman, M. (2000). *The syntactic process*. MIT Press, Cambridge, MA.

Evaluation of an Information State-Based Dialogue Manager

Antonio Roque, Hua Ai[†], and David Traum

USC Institute for Creative Technologies
13274 Fiji Way
Marina Del Rey, CA 90292
`{roque,traum}@ict.usc.edu`

[†]Intelligent Systems Program
University of Pittsburgh
210 S. Bouquet, Pittsburgh PA 15260
`hua@cs.pitt.edu`

Abstract

We describe an evaluation of an information state-based dialogue manager by measuring its accuracy in information state component updating.

1 Introduction

Evaluation of dialogue managers is essential for the development of dialogue systems. However, it can be difficult to separate the performance of a dialogue manager from the performance of the system as a whole. Here we describe an approach towards evaluating the performance of an Information State-based dialogue manager separately from the other components of the dialogue system and the system as a whole.

Our testbed system, Radiobot-CFF (Roque et al., 2006), is a military virtual reality environment designed to train soldiers in artillery strike requests. The trainees hold a radio dialogue with Radiobot-CFF during which an enemy target is located and attacked. Radiobot-CFF includes a speech recognition component, a dialogue move interpreter, and an information state-based dialogue manager (Roque and Traum, 2006). We ran an evaluation of the system from which we calculated task completion rates and time-to-task measures for the system as a whole, as well as error rates for the speech recognition and interpreter components (Robinson et al., 2006). However, we lacked an analysis of the dialogue manager component's performance.

2 Evaluation

Radiobot-CFF uses an information state-based (Traum and Larsson, 2003) dialogue manager, and therefore works by firing update rules which are dependent on and which change information state components. For example, Radiobot-CFF uses information state components to track whether it has received a target's location and what that target location is, as well as whether it has enough information to send a fire. To evaluate the performance of our dialogue manager, we studied how well it updated its information state components.

2.1 Approach

Our approach is to use human coders to decide how the information state components should be updated, given a sequence of utterances, and to compare that to how the system actually does update its information state components.

We develop a coding manual of guidelines for updating the information state components based on the kind of input received. We then use a sequence of trainee utterances (produced by hand-transcribing audio logs and hand-correcting system dialogue move interpretations of those utterances) to produce a sequence of hand-coded information state components. That sequence is our gold standard, and represents the output of the dialogue manager if the speech recognition, interpreter, and dialogue manager components are all performing to the level of a human.

We compare our system's performance to this gold standard corpus in two conditions. First, we run the dialogue manager on perfect input by feeding it the hand-corrected interpreter output, recording the information state components after every utterance, and comparing that to our gold standard. This allows us to evaluate the dialogue manager separately from the rest of the system, so that errors in the speech recognition and interpreter components do not affect its performance. Secondly, we compare the gold standard to the system's information state components when updated by the system on actual speech recognition and interpreter input. This allows us to evaluate the dialogue manager's performance given noisy input.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 181–182.

IS Component	*Accuracy, corrected input*	*Accuracy, noisy input*
has warning order	0.76	0.67
has target location	0.98	0.90
has grid location ‡	0.99	0.96
has polar direction	0.83	0.80
has polar distance	0.99	0.91
has target descript.	0.93	0.76
has enough to fire	0.99	0.52
method of control	0.71	0.71
method of fire †	0.38	0.44
grid value ‡	0.98	0.96
direction value	0.83	0.79
distance value	0.99	0.91
adjust fire	0.88	0.65
repeat FFE *	0.89	0.97
LR adjustment	0.99	0.92
AD adjustment	1.00	0.97
end of mission	0.93	0.91
disposition	0.93	0.78
number of casualties	0.95	0.83
mission is polar	0.99	0.85
last method of fire †	0.90	0.61
missions active	0.81	0.67

† Kappa was less than 0.8 and greater than 0.67

‡ Kappa was less than 0.67

* Kappa could not be calculated, as its value never changed in the data over which kappa was measured.

Table 1: Accuracy per IS Component

2.2 Results

We worked with a corpus of 17 sessions consisting of 407 utterances, representing a total of 8954 information state components to be updated. A pair of human coders coded several sessions by consensus to develop a set of guidelines, then individually coded the rest of the corpus. Several sessions were held out for concurrent coding by both coders, from which a kappa score was calculated per information state component. Components had kappa values above 0.8 except as noted in Table 1.

We then fed the corrected utterance interpretations into the dialogue manager to get sequences of IS component updates for corrected interpretations, and processed log files from the full system evaluation to get sequences of IS component updates for noisy interpretations. Accuracy results (measured by number of times the dialogue manager agreed with the human coder) for both are shown in Table 1.

3 Future Work

Because the input used in the corrected input condition is not reacting to the dialogue manager's responses, the dialogue may take an unnatural direction; for example, in which the dialogue manager is repeatedly prompting or correcting the trainee, but the trainee is proceeding as if there is no problem.

Also, a component's value may be more important at certain parts of a dialogue than at others. For example, as shown in Table 1, the "method of fire" component's accuracy is low, but the dialogue manager and humans disagree on its value most often at a phase of the dialogue in which the "method of fire" value is never used in decisions or output.

We hope to quantify and address these problems in future work.

Acknowledgments

This work has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Susan Robinson, Antonio Roque, Ashish Vaswani, Charles Hernandez, Bill Millspaugh, and David Traum, "Evaluation of a Spoken Dialogue System for Virtual Reality Call For Fire Training," Submitted, 2006.

Antonio Roque, Anton Leuski, Vivek Rangarajan, Susan Robinson, Ashish Vaswani, Shri Narayanan, David Traum, "Radiobot-CFF: A Spoken Dialogue System for Military Training," 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), Pittsburgh, PA, September 17-21, 2006.

Antonio Roque and David Traum, "An Information State-Based Dialogue Manager for Call for Fire Dialogues," 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia, July 15-16, 2006.

David Traum and Staffan Larsson, 2003. The Information State Approach to Dialogue Management. In R. Smith & J. van Kuppevelt (eds.) Current and New Directions in Discourse and Dialogue. Dordrecht: Kluwer, 325-353.

Dialogue management for cooperative, symmetrical human-robot interaction

Mary Ellen Foster
Robotics and Embedded Systems Group
Department of Informatics, Technical University of Munich
Boltzmannstraße 3, 85748 Garching, Germany
foster@in.tum.de

Abstract

We describe the JAST human-robot dialogue system, which supports fully symmetrical collaboration between a human and a robot on a joint construction task. We concentrate on the dialogue manager, which is based on Blaylock and Allen's (2005) collaborative problem-solving model of dialogue and which supports joint action between the dialogue participants at both the planning and the execution levels.

1 Human-robot dialogue in JAST

The overall goal of the JAST project ("**J**oint **A**ction **S**cience and **T**echnology"; http://www.euprojects-jast.net/) is to investigate the cognitive and communicative aspects of jointly-acting agents, both human and artificial. The JAST human-robot dialogue system (Foster et al., 2006) is designed as a platform for integrating the project's empirical findings on cognition and dialogue with its work on autonomous robots, by supporting symmetrical human-robot collaboration on a joint construction task.

The robot (Figure 1) consists of a pair of mechanical arms, mounted to resemble human arms, and an animatronic talking head capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The system input channels are speech recognition, object recognition, and face tracking; the outputs include synthesised speech, facial expressions and rigid head motion, and robot actions. The human user and the robot work jointly to assemble a Baufix wooden construction toy (Figure 2), coordinating their actions through speech, gestures, and facial motions. Joint action may take several forms in the course of an interaction: for example, the robot may ask the user to provide assistance by holding one part of a larger assembly, or may delegate entire sub-tasks to be done independently. In the current version of the system, the robot is able to manipulate objects in the workspace (e.g., picking them up, putting them down, or giving them to the user) and to perform simple assembly tasks.

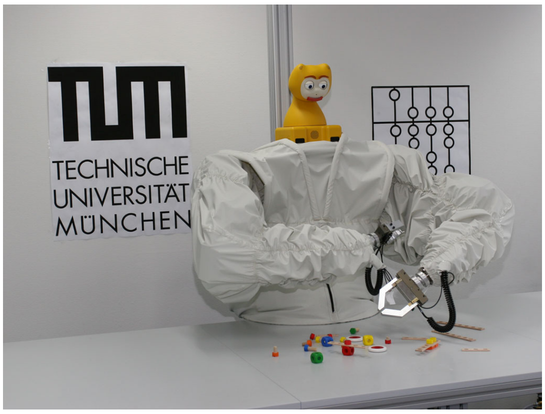


Figure 1: The JAST human-robot dialogue system

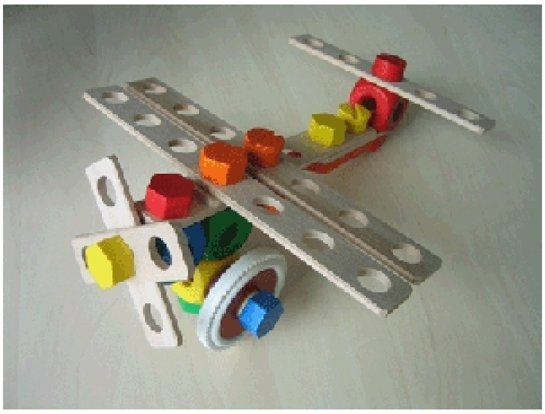


Figure 2: Assembled Baufix airplane

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 183–184.

2 Dialogue management in JAST

The JAST human-robot dialogue system has several features that distinguish it from many existing dialogue systems. First, the roles of the user and the robot are, in principle, completely symmetrical at all levels: either agent may propose a goal or a strategy for addressing one, and either—or both—may perform any of the actions necessary to achieve it. Also, the interaction must deal with both the selection of the actions to take in the execution of those actions, and may switch between the two tasks at any point. Finally, *joint action* is central to the dialogue at all levels: the participants work together to create domain plans, and also jointly execute the selected plans.

The distinctive requirements of the JAST dialogue system are most similar to those addressed by Blaylock and Allen (2005) in their collaborative problem-solving (CPS) model of dialogue. In collaborative problem solving, multiple agents jointly select and pursue goals, in three interleaved phases: selecting the goals to address, choosing procedures for achieving the goals, and executing the selected procedures. The central process in the CPS model is the selection of values (or sets of values) to fill roles, such as the goal to pursue or the allocation of sub-tasks among the participants. Slot-filler negotiations of this sort make up a large part of collaborative communication.

Dialogue management in the JAST system is based on this CPS model. As in COLLAGEN (Rich et al., 2001), the JAST dialogue state consists of three parts: the active set of goals and procedures, a set of open issues, and the interaction history. An *open issue* corresponds to any request, proposal or action that has occurred during the course of the dialogue and that has not yet been fully addressed; these are essentially the same objects as Ginzburg's (1996) *questions under discussion* (QUD). As an interaction proceeds, two parallel processes are active: the participants must complete domain goals such as locating and assembling objects, and must also address open issues that arise during the conversation. These two processes are tightly linked; for example, if an agent proposes a procedure for a particular subgoal and the other agrees (and closes the open issue), the next step in the interaction is likely to be executing the agreed-upon sequence of actions. Similarly, when an sub-goal is completed, the participants must address the open issue of how to proceed. The dialogue manager therefore maintains explicit links between the open issues and the current state of the domain plan to enable information to flow in both directions.

3 Current status and future work

At the moment, an initial dialogue-manager prototype based on the CPS model has been implemented in Java. This prototype supports a limited range of simple interactions with a cooperative user, using template expansion to create the domain plans. We are currently developing a more full-featured interaction manager, using a hierarchical planner to create the action sequences. As the system develops, we aim to expand its coverage to support phenomena such as failed actions and incorrect beliefs about the world, and to increase its robustness on incomplete or ill-formed messages from the input-processing modules.

Once a full working dialogue system has been developed, we intend to use it to implement and test the findings from the human-human joint-action dialogues that are currently being recorded and analysed by other participants in the JAST project; for example, we hope to derive strategies for confirmation, grounding, role assignment, and error handling. We will then perform a range of user studies to compare the success of the different strategies, as well as to measure the impact of factors such as feedback from the talking head, using both objective task-success measures and subjective measures of satisfaction and engagement.

Acknowledgements

This work is supported by the EU FP6 IST Cognitive Systems Integrated Project "JAST" (FP6-003747-IP). Thanks to Candy Sidner, Ron Petrick, Oliver Lemon, Geert-Jan Kruijff, Colin Matheson, Jon Oberlander, and the BranDial 2006 reviewers for useful comments and advice.

References

N. Blaylock and J. Allen. 2005. A collaborative problem-solving model of dialogue. In L. Dybkjær and W. Minker, editors, *Proceedings, 6th SIGdial Workshop on Discourse and Dialogue*, pages 200–211.

M. E. Foster, M. Rickert, and A. Knoll. 2006. Human-robot dialogue for joint-action construction tasks. In *Proceedings, 8th International Conference on Multimodal Interfaces (ICMI 2006)*. To appear.

J. Ginzburg. 1996. Dynamics and the semantics of dialogue. In J. Seligman and D. Westerstahl, editors, *Language, Logic and Computation, Volume 1*, CSLI Lecture Notes.

C. Rich, C. L. Sidner, and N. Lesh. 2001. COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, 22(4):15–25.

Hierarchical Reinforcement Learning of Dialogue Policies in a development environment for dialogue systems: REALL-DUDE

Oliver Lemon and Xingkun Liu
School of Informatics
Edinburgh University
`olemon,xliu4@inf.ed.ac.uk`

Daniel Shapiro and Carl Tollander
CSLI/Applied Reactivity
Stanford University
`dgs,carl@appliedreactivity.com`

Abstract

We demonstrate the REALL-DUDE system[1], which is a combination of REALL, an environment for Hierarchical Reinforcement Learning, and DUDE, a development environment for "Information State Update" dialogue systems (Lemon and Liu, 2006) which allows non-expert developers to produce complete spoken dialogue systems based only on a Business Process Model (BPM) and SQL database describing their application (e.g. banking, cinema booking, shopping, restaurant information, ...). The combined system allows rapid development and automatic optimization of spoken dialogue systems. Hierarchical Reinforcement Learning (RL) has not been applied to the problem of dialogue management before. It provides a way of dramatically reducing the size of the state space to be considered in RL problems. REALL-DUDE thus allows iterative development of dialogue policies through Hierarchical RL to be combined with a development environment for complete dialogue systems, encompassing parsing, speech recognition, synthesis, and dialogue management.

1 Introduction

It has been shown in previous work (Singh et al., 2002) that dialogue policies obtained by Reinforcement Learning (RL) can improve over hand-coded dialogue managers. However, a key problem in RL applied to dialogue management is the very large policy spaces generated by the dialogue management problem. REALL 's key source of power is its ability to constrain learning with background knowledge, within a principled framework. It has been shown (Shapiro and Langley, 2002) that this approach generates three order of magnitude reductions in problem size, and two order of magnitude improvements in learning rate, relative to the common formulation of RL tasks which offers all feasible options in all possible situations.

We demonstrate a development environment for dialogue systems which allows iterative development and refinement of dialogue policies through Hierarchical RL. We present the concepts behind REALL and DUDE , and show how to use DUDE to generate complete spoken dialogue systems (Lemon and Liu, 2006). We then demonstrate learning experiments that explore dialogue policies in the presence of different reward signals and channel noise characteristics, and show how the learner acquires different optimized policies.

2 REALL – Reactive Planning and Hierarchial RL

REALL is a language for defining extremely reactive agent behavior. It consists of a representation for expressing hierarchical, goal-oriented plans, together with an interpreter for evaluating those plans that operates in a repetitive loop. This iteration supplies reactivity: even if the world changes radically between two execution cycles, REALL will find a goal-relevant action to employ.

REALL is also a learning system. Because its interpreter contains a model-free reinforcement learning algorithm, every REALL agent has the ability to acquire an action policy from delayed reward. Programmers can access this capability

---

[1] This research is supported by Scottish Enterprise under the Edinburgh-Stanford Link programme.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 185–186.

by writing plans with disjunctive elements, and by embedding those choice points in hierarchical plans. As a result, REALL offers a means of invoking learning in the context of background knowledge, and this constrains the learning task.

Because REALL is a learning system, it supports a novel development metaphor called programming by reward. Here, the programmer may encode a dialogue strategy with options, and specify reward functions that serve as the targets of optimization. Via a training period, the reward functions select one of the many policies implicitly contained in the REALL plan, and developers can obtain distinct behaviors by making small changes to the reward functions (Shapiro et al., 2001).

REALL learns a policy by finding the best action to take in every state. It learns the value of a given state-action pair by sampling its future trajectory, and it represents this value using a linear function of currently observable features. REALL bootstraps: it updates the estimate for a state-action pair using its current value, the current reward, and the estimate associated with the next state-action pair. Over time, these estimates converge to their appropriate values.

3 The DUDE development environment

The contribution of DUDE (Lemon and Liu, 2006) is to allow non-expert developers to build ISU dialogue systems using only the Business Process Models (BPMs) and databases that they are already familiar with, as shown in figure 1.

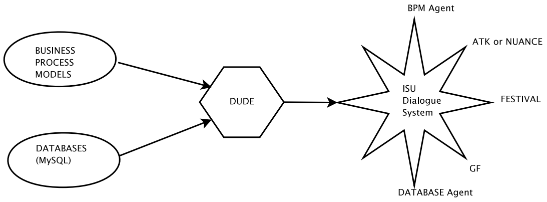


Figure 1: The DUDE development process

The environment includes a development GUI, automatic generation of Grammatical Framework (GF) grammars for robust interpretation of spontaneous speech, and uses application databases to generate lexical entries and grammar rules. The GF grammar is automatically compiled to an ATK or Nuance language model for speech recognition. See (Lemon and Liu, 2006) for details.

The power of REALL-DUDE is to embed Hierarchical Reinforcement policy learning and optimization from REALL within the rich development environment supplied by DUDE .

4 Demonstrating learning

We will present a REALL program, `Slotfiller`, embedded in the DUDE environment, which contains a scaffolding of required dialogue behavior (e.g., confirmations, clarifications, mixed-initiative questions). The demonstration presents a variety of learning experiments that explore these decisions in the presence of different reward signals and channel noise characteristics. We will show how the learner acquires and optimizes distinct dialogue policies in each case.

5 Conclusion

Hierarchical RL has not been applied to the problem of dialogue management before. It provides a principled way of dramatically reducing the size of the state space to be considered in RL of dialogue management. Here we demonstrate a development environment, REALL-DUDE , which combines RL for optimization of dialogue policies with a full development environment for automatic generation of spoken dialogue systems. We will demonstrate how to develop complete spoken dialogue systems using DUDE and then we will demonstrate strategy learning for those systems using REALL, which optimizes policies for different noise and reward conditions in dialogue.

References

Oliver Lemon and Xingkun Liu. 2006. DUDE: a Dialogue and Understanding Development Environment, mapping Business Process Models to Information State Update dialogue systems. In *Proceedings of EACL (demonstration systems)*.

D. Shapiro and P. Langley. 2002. Separating skills from preference: using learning to program by reward. In *Nineteenth International Conference on Machine Learning*.

Dan Shapiro, Pat Langley, and Ross Shachter. 2001. Using background knowledge to speed reinforcement learning. In *Fifth International Conference on Autonomous Agents*.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research (JAIR)*.

Dico: A Multimodal Menu-based In-vehicle Dialogue System

Jessica Villing
Department of Linguistics
Göteborg University
Sweden
jessica@ling.gu.se

Staffan Larsson
Department of Linguistics
Göteborg University
Sweden
sl@ling.gu.se

Abstract

We present a demo of a set of dialogue system applications for controlling various devices in a truck. All applications use the GoDiS dialogue manager and implement multimodal menu-based dialogue (MMD) based on the menu structures of existing GUI interfaces.

1 Introduction

Dico is a multimodal in-car dialogue system application[1]. An obvious advantage of spoken dialogue in the vehicle environment is thath the driver does not have to take the eyes - and the attention - off the road.

In the original Dico application (Olsson and Villing, 2005), the dialogue system was able to control a cellphone. The main goal was to develop an interface that is less distracting the the driver, and thus both safer and easier to use than existing interfaces. In an in-vehicle environment, it is cruicial that the system is intuitive and easy to use. GoDiS' dialogue manager allows the user to interact more flexibly and naturally with menu-based interfaces to devices.

Today's vehicles typically contain several devices that the driver needs to control, many of them with menu-based interfaces. To show how a multimodal dialogue system can help when controlling several devices, Dico has been extended with a DID (Driver Information Display) and a radio.

2 GoDiS and TrindiKit

Dico is developed using the dialogue system GoDiS (Larsson, 2002). GoDiS is implemented using TrindiKit (Traum and Larsson, 2003). General dialogue management issues such as feedback, grounding, question accommodation and task switching are handled by the application-independent dialogue manager. Re-using these technologies in new applications enables rapid prototyping of advanced dialogue applications. GoDiS has been adapted to several different dialogue types, domains, and languages, including menu-based multimodal dialogue when acting as an interface to an mp3 player (Hjelm et al., 2005).

3 The Dico application

When the user starts Dico she is asked to choose one of the three applications. Later in the interaction she is allowed to switch to any application without having to return to the top application (Dico) first. The output in the radio and the DID applications is multimodal. The modalities, visual and acoustic, work in parallel in the sense that the same information is given on the screen and verbally.

The services offered by the cellphone application includes calling a name or a number, managing the phonebook, using the helper and changing the settings in the cellphone application. The radio station and the volume can be changed. It is also possible to ask for the current radio station and for available radio stations. In the DID application the truck driver can get information about the orders (such as where to pick-up and deliver goods) and information from the tachograph (driving, resting, working and waiting time).

---

[1] The work reported here was funded by TALK (Talk And Look, Tools for Ambient Linguistic Knowledge), EC project IST-507802, and DICO, Vinnova project P28536-1

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 187–188.

4 GoDiS features in Dico

To enable a more natural and intuitive dialogue, GoDiS supports (among other things) accommodation, task switching and grounding.

4.1 Accommodation

Since the applications are based on existing menu interfaces, it is possible for the novice user to let the system take the initiative and guide the user through e.g. the phonebook menu.

For expert users, accommodation enables skipping through the menus and getting right to the point. Instead of navigating through a complicated and tiresome menu structure it is possible for the expert user to give all information in one utterance:

Ex. 1: "Call Lisa's home number"

4.2 Multiple simultaneous tasks and task switching

It is possible to start one dialogue to perform a task, and then start a subdialogue before the first task is completed. When the second task is completed the system automatically returns to the first task. This gives the user freedom to switch task at any time:

Ex. 2:

U: "Change Lisa's home number."
S: "Okay. Let's see. What phonenumber do you want instead?"
U: "Check my messages."
S: "You have got this message. . Hi! I have got a new home number, it is (031)234567. Best regards, Lisa."
S: "Returning to change an entry. What phone number do you want instead?"
U: "oh three one two three four five six seven."
U: "Okay. Changing Lisa's home number to oh three one two three four five six seven."

4.3 Feedback and grounding

The GoDiS dialogue manager provides feedback to make sure that the dialogue partners have contact, that the system can can hear what the user says, understands the words that are spoken (semantical understanding), understands the meaning of the utterance (pragmatical understanding) and accepts the dialogue moves performed in utterances.

Combining feedback on different levels makes it possible for the system to give explicit feedback on the user's input. E.g. the single user utterance "Lisa" gives positive grounding on the semantic level but negative on the pragmatic, resulting in a system utterance consisting of two feedbac moves and a clarification question: "Lisa. I don't quite understand. Do you want to add an entry to the phonebook, call a person, change an entry in the phonebook, delete an entry from the phonebook or search for a name?".

5 Future work

We plan to extend Dico to handle integrated multimodality on the input side. By "integrated multimodality" we mean that different modalities contribute with different parts of the dialogue, i.e. *"Play this [click]"* where the [click] is a mouse click at the selected song. Technolgies for integrated multimodality in menu-based applications have already been developed for other GoDiS applications (Hjelm et al., 2005) and these solutions will be re-used in Dico.

Ko (2006) found that a context-aware dialogue system can reduce the degree of user distraction while driving. Techniques for adding awareness of the driver's level of distraction to Dico, and adapting the dialogue accordingly (e.g. by pausing the dialogue when the driver's distraction level is high and resuming it later), will be investigated in the ongoing DICO project in cooperation withVolvo, TeliaSonera and KTH.

References

Jeongwoo Ko, Fumihiko Murase, Teruko Mitamura, Eric Nyberg, Masahiko Tateishi and Ichiro Akahori (2006). *Analyzing the Effects of Spoken Dialog Systems on Driving Behavior.* LREC 2006 Conference.

David Hjelm et. al. (2005). DJ GoDiS: Multimodal Menu-based Dialogue in an Asychronous ISU System in Gardent and Gaiffe (eds.) *Proceedings of the ninth workshop on the semantics and pragmatics of dialogue.*

Staffan Larsson (2002). *Issue-Based Dialogue Management*, PhD thesis, Department of Linguistics, Goteborg University.

Anna Olsson and Jessica Villing 2005 *Dico - a Dialogue System for a Cell Phone*, Master thesis, Department of Linguistics, Goteborg University.

David Traum and Staffan Larsson 2003 *The Information State Approach to Dialogue Management.* In Smith and Kuppevelt (eds.): Current and New Directions in Discourse I& Dialogue, Kluwer Academic Publishers, 325–353.

Accommodating Social Relationships

Joris Hulstijn
Vrije Universiteit, Amsterdam
`jhulstijn@feweb.vu.nl`

Abstract

Many social aspects of a dialogue context are determined by the way an addressee 'takes up' an utterance of the speaker. We show that inferences about the dialogue context based on uptake, are essentially a form of *presupposition accommodation*. The account is illustrated by an analysis of commands, advice and threats.

1 Introduction

When describing the semantics and pragmatics of dialogue, social aspects of the dialogue context are crucial. Social roles and relationships are often part of the felicity conditions of a speech act (Austin, 1962). For example, a command is only warranted when the speaker has some power or authority over the addressee. Advice requires that the speaker has expertise. For threats to be convincing, the threatener must appear to be willing and able to actually carry out the threat, which must be feared by the addressee. Thus, by making an utterance of a certain type, requirements are put on the social relationships between the dialogue participants. Witness the following examples.

(1) S: Clean up the floor!
A.1: Yes, sir.
A.2: Do it yourself.

(2) S: You should do the literature review, before collecting your data.
A.1: All right.
A.2: Why?

(3) A: If you haven't paid by Tuesday, I'm gonna break your balls.
S.1: Yes.
S.2: We'll see.

When the social relations are partially unknown, the response of the addressee further determines the common ground in this respect. For example, by accepting the command in (1), addressee A helps to establish a power relation; and by rejecting the advice in (2), A challenges S's expertise, undermining the supervisor-student relationship. So the way in which an addressee 'takes up' the utterance of the speaker helps to determine the dialogue context. This process may be called *uptake* (Austin, 1962; Hulstijn and Maudet, 2006).

This paper explores the idea that the inferences that can be made as a result of uptake, are the result of *presupposition accommodation* (Lewis, 1979). The speaker presupposes that the felicity conditions of the speech act are part of the common ground (Austin, 1962, p.50,51). If they are not, the addressee will adapt his or her version of the common ground to accommodate the felicity conditions, provided that there is no information to the contrary. Otherwise, the dialogue becomes awkward, typically indicated by the addressee.

2 Roles in Dialogue

Dialogue participants are executing some social activity, the conventional rules of which may be expressed as a *dialogue game*. Crucial are the roles of the participants. Roles *prescribe* obligations and permissions, but like stereotypes, roles also *describe* expectations, allowing others to predict behavour. We distinguish three kinds of roles (Hulstijn, 2003): (i) *Turn taking roles*, such as speaker, addressee or (over)hearer, alternate repeatedly. (ii) *Participant roles*, such as expert and novice in an information exchange, remain stable during a dialogue. (iii) *Social roles*, like teacher and pupil, extend beyond single dialogues. Their scope depends on the social setting.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 189–190.

We specify the social aspects of the felicity conditions of our examples, with notation s: speaker, a: addressee, and pres(.): presuppositions.

(4) $\mathrm{pres}(\mathrm{command}(s, a, \varphi)) =$
$\mathrm{authority}(s, a) \wedge$
$\mathrm{capable}(a, \varphi)$

(5) $\mathrm{pres}(\mathrm{advice}(s, a, \varphi)) =$
$\mathrm{expertise}(s, \varphi)$

(6) $\mathrm{pres}(\mathrm{threat}(s, a, \varphi, \psi)) =$
$\mathrm{fear}(a, s) \wedge$
$\mathrm{goal}(s, \neg\varphi \rightarrow \psi) \wedge$
$\mathrm{capable}(s, \psi)$

3 Presuppositions and Common Ground

A *presupposition* is whatever the speaker takes to be part of the common ground, when making an utterance (Stalnaker, 1974). The hypothesis is that this can be generalised to addressees, who reveal their version of the common ground by responding in a certain way.

Suppose we represent belief with a standard KD45 operator $\mathrm{B}_i\varphi$. We use $\mathrm{C}\varphi$ to represent that φ is common belief in group $\{1...n\}$. Let $\mathrm{E}^1(\varphi) = \mathrm{B}_1\varphi \wedge ... \wedge \mathrm{B}_n\varphi$, and define recursively $\mathrm{E}^{m+1}(\varphi) = \mathrm{E}^1(\mathrm{E}^m(\varphi))$. Define $\mathrm{C}\varphi = \mathrm{E}^k\varphi$, for all $k \geq 1$ (Fagin et al., 1995). Now a participant's presupposition can be represented by $\mathrm{B}_i\mathrm{C}\varphi$.

However, Stalnaker (2002) proposes to use the notion of *acceptance*. To accept a proposition, is to treat it publicly as true. Let $\mathrm{A}_i\varphi$ mean that i accepts φ. Acceptance has the same logic as belief, namely KD45. The common ground is understood as common belief about what is accepted: $\mathrm{CG}\varphi \equiv \mathrm{C}(\mathrm{A}_1\varphi \wedge ... \wedge \mathrm{A}_n\varphi)$. Presuppositions then are beliefs about the common ground: $\mathrm{B}_i\mathrm{CG}\varphi$.

To become common ground, acceptance must be indicated by explicit or implicit acknowledgements. This process is called *grounding* (Clark and Schaefer, 1989). The recent dialogue history can serve as a kind of *basis* for establishing the common ground (Lewis, 1969).

Given an utterance with a presupposition, how should the addressee respond? There are three possibilities, inspired by the satisfaction theory of presupposition (Beaver, 1996): (i) the presupposition is already part of the (addressee's version of the) common ground, (ii) the presupposition is not yet part of the common ground, but can be added consistently, or (iii) the common ground contains information that contradicts the presupposition.

4 Conclusion

Presupposition accommodation provides a dialogue game rule, similar to the Gricean maxims. In case (i) and (ii) the addressee should acknowledge acceptance of the utterance. In case (iii) the addressee should explicitly reject the utterance.

(7) Suppose $\mathrm{pres}(\alpha(s, a, \varphi)) = \psi$.
If $\neg\mathrm{B}_a\mathrm{CG}\neg\psi$, then $\mathrm{ack}(a, s, \alpha(s, a, \varphi))$,
else $\mathrm{reject}(a, s, \alpha(s, a, \varphi))$.

Here, α stands for any speech act, and 'ack' for an acknowledgement. Because of KD45, case (i) $\mathrm{B}_a\mathrm{CG}\psi$, is included under (ii) $\neg\mathrm{B}_a\mathrm{CG}\neg\psi$.

Suppose a hears utterance $\mathrm{command}(s, a, \varphi)$, $\mathrm{advice}(s, a, \varphi)$ or $\mathrm{threat}(s, a, \varphi)$, as in (1) - (3). Based on the response, A.1 or A.2, we can derive that the presupposed social relationship is either accommodated to the common ground, or denied.
– Response A.1 indicates acceptance. Because this response forms a sufficient basis to establish a common ground, we can derive $\mathrm{CA}_a\psi$. Similarly, we can derive $\mathrm{CA}_s\psi$ from the speaker presupposition. By distribution we get $\mathrm{C}(\mathrm{A}_a\psi \wedge \mathrm{A}_s\psi)$ and by definition $\mathrm{CG}\psi$.
– Response A.2 indicates rejection. Again we derive $\mathrm{C}\neg\mathrm{A}_a\psi$, so $\neg\mathrm{CA}_a\psi$ by and therefore $\neg\mathrm{CG}\psi$.

This shows that presupposed social relations are indeed accommodated or rejected, and that we can reason about this on the basis of the 'uptake' by the addressee. Exploring alternative ways of dealing with presupposition and grounding, remains a topic for further research.

References

Austin, J. (1962). *How to do things with words*. Harvard University Press, Cambridge Mass.

Beaver, D. (1996). Presupposition. In *Handbook of Logic and Language*, 939–1008. Elsevier, Amsterdam.

Clark, H. H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. MIT Press, Cambridge,MA.

Hulstijn, J. (2003). Roles in dialogue. In *Proceedings Diabrück'03*, 43 – 50. Universität des Saarlandes.

Hulstijn, J. and Maudet, N. (2006). Uptake and joint action. *Cognitive Systems Research*, 7(2-3):175–191.

Lewis, D. (1969). *Convention: A Philosophical Study.* Harvard University Press, Cambridge.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 9:339–359.

Stalnaker, R. C. (1974). Pragmatic presupposition. In *Semantics and Philosophy*, 197–213. New York Univ. Press.

Stalnaker, R. C. (2002). Common ground. *Linguistics and Philosophy*, 25:701-7219.

The Effect of Multiple Modalities in Dialogue Act Annotation

Crystal Nakatsu and **Chris Brew**
Department of Linguistics
The Ohio State University
Columbus, OH 43210 USA
`{cnakatsu,cbrew}@ling.ohio-state.edu`

1 Introduction

In previous work, SWBD-DAMSL (Jurafsky et al., 1997) showed that interrater reliability could be improved by decreasing the set of possible tag combinations. However, this solution may not be an option for researchers wishing to study dialogue act phenomena at a lower level of detail. So how can we continue to improve interrater reliability scores without modifying the annotation scheme?

As an alternative to modifying the tagset, one might instead alter the process of annotation. Very few corpus developers mention whether they allow for audio playback of an utterance during the coding process. In this work, we investigate the effects of dual modality annotation on both annotation rate and interrater reliability.

2 Annotation Experiment

2.1 Annotation Procedure

Two native speakers of English independently labeled the Trains 93 corpus (Heeman and Allen, 1994), using a (re-)modified version of Doran et. al.'s (2001) modified C-Star dialogue act tagset.

The first 43 dialogues (2961 utterances) were annotated through reading the transcripts (i.e. text) only, while the remaining 52 dialogues (3875 utterances) were annotated by listening to the corresponding audio file while viewing the transcripts.

2.2 Annotation Rate

The text-only utterances were annotated at an average rate of 0.121 utt/s and the text-audio utterances at at rate of 0.157 utt/s. Initially, these rates imply that it is the use of audio that increases the annotation rate. However, since the utterances in the text-only condition are annotated before the utterances in text-audio condition, the increased rate could be attributed to increased familiarity with the tag set.

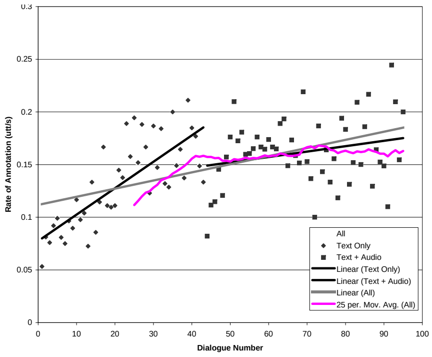


Figure 1: Rate of Annotation (seconds/utterance) with linear interpolation.

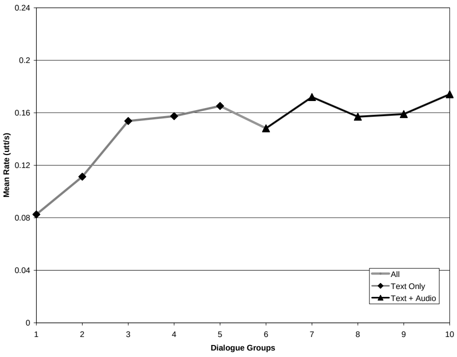


Figure 2: The Group Mean Annotation Rate (seconds/utterance)

The correlation analysis supports the influence of the familiarity effect, showing a significantly strong positive correlation ($r = .77, p < .001$) between dialogue number and annotation rate in the text-only condition, and a non-significant weak correlation ($r = .24, p < .1$) between the same two variables in the text-audio condition. Further analysis by a 2-factor ANOVA (F = 6.6, df = 8, $p <$ 1x10^{-6}), using dialogue number and modality as independent factors (depicted in Figure 2) more clearly indicates that the rising rate occurs mostly in the first two groups of the text-only dialogues

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 191–192.

and then flattens out in last 3 groups[1]. Furthermore, the rate from Groups 4 & 5 are maintained in groups 6-10 (with minor variance). This finding suggests that the addition of audio is not a factor in the increased annotation rate, but rather that annotation rate increases sharply at the onset of the annotation process as a result of some other factor that changes over time, such as an increase in familiarity of the tagset, and then flattens out, likely due to the annotators reaching maximum familiarity with the tagset.

Also, although the annotation rate is flattened in the later dialogues (Groups 4-10), it is maintained throughout the text-audio condition at about the same rate as the latter text-only dialogues. Thus, while annotation rate is not positively affected by the use of additional media, neither is it negatively affected.

2.3 Interrater Reliability

Raw agreement for all the utterances in the text-only modality is 66.7%, with $\kappa = 0.623$. This is slightly lower than the $\kappa = 0.66$ reported in (Doran et al., 2001) using their modified C-star tagset, but higher than the averaged $\kappa = 0.54$ achieved by the Trains 93 corpus annotators using the DAMSL scheme (Allen and Core, 1997) which allowed use of audio during annotation. In comparison, adding audio during our annotation resulted in an even higher raw agreement of 74.5% and $\kappa = 0.701$.

Again, at first glance these scores indicate that the increase in reliability is due to the use of the utterances' audio recordings during annotation. However, as before, due to the order of annotation, the increase in reliability could be due to increasing familiarity with the tagset.

The significant negative correlation (r = -0.45, $p <$ .005) in the text-only condition (in Figure 3) would seem to strongly indicate that κ did not improve as a result of familiarity, since we would expect a positive correlation in that case. This is further supported by a 2-factor ANOVA (F = 3.0, df = 8, $p <$ .005), which shows that κ decreases over time in the text-only condition, but is mostly level in the text-audio condition (Figure 4).

Having ruled out familiarity as a possibility for the improvement in interrater reliability, it seems that the improvement can indeed be correlated with the use of the corresponding audio recording of the utterance during annotation. In addition, this improvement comes at no obvious detriment to the annotation rate, since the annotation rate does not decline but rather remains somewhat steady throughout the text-audio condition.

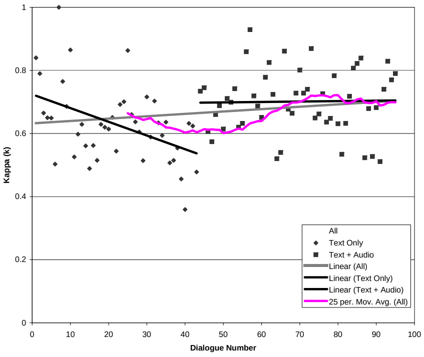


Figure 3: Interrater Reliability (κ) with linear interpolation and moving average.

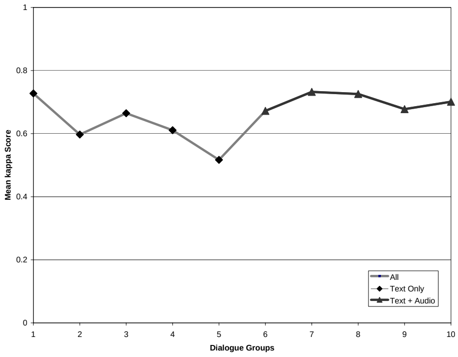


Figure 4: The Group Mean Interrater Reliability (kappa) Score

3 Acknowledgements

Thank you to Donna Byron, Michael White, and the Clippers Discussion group at OSU.

References

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers. Available at http://www.cs.rochester.edu/research/trains/annotation.

C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, pages 48–57.

Peter A. Heeman and James Allen. 1994. The TRAINS93 dialogues. Technical Report TRAINS TN 94-2, University of Rochester.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science, August.

---

[1] Groups 1-5 (text-only condition) were annotated first, and Groups 6-10 (text-audio condition) were annotated last.

Building a Multi-Lingual Interactive Question-Answering System for the Library Domain

Manuel Kirschner
KRDB, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy
`kirschner@inf.unibz.it`

1 Introduction

Lately, there has been increasing interest in how to best enrich Question-Answering (QA) applications with dialogue capabilities[1]. While classical QA is concerned with questions posed in isolation, its interactive variant keeps track of the QA process and supports the user in finding the exact solution via natural-language dialogue. The context of each utterance must be considered for handling clarification sub-dialogues and to resolve anaphora, ellipses or fragmentary utterances.

We have started to develop an Interactive QA system for the university library domain. We see it as a fusion of the QA scenario with robust dialogue systems techniques. From the QA point of view, the system's task is to retrieve answers to user questions using a knowledge base, where the user holds the initiative throughout the exchange. Research in discourse structure and dialogue management can provide us with models for properly dealing with co-reference, keeping track of the current topic, and even temporally switching to system initiative when the user is lost. We adopt a bottom-up approach, starting with data collection, and building a simple baseline system. As more data become available, we plan to iteratively add support for more sophisticated discourse phenomena.

2 Library domain

The University library is striving to extend services that help their users find information over the web site in their preferred language. To this end, an interactive QA system will provide permanent and instant multi-lingual access to library-specific information. As the experiences of other libraries have shown, these systems surpass FAQ lists by guiding the user towards answers when initially she does not know how to formulate (and search for) the exact question.

In cooperation with library staff, we drew up a specification of the scope of our conversational agent, i.e., the topics for possible human-computer dialogues. Our librarians analyzed archived e-mail with past user queries and compiled a typology of frequent questions. Existing FAQ lists from the library web-site were also incorporated. The resulting library domain is structured into topics and sub-topics as follows[2]:

General: library buildings, organization, services (e.g., opening hours, library card, technical infrastructure, library web-site), FAQ
Inquiry: catalog query, general, books, journals, articles/theses, literature topics
Lending: lending, inter-library loan

3 Wizard-of-Oz study

Under the bottom-up paradigm described in the introduction, we first need conversation data, gained through conducting a Wizard-of-Oz (WoZ) study with actual library users. The study has two goals:

1. Verify topic scope of knowledge base
2. Explore dialogue-specific discourse phenomena present in WoZ data

We designed a simple chat-like interface that was accessible to all users on the university intranet via a link from the main library web-site. The system establishes a text-based conversation between users and library staff, while the users are made believe that they are actually talking to an automatic dialogue system. The users were given only a rough specification of possible query topics as a guideline, but were not provided with a task description, as is usually the case in WoZ studies.

In a first run, the WoZ system was online during office hours for two weeks. We collected an overall of 64 dialogues, with an average of 6 turns per

---

[1] E.g., Workshop on Interactive Question Answering (IQA'06), in HLT-NAACL'06.

[2] As a side task, the librarians are annotating electronically available library documents with the above topic structure. The marked-up documents will prove useful for providing users with additional source evidence, and for later Information Retrieval experiments.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 193–194.

dialogue. We plan to conduct more WoZ runs in the near future.

Topic scope Analyzing the coverage of our knowledge base with regard to the collected WoZ dialogues, roughly two thirds are within the scope of library topics we wanted to cover (see box). The remaining logs exhibited out-of-scope user queries, about half being nonsense words, the other half typical "small talk" dialogue. Thus, if only to keep the conversation going and to allow for a playful exploration of the task domain, the dialogue system should exhibit some small talk skills as well.

Observed discourse phenomena Of the initially collected 64 dialogues, 12 contain some kind of discourse phenomenon. This ratio (18,75%) is on the lower end of the spectrum reported in the literature, where information-seeking tasks in particular have been shown to contain the lowest share of context-dependent turns (Bertomeu et al., 2006). Also, in that study, participants were given a complex task: asking for additional information about previously retrieved entities from a database. In our view, such differences in system and user tasks can explain the lower number of discourse phenomena in our data. We annotated these data using the classification scheme for fragments/referential devices proposed in Bertomeu et al. (2006). In our study, the three most frequent discourse phenomena were anaphoric pronouns[3] (11 times), elided NPs[4] (10) and fragments[5] (5).

4 Implementation status and next steps

Having implemented the WoZ application, our next goal is to build a baseline dialogue system. We have developed software tools that enable our librarians to construct a hierarchical knowledge base of library topics in XML (with main topics as listed in section 2). For each topic (and subtopic), this hierarchy contains keywords for mapping a user query to an appropriate canned-text answer or to a clarification request (if the query does not contain all the keywords required to traverse the hierarchy until reaching a terminal topic).

The baseline system will cover a sub-set of the discourse phenomena encountered in our current WoZ data. The underlying notion is that in our hierarchy, the topic that matches a user query represents the current conversational context. Follow-up questions as in footnotes 3 and 5 can lead to a correct answer if the identity of the previous topic is kept as the current dialogue state. A follow-up question is then matched against the keywords stored under that specific topic. For the example of footnote 3, the active topic just before the follow-up question is *inter-library loan*; then, provided that *cost* is defined as a sub-topic, the follow-up question (containing "cost") matches a local keyword, yielding the desired topic-related answer.

Because of the dual role of our topic hierarchy—knowledge base and data structure for keeping dialogue state—we see it as a simplified version of the *topic structure* described in Stede and Schlangen (2004). At the time of writing, library staff have started filling the knowledge base with topics, keywords and answers. By analyzing log-files gained from the baseline system (i.e., checking where user input failed to match (correct) patterns in the knowledge base), librarians will be able to iteratively improve the coverage by adding new keywords or new topics.

We plan to support the full regular expression syntax for defining keywords, and to let the librarians define procedural sub-dialogues to complement the baseline keyword search algorithm for the topic hierarchy[6]. Through the analysis of log files, we will study how users keep/change topic focus, and how mixed-initiative dialogue can help the user in finding the answer to a complex query.

References

Núria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 2006. Contextual phenomena and thematic relations in database QA dialogues: results from a Wizard-of-Oz Experiment. *Proc. of HLT-NAACL'06*, New York City, NY.

Manfred Stede and David Schlangen. 2004. Information-Seeking Chat: Dialogues Driven by Topic Structure. *Proc. of Catalog'04 (Semdial'04)*, Barcelona, Spain.

---

[3] User: "wie funktioniert Fernleihe? [how does inter-library loan work?]" – Wizard: "(...)" – User: "und wieviel kostet *es*? [and how much does *it* cost?]"

[4] U: "Nur *die Diplomarbeiten* der Informatik-Fakultät? Nicht auch ∅ von der Wirtschaftsfakultät? [Only *the masters theses* from the faculty of computer science? Not also ∅ from the faculty of Economics?]"

[5] W: "Zum Kopieren verwendet man die Student Card. [For making photocopies use your student card.]" – U: "zum Drucken? [for printing?]"

[6] These features were deemed important in "Stella" (`http://www.sub.uni-hamburg.de/`), a proprietary German dialogue system built for the Hamburg University library (Anne Christensen, pers. comm.).

Referring Expressions and the Local Coherence of Discourse in a Parallel Corpus of English and Japanese Map Task Dialogues

Etsuko Yoshida
Mie University/University of Edinburgh
1577 Kurimamachiya-cho, Tsu, Mie
514-8507 Japan
tantan@human.mie-u.ac.jp

Abstract

This paper examines how discourse entities are linked with the center (Cb) in the Centering Model in the construction of coherence in discourse, by looking at how the choice and distribution of referring expressions are correlated with the center transition patterns. Based on the analysis of parallel dialogue data in English and Japanese, the proportion of center transition patterns is investigated. More specifically, despite the grammatical differences in the form of references between the two languages, the ways of discourse development in both sets of data show distinctive similarities in the process in which topic entities are introduced, established, and shifted away to subsequent topic entities. There is little doubt that full NPs are the main conduit for the topic center, and continue to be used both within and across discourse segment boundaries, while the (zero) pronoun can carry the topic center only in limited stretches of discourse and is likely to be discontinued at the end of the discourse segment.

1. Outline

In this paper, I examine how discourse entities are linked with the center (Cb) in the Centering Model in the construction of coherence in discourse, by looking at how the choice and distribution of referring expressions are correlated with the center transition patterns. This may suggest that there is a universal feature in which noun phrases contribute to the center transition of Cb in developing the topic entities in dialogic discourse.

Centering is the model in which discourse coherence is predicted by the transitions of discourse entities and the local focus of attention within the discourse segment (Grosz et al. 1995, Walker et al. 1998). I first analyse the distribution of center (Cb) transition patterns and the types of referring expressions in each center transition pattern. Then I clarify the preferred combination of transition sequence patterns by investigating the relationship between the types of referring expressions and the way of constructing topic chains in particular transition sequence patterns.

2. Example

Let us look at the following extract in CONTINUE transition in English. The chain of NPs tends to be combined with the chain of pronouns and zero pronouns: *this grass* is immediately replaced with PRO *it* and ZERO and established as a Cb in CONTINUE transition until the new entity *a house* is introduced. This new entity can predict the change of Cb in RETAIN transition, but the current Cb still continues as a noun *the grass*.

(1) (English: Lleq4c2)
TA109: And then you're turning up towards <u>this grass</u> **CON**
TB110: And going round <u>it</u> **CON**
TA111: Yeah round the left hand side <u>[Ø]</u> and then over the top <u>[Ø]</u> **CON** **CON**
TB112: Right
TA113: And then stop when you get to the/
TB114: Edge of <u>it</u> **CON**
TA115: Edge of <u>it</u> **CON**
TB116: Okay

TA117: And then do you have a house with on it? **RET**

Just up from the grass on the right **CON**

TB118: No

In Japanese, it is possible to see that temporary chains of zero pronouns do occur, but the topic chains of NPs tend to link the entities in CONTINUE transition with these zero entities. Thus, the entities that are a current topic are consistent, but the types of expressions can be replaced depending on the different context of the situation in discourse.

3. Summary and Further Research

It has been clarified that despite the difference in grammatical forms of referring expressions between English and Japanese, there are similar patterns in the way the discourse entities are introduced, established as a focus of attention, and are shifted away to the new entities in discourse development. It can be clearly seen that the chain of NPs can contribute to discourse coherence in both English and Japanese, but their distribution shows that NPs have different roles in different types of Cb transition patterns. That is, NULL and SHIFT are the transitions in which the entities are introduced as a full NP and the expressions can be extended according to the understanding of the entities for the participants. In CONTINUE transitions the entities contribute to the topic chains of NPs, mainly as bare nouns in Japanese in the same way that pronouns contribute to topic chains in combining NPs in English data.

It is still difficult to evaluate the results as reliable, but the findings themselves are interesting. They suggest that the center of attention is maintained by the chain of NPs rather than (zero) pronouns, and that the chain of NPs is correlated with the local and global focus of discourse coherence. There is little doubt that full NPs are the main conduit for the topic center, and continue to be used both within and across discourse segment boundaries, while the (zero) pronoun can carry the center of topic only in limited stretches of discourse and is likely to be discontinued at the end of the discourse segment.

These observations are not fully explained by existing theories of anaphora resolution and it is difficult to predict typical patterns of referential transitions in naturally occurring discourse. Two alternative perspectives appear to be particularly promising. First, it is worth noting that such repeated NPs tend to function as proper names in the discourse; second, the speaker's short-term memory is repeatedly activated by the combination of these expressions. Both phenomena call out for further research.

References

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. (1995) Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21/2, 203-225.

Gundel, J. K., N. Hedberg and R. Zacharski (1993) "Cognitive status and the form of referring expressions in discourse." *Language*, 69, 2: 274-307.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. (2004) Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30/3, 309-363.

Prince, E. (1981) 'Toward a taxonomy of given-new information' In P. Cole (ed.), *Radical Pragmatics*. New York: Academic Press. 223-56.

Walker M. A., A. K. Joshi and E. Prince (eds.) (1998) *Centering Theory in Discourse*. Oxford: Clarendon Press.

Walker, M. A. (1998) "Centering, Anaphora Resolution, and Discourse Structure." In Walker M. A., A. K. Joshi and E. Prince (eds.).401-436.

Walker, M.A. (2000) 'Toward a Model of the Interaction of Centering with Global Discourse Structure' Verbum.

Yamura-Takei, Mitsuko. (2005) *Theoretical, Technological and Pedagogical Approaches to Zero-Arguments in Japanese Discourse: Making the Invisible Visible.* Doctoral thesis, Hiroshima City University.

How speakers represent size information in referential communication for knowing and unknowing recipients

Judith Holler & Rachel Stevens
School of Psychological Sciences
University of Manchester
Manchester M13 9PL, U.K.
`judith.holler@manchester.ac.uk`

1 Introduction

Past research has shown that in order to communicate efficiently and effectively participants in conversation take into account their ‘common ground’ (e.g. Clark & Schäfer, 1989; Clark & Wilkes-Gibbs, 1986; Fussell & Krauss, 1989; Isaacs & Clark, 1987). Speakers tailor their utterances with respect to what their interlocutors already know about the respective topic of conversation and what they don’t know. This concerns both common ground that exists from the outset, as well as the common ground that is accumulated over the course of a conversation. Most of the research focussing on this topic has considered exclusively verbal communication.

However, one major element of human communication is constituted by speakers’ imagistic hand movements that accompany much of everyday talk. The gestural system bears a very close connection to the speech system and is thus considered as representing a core part of language; speakers’ utterances are seen as comprising both a linguistic as well as an imagistic side (McNeill, 1992). When researching certain aspects of human communication it appears therefore essential that we apply a wider perspective which takes into account the gestural component of utterances.

To date, within the field of gesture, the connection between social processes and gesture use has been investigated to some extent. For example, some studies have explored the effect of gestures being visible to an interlocutor, or limited verbal interaction with an interlocutor, on the occurrence of gestures (Alibali, Heath & Meyers, 2001; Bavelas, Kenwood, Johnson & Phillips, 2002, Beattie & Aboudan, 1994). One study has focused on whether speakers’ gesturing is influenced by them anticipating their listeners’ potential understanding problems (Holler & Beattie, 2003). However, too little is still known about the communicational functions gesture fulfils in conjunction with speech in the context of conversational interaction.

One interesting starting point in this respect is the investigation of common ground. A study by Gerwing & Bavelas (2004) has already analysed imagistic gestures in association with common ground. The findings revealed that the gestures produced when common ground did exist were less precise, complex and informative than corresponding gestures produced without common ground.

The present study is a first attempt to find out more about how exactly the semantic information encoded in gesture and speech is affected when common ground does or does not exist. In the first instance, the analysis focuses on the communication of size information only.

2 Empirical Study

2.1 Experimental design and procedure

40 participants took part in this study, who were randomly allocated to one of two groups, a common ground (CG) and a no common ground (NCG) condition. The participants took part in pairs, and took on either the role of the speaker or the role of the listener.

The pairs were asked to collaborate on a referential communication task, which involved the speakers describing the location of a certain target entity in three different pictures which also showed many other entities. The listener was asked to later locate the target entity based on the speaker’s description. The two experimental conditions differed in that in the CG condition

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 197–198.

speaker and listener were given the opportunity to jointly examine the picture without the target entity in it before the start of the actual task. This was thought to induce common ground (i.e. knowledge about the entities shown in each picture).

2.2 Analysis

The speech and gesture data were analysed with respect to how the speakers encoded size information relating to certain particularly large entities shown in each picture. Furthermore, two independent judges were asked to rate the size of the respective entities represented by the gestures that accompanied these references (agreement = 74%).

2.3 Results and Conclusion

The results revealed that how speakers encoded size information in the NCG and CG conditions differed significantly. When size information was encoded verbally, speakers from the NCG condition were more likely to accompany their references with gestures, while in the CG condition they tended to produce purely verbal utterances. Further, when we considered only those gestures that did represent the accurate size of the respective entities, it was found that they were significantly more likely to occur in the NCG condition than in the CG condition. Overall, speakers in the NCG condition represented size predominantly either in gesture only, or in gesture and speech, whereas speakers in the CG condition represented the information mainly exclusively verbally.

The findings are interpreted with respect to the pragmatic functions gesture may fulfil during dialogue. They provide important insights regarding gesture production theories and they advance our understanding of how people in talk use language.

3 Work in Progress

The present study is a first attempt to determine how the semantic information conveyed by gesture and speech is affected by common ground. Work currently in progress builds on this study by focusing on a more collaborative, interactive setting. Using a variation of Clark and Wilkes-Gibbs' (1986) tangram task it explores the role of gestural communication in the accumulation of common ground over the course of a conversation, focussing in particular on gestural and verbal alignment as well as systematic changes in how information is encoded in gesture and speech over the consecutive trials. This work will provide us with further insights into how speakers use language to collaborate in talk.

References

Alibali, M.W., Heath, D.C. & Meyers, H.J. (2001). Effects of visibility between speakers and listeners on gesture production: Some gestures are meant to be seen. *Journal of Memory & Language*, *44*, 169-188.

Bavelas, J.B., Kenwood, C., Johnson, T. & Phillips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, *2*, 1–17.

Beattie, G. & Aboudan, R. (1994). Gestures, pauses and speech: An experimental investigation of the effects of changing social context on their precise temporal relationships. *Semiotica*, *99*, 239-272.

Clark, H.H., & Schäfer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*, 259-294.

Clark, H.H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1-39.

Fussell, S.R. & Krauss, R.M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*, 203-219.

Gerwing, J. & Bavelas, J.B. (2004). Linguistic influences on gesture's form. *Gesture*, *4*, 157–195.

Holler, J. & Beattie, G. (2003b). Pragmatic aspects of representational gestures: do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, *3*, 127-154.

Isaacs, E.A., & Clark, H.H. (1987). References in conversations between experts and novices. *Journal of Experimental Psychology: General*, *116*, 26-37.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Programming by Voice: enhancing adaptivity and robustness of spoken dialogue systems

Kallirroi Georgila and Oliver Lemon
School of Informatics, University of Edinburgh
`kgeorgil,olemon@inf.ed.ac.uk`

Abstract

This demonstration system allows users to reconfigure dialogue systems by using speech dialogues to build simple programs for devices and services. This is a novel type of adaptivity – where the user is able to explicitly adapt some aspects of the dialogue system to their own needs, and is a capability beyond any commercially deployed systems. The main idea is to extend command-based and information-seeking dialogue systems so that users can reconfigure them to perform common tasks, or to behave in specific ways in certain contexts that are of interest to the user. We present a voice-programming (VP) system for device control and information seeking, using the extended in-car "TownInfo" dialogue system (Lemon et al., 2006) developed in the TALK project[1] and built using DIPPER (Bos et al., 2003) and ATK (Young, 2004).

1 Introduction

Most users do not want to learn complex operating instructions for devices and services, so an alternative is to allow them to create their own commands and programs. Users may also wish to configure their devices to carry out specific actions which are tailored to their needs and preferences. One way of doing this is to allow users a level of "programming" access to the interfaces themselves.

For example, in an automated home environment, by uttering a simple phrase such as "I want to relax" a user may request the home control system to perform a number of pre-defined tasks such as turning down the lights, playing classical music, and switching the telephone off. In a similar way, users could call their house when away and define external-event-driven programs such as "Only turn the heating on if the temperature falls below 10 degrees". Similar scenarios can be imagined for in-car device-control dialogues, for example "'I need some peace' means turn the stereo off and close the windows", or "Open the sun roof if the temperature goes above 20".

The idea of voice programming for services is similar – it is much faster, easier, and more robust for the user of a tourist information service to say, for example "show me my favourites" than "show me all expensive French restaurants in the centre of the city".

1.1 Improved robustness

Voice programming is not only a matter of convenience and efficiency for the user but also leads to potential robustness gains. Considering speech recognition limitations, especially in noisy environments such as cars, shorter and more precise commands will in general lead to fewer errors and increase overall user satisfaction. Likewise, if users can define the semantics of their utterances through voice programs, fewer clarifications and confirmations will be required in dialogues.

1.2 Related work

The Metafor project (Liu and Lieberman, 2005) explored the idea of using descriptions in natural language as a representation for programs (Python code). Metafor does not convert arbitrary English to fully specified code, but uses a reasonably expressive subset of English as a visualization tool. Simple descriptions of program objects and their behaviour generate scaffolding (underspecified) code fragments, that can be used as feedback for the designer. In contrast, our system allows users to generate fully working programs via speech dialogues alone.

---

[1] http://www.talk-project.org

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 199–200.

2 The demonstration system

The current system focuses on controlling devices and services using programs which are:

- activated by speech commands or environmental events
- defined by the user via speech dialogues.

The basic system that we will demonstrate shows programming by voice of macros and conditionals for a tourist information service, and uses the ATK speech recogniser (Young, 2004) and DIPPER dialogue manager (Bos et al., 2003).

The capabilities of the demonstration system[2] are implemented by extending the Information State definitions with fields for macro and conditional names, which can take appropriate arguments (sequences of commands and/or slot values), and adding update rules for interpreting and processing voice programming utterances. In addition, we compiled a language model for voice programming from a GF grammar (Ranta, 2004), and extended the system's parser.

2.1 Defining Macros

A macro is a way for the user to automate a complex task that he/she performs repeatedly or on a regular basis. It is a series of commands or information slots that can be stored and run/accessed whenever the user needs to perform the task. The user can record or build a macro, and then play the macro to automatically activate the series of actions.

The syntax for a macro is:
`macro_name = slotValue/command_1 and ... slotValue/command_N`

In the tourist-information service demo a supported macro is: "When I say 'romantic dinner' I mean an expensive Italian restaurant in the town centre".

After a Wizard-of-Oz data collection for voice programming dialogues, we have extended the coverage of our system (Lemon et al., 2006) to interpret some types of user utterances as macro definitions.

For example user inputs such as:

- When/If/Whenever I say T, it means/I mean X_1 ... X_n
- X_1 ... X_n when/if/whenever I say T

are interpreted as defining a macro with trigger phrase T and which stands for commands/information slot values X_1 ... X_n.

Note however that in terms of the dialogue context, the effect of "X_1 ... X_n" is not the same as if the user had actually uttered the individual X_i. For example the salient NPs in each X_i are not available for anaphoric reference. Exactly what the effects on the dialogue context should be is a matter for ongoing research.

The previous example ('romantic dinner') is stored in the information state as a list. When the user utters the macro name, the system will retrieve the associated slots with their values and try to satisfy the user's request.
`[macro, 'romantic dinner', restaurant, [[price_range],[food_type],[location]], [[expensive],[italian],[central]] ]`

2.2 Defining Conditionals

The syntax for conditionals is: `if condition=true then slotValue/command_1 ...slotValue/command_N` or `execute macro_name`

A typical example of a conditional for programming services in the demo system is: "When I ask for pizza make it expensive". This conditional is stored in the information state as follows:
`[cond, restaurant, [food_type], [pizza], [[price_range]], [[expensive]] ]`

3 Summary

We demonstrate a novel dialogue system for Programming by Voice which leads to enhanced adaptivity and robustness of spoken dialogue systems.

References

Johan Bos, Ewan Klein, Oliver Lemon, and Tetsushi Oka. 2003. DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. In *4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124, Sapporo.

Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle. 2006. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system. In *Proceedings of EACL*.

Hugo Liu and Henry Lieberman. 2005. Metafor: Visualizing stories as code. In *Proceedings of the ACM International Conference on Intelligent User Interfaces, IUI*. ACM.

Aarne Ranta. 2004. Grammatical framework: A type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189.

Steve Young, 2004. *ATK: An Application Toolkit for HTK, version 1.4.*

---

[2] Macros and Conditionals are functional at the time of writing, and we expect Loops and Iteration to be supported by the time of the conference.

Author Index

Ai, Hua, 181
Alexandersson, Jan, 2
Allen, James F., 1
André, Elisabeth, 80
Arstein, Ron, 56

Bard, Ellen Gurman, 179
Becker, Stephanie, 171
Bergmann, Kirsten, 90
Beyssade, Claire, 42
Brew, Chris, 191
Brown-Schmidt, Sarah, 177

Callaway, Charles B., 10
Cann, Ronnie, 155

DeVault, David, 139
Diderichsen, Philip, 98
Dowding, John, 173
Dzikovska, Myroslava O., 10

Ehlen, Patrick, 173

Fischer, Kerstin, 163
Foster, Mary Ellen, 183

Georgila, Kallirroi, 199
Ginzburg, Jonathan, 114
Guhe, Markus, 179
Gunlogson, Christine, 177
Guyomard, Marc, 131

Hara, Yuri, 50
Healey, Patrick G.T., 122
Holler, Judith, 197
Hulstijn, Joris, 189

Inoue, Naomi, 165
Ito, Sadanori, 165
Iwasawa, Shoichiro, 165

Karagjosova, Elena, 26
Keizer, Simon, 169
Kelleher, John D., 81

Kempson, Ruth, 155
Kirschner, Manuel, 193
Kleinbauer, Thomas, 171
Knoeferle, Pia, 175
Knott, Alistair, 18
Kogure, Kiyoshi, 165
Kopp, Stefan, 90
Kranstedt, Alfred, 82
Krifka, Manfred, 55
Kruijff, Geert-Jan M., 81
Kruijff-Korbayová, Ivana, 147

Laidebeure, Stéphane, 173
Larsson, Staffan, 187
Lascarides, Alex, 64
Lemon, Oliver, 185, 199
Lesch, Stephan, 171
Liu, Xingkun, 185
Louwerse, Max M., 179
Lücking, Andy, 72, 82, 106

Macura, Zoran, 114
Marandin, Jean-Marie, 42
Mase, Kenji, 165
Mills, Gregory J., 122
Mironovschi, Lilia, 167
Moore, Johanna D., 10
Morante, Roser, 169

Nakatsu, Crystal, 191
Niekrasz, John, 173

Peters, Stanley, 173
Pfeiffer, Thies, 82
Pfleger, Norbert, 2
Poesio, Massimo, 56
Purver, Matthew, 173

Rieser, Hannes, 72, 82, 106
Rodriguez, Monica, 175
Roque, Antonio, 181

Saget, Sylvie, 131

Schegloff, Emanuel, 130
Shapiro, Daniel, 185
Shi, Hui, 163
Staudacher, Marc, 72, 82, 106
Steedman, Markus, 179
Stevens, Rachel, 197
Stone, Matthew, 10, 64, 139
Suzuki, Noriko, 165

Tanenhaus, Michael T., 177
Tenbrink, Thora, 163
Thomas, Kavita, 34
Tollander, Carl, 185
Toriyama, Tomoji, 165
Traum, David, 181

Umata, Ichiro, 165

Villing, Jessica, 187
Vlugter, Peter, 18

Watson, Duane G., 177
Wolska, Magdalena, 147

Yoshida, Etsuko, 195