PROCEEDINGS OF

DiaHolmia

2009 WORKSHOP ON THE SEMANTICS AND PRAGMATICS OF DIALOGUE

ROYAL INSTITUTE OF TECHNOLOGY

24-26 JUNE, 2009

STOCKHOLM, SWEDEN

Proceedings of DiaHolmia

2009 Workshop on the Semantics and Pragmatics of Dialogue

Edited by

Jens Edlund, Joakim Gustafson, Anna Hjalmarsson and Gabriel Skantze

Sponsored by

Vetenskapsrådet

Endorsed by

SigDial

ISCA

SigSem

24-26 June 2009

Stockholm, Sweden

Royal Institute of Technology (KTH)
Department of Speech Music and Hearing
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

Stockholm 2009

Preface

We are very pleased to present the proceedings of DiaHolmia – the 2009 Workshop on the Semantics and Pragmatics of Dialogue (SemDial). The SemDial series of workshops brings together researchers working on the semantics and pragmatics of dialogue in fields such as artificial intelligence, computational linguistics, formal semantics/pragmatics, philosophy, psychology, and neural science. The 2009 workshop is hosted by the Department of Speech Music and Hearing, KTH (Royal Institute of Technology). KTH is Scandinavia's largest institution of higher education in technology and is located in central Stockholm (Holmia in Latin).

We received a total of 20 full paper submissions for the workshop, of which 13 were accepted and are included in this volume. In addition, 7 posters and 5 demos were accepted, and descriptions of these are also included. Finally, abstracts and/or full articles from four invited keynote speakers are included. We would like to thank all authors for the effort they spent on their submissions.

We are grateful for the work of the Programme Committee and for their advice in selecting papers for the workshop. The review process was facilitated by the EasyChair system. The names of the Programme Committee members are listed on the next page.

We wish to thank the people at the Department of Speech Music and Hearing who have all helped organise the event. We are also grateful for financial support from Vetenskapsrådet (the Swedish research council).

Last but not least we would like to thank our invited speakers: Harry Bunt from Tilburg University, Nick Campbell from TCD, Julia Hirschberg from Columbia University, and Sverre Sjölander from Linköping University. We are convinced that their contributions will be very valuable to the workshop.

We wish all workshop participants an enjoyable and fruitful three days and we hope that all readers of the proceedings will benefit from the contents.

Jens Edlund, Joakim Gustafson, Anna Hjalmarsson and Gabriel Skantze (KTH)

Organising Committee

Programme Committee

Jan Alexandersson

Srinivas Bangalore

Ellen Gurman Bard

Anton Benz

Johan Bos

Johan Boye

Harry Bunt

Donna Byron

Jean Carletta

Rolf Carlson

Robin Cooper

Paul Dekker

Giuseppe Di Fabbrizio

Raquel Fernández

Claire Gardent

Simon Garrod

Jonathan Ginzburg

Pat Healey

Peter Heeman

Mattias Heldner

Joris Hulstijn

Michael Johnston

Kristiina Jokinen

Arne Jönsson

Alistair Knott

Ivana Kruijff-Korbayova

Staffan Larsson

Oliver Lemon

Ian Lewin

Diane Litman

Susann Luperfoy

Colin Matheson

Nicolas Maudet

Michael McTear

Wolfgang Minker

Philippe Muller

Fabio Pianesi

Martin Pickering

Manfred Pinkal

Paul Piwek

Massimo Poesio

Alexandros Potamianos

Matthew Purver

Manny Rayner

Hannes Rieser

Laurent Romary

Alex Rudnicky

David Schlangen

Stephanie Seneff

Ronnie Smith

Mark Steedman

Amanda Stent

Matthew Stone

David Traum

Marilyn Walker

Mats Wirén

Workshop Programme

Wednesday, June 24

14:00-15:00 **Registration**

15:00-16:00 **Keynote**: Julia Hirschberg
Turn-taking vs. backchanneling in spoken dialogue systems

16:00-16:40 Volha Petukhova and Harry Bunt
Who's next? Speaker-selection mechanisms in multiparty dialogue

16:40-17:20 Anna Hjalmarsson
On cue - additive effects of turn-regulating phenomena in dialogue

18:00 **Reception**

Thursday, June 25

09:00-10:00 **Keynote**: Harry Bunt
Multifunctionality and multidimensional dialogue semantics

10:00-10:30 **Coffee break**

10:30-11:10 Massimo Poesio and Hannes Rieser
Anaphora and Direct Reference: Empirical Evidence from Pointing

11:10-11:50 Ron Artstein, Sudeep Gandhe, Michael Rushforth and David Traum
Matching User Questions to Domain Speech Acts for a Tactical Questioning Dialogue System

11:50-13:00 **Lunch**

13:00-14:00 **Keynote**: Sverre Sjölander
Animal communication - bluffing, lying, impressing, and sometimes even information

14:00-14:40 David Schlangen
What we can learn from dialogue systems that don't work

14:40-15:10 **Coffee break**

15:10-15:50 Robin Cooper and Staffan Larsson
Compositional and ontological semantics in learning from corrective feedback and explicit definition

15:50-16:30 Ruth Kempson, Eleni Gregoromichelaki, Matt Purver, Greg Mills, Andrew Gargett and Christine Howes
How mechanistic can accounts of interaction be?

18:00 **Workshop dinner**

Friday, June 26

09:00-10:00 **Keynote**: Nick Campbell
The expanding role of prosody in speech communication technology

10:00-10:30 **Coffee break**

10:30-11:10 Elena Karagjosova
A monotonic model of denials in dialogue

11:10-11:50 Peter Ljunglöf
Dialogue Management as Interactive Tree Building

11:50-13:00 **Lunch**

13:00-14:00 **Demos**

Peter Ljunglöf
TRIK: A Talking and Drawing Robot for Children with Communication Disabilities

Staffan Larsson and Jessica Villing
Multimodal Menu-based Dialogue in Dico II

Ron Artstein, Sudeep Gandhe, Michael Rushforth and David Traum
Demonstration of the Amani Tactical Questioning Dialogue System

Gabriel Skantze and Joakim Gustafson
Multimodal interaction control in the MonAMI Reminder

Jens Edlund
Spontal - a first glimpse of a Swedish database of spontaneous speech

Posters

Samer Al Moubayed
Prosodic Disambiguation in Spoken Systems Output

Srinivasan Janarthanam and Oliver Lemon
Learning Adaptive Referring Expression Generation Policies for Spoken Dialogue Systems using Reinforcement Learning

Nuria Bertomeu and Anton Benz
Ontology Based Information States for an Artificial Sales Agent

Elena Andonova and Kenny R. Coventry
Alignment and Priming of Spatial Perspective

Jenny Brusk
Using Screenplays as Corpus for Modeling Gossip in Game Dialogues

Lluís F Hurtado, Encarna Segarra, Fernando Garcia, Emilio Sanchis and David Griol
The Acquisition of a Dialog Corpus with a Prototype and two WOz

Timo Baumann
Integrating prosodic modelling with incremental speech recognition

14:00-14:40 Vladimir Popescu and Jean Caelen
The Non-Individuation Constraint Revisited: When to Produce Free Choice Items in Multi-Party Dialogue

14:40-15:20 Rieks op den Akker and David Traum
A comparison of addressee detection methods for multiparty conversations

15:20-15:50 **Coffee break**

15:50-16:30 Jan Kleindienst, Jan Curin and Martin Labsky
A domain ontology based metric to evaluate spoken dialog systems

16:30-17:10 Robert Ross and John Bateman
Agency & Information State in Situated Dialogues: Analysis & Computational Modelling

17:10 **Closing**

Table of Contents

Keynotes

Full papers

Demo descriptions

Poster descriptions

Turn-taking vs. backchanneling in spoken dialogue systems

Julia Hirschberg
Columbia University
New York, United States
julia@cs.columbia.edu

Abstract

Listeners have many options in dialogue: They may interrupt the current speaker, take the next turn after the speaker has finished, remain silent, or backchannel, to indicate that they are attending, without taking the turn. In this talk I will discuss two of these options which are particularly difficult, yet particularly important, to model in spoken dialogue systems: taking the turn vs. backchanneling. How can the system determine which option the user is taking? How can the system decide which option it should take, and when? I will describe results of an empirical study of these phenomena in the context of a larger study of human-human turn-taking behavior in the Columbia Games Corpus. This is joint work with Agus Gravano.

Multifunctionality and multidimensional dialogue semantics

Harry Bunt
Tilburg Center for Creative Computing
Tilburg, the Netherlands
harry.bunt@uvt.nl

Abstract

This paper addresses the following questions: (1) Is it true, as is often claimed, that utterances in dialogue tend to have multiple functions? (2) If so, then what are the reasons for that? (3) How many functions does a dialogue utterance typically have, and which factors determine this? (4) What consequences does this have for the computational semantics of dialogue utterances? Answers to these questions are sought by investigating a dialogue corpus annotated with communicative functions using various segmentation and annotation strategies.

1 Introduction

Traditional approaches to the analysis of sentence meaning notoriously fail when applied to dialogue utterances. This is partly because these approaches are rooted in the truth-conditional view of meaning, while dialogue utterances like *Good morning?*, *Yes okay* and *Let me see...* have meanings that cannot be captured in terms of the truth or falsity of propositions.

Alternatively, the semantics of dialogue utterances has been studied in terms of information-state update (ISU) or context-change (CC) approaches Traum & Larsson, 2003) , which view utterance meanings in terms of changes in the information states (or 'contexts') of the dialogue participants. These approaches closely relate to the ideas of speech act theory, which regard the use of language as the performance of communicative actions.

A complication that these approaches have to face is that, contrary to what speech act theory tells us, dialogue utterances often have multiple communicative functions, such as answering a question but also providing feedback on the understanding of the question, and also taking the turn. The following example illustrates this.

(1)
1. A: What time is the next train to Amersfoort?
2. B: Let me see.... That will be at 11:25.
3. A: Is there no train to Amersfoort before 11:25?
4. B: Amersfoort? I'm sorry, I thought you said Apeldoorn.

Utterance 3 shows that A assumes that B understood the question 1, when he answered it in 2. He did not question B's understanding of the question, even though the answer surprised him.

The first part of B's utterance 2 is also worth considering: why does B stall for time by saying *Let me see....*? This is because he needs a bit of time to find the information that A asked for, but then why doesn't he just wait until he has found that information before starting to speak? This must be because he has decided to take the turn, so the utterance *Let me see* in fact has two functions: B signals that (1) he takes the turn; and (2) that he needs a bit of time to formulate his contribution (the answer to A's question).[1]

This example illustrates that dialogue utterances often do not correspond to a single speech act, but to sets of speech acts. Moreover, some of these speech act types, such as feedback acts and turn-taking acts have hardly if at all been studied in speech act theory, and do not easily fit within that theory. Approaches to dialogue semantics in terms of updating models of information states or dialogue contexts have therefore in fact not related closely to speech act theory, but rather to modern, data-driven versions of 'dialogue act' theory, such as DIT (see Section 2).

[1] This is common for a turn-initial stalling act. A turn-*internal* stalling act, by contrast, usually has a turn-*keeping* rather than a turn-*taking* function, as in *That will be... let me see... at 11:25.*

One of the reasons why dialogue utterances often have multiple communicative functions is that, in addition to the functions which are signaled through observable utterance features (choice of words, word order, intonation, accompanying gestures,...), other functions are often *implied* by what is signaled. Example 1 illustrates this as well: in the first part of B's utterance 2 the speaker signals that he is stalling for time through the use of the expression *Let me see* and slowing down; by implication the utterance also constitutes a turn-taking act. The second part constitutes an answer due to its form and content plus the fact that it follows a question; by implication it also gives the feedback information that A's question was well understood. In Section 3 we will discuss the issue of implied functions in more detail, as well as other reasons why dialogue utterances often have multiple functions.

In the literature, claims about the multiple functionality of dialogue utterances are often motivated by isolated examples like (1), rather than by quantitative studies of corpus data; moreover, the claimed multifunctionality of utterances is highly dependent on what is meant by 'utterance', as well as by the spectrum of communicative functions that is considered. In Section 3 we will discuss the definition of 'utterance' in the light of segmenting a dialogue into meaningful units, and in Section 2 we will introduce a rich, well-motivated taxonomy of communicative functions for the analysis in the rest of the paper. In Section 4 we discuss the various ways in which one dialogue act may imply another. Section 5 is devoted to an empirical study of the multifunctionality of utterances in a dialogue corpus, and Section 6 ends the paper by summarizing the answers to the questions that were raised in the abstract.

2 Theoretical framework

2.1 Dialogue acts and utterance meanings

The semantic framework of Dynamic Interpretation Theory (DIT, see Bunt, 2000; 2009)) takes a multidimensional view on dialogue in the sense that participation in a dialogue is viewed as performing several activities in parallel, such as pursuing a task or activity that motivates the dialogue, providing and eliciting communicative feedback, taking turns, managing the use of time; and taking care of social obligations. The activities in these various dimensions are called dialogue acts and are formally interpreted as update operations on the information states (or 'context models')[2]; of the dialogue participants. Dialogue acts have two main components: a semantic content which is to be inserted into, to be extracted from, or to be checked against the current information state; and a communicative function, which specifies more precisely how an addressee updates his information state with the semantic content when he understands the corresponding aspect of the meaning of a dialogue utterance.

DIT distinguishes the following 10 dimensions (for discussion and justification see Petukhova & Bunt 2009a; 2009b):

1. Task/Activity: dialogue acts whose performance contributes to performing the task or activity underlying the dialogue;
2. Auto-Feedback: dialogue acts that provide information about the speaker's processing of the previous utterance(s);
3. Allo-Feedback: dialogue acts used by the speaker to express opinions about the addressee's processing of the previous utterance(s), or that solicit information about that processing;
4. Contact Management: dialogue acts for establishing and maintaining contact;
5. Turn Management: dialogue acts concerned with grabbing, keeping, giving, or accepting the sender role;
6. Time Management: dialogue acts signalling that the speaker needs a little time to formulate his contribution to the dialogue;
7. Discourse Structuring: dialogue acts for explicitly structuring the conversation, e.g. announcing the next dialogue act, or proposing a change of topic;
8. Own Communication Management: dialogue acts where the speaker edits the contribution to the dialogue that he is currently producing;
9. Partner Communication Management: the agent who performs these dialogue acts does not have the speaker role, and assists or corrects the speaker in formulating a contribution to the dialogue;

[2] In the rest of this paper, we will use the terms 'information state', and 'context' (or 'context model') interchangeably, as also the terms 'information state update, 'context change' and 'context model update'.

- *Information Transfer Functions*
  - *information-seeking functions*
    - *Direct Questions*
      - propositional question, set question, alternatives question, check question, etc.
    - *Indirect Questions*
      - indirect propositional question, set question, alternatives question, check question, etc.
  - *information-providing functions:*
    - *informing functions:*
      - inform, agreement, disagreement, correction;
      - *informs with rhetorical functions such as:*
    - *answer functions:*
      - propositional answer, set answer, confirmation, disconfirmation
- *Action Discussion Functions*
    - *Commissives*
      - offer, promise, address request
      - other commissives, expressable by means of performative verbs
  - *Directive functions:*
    - instruction, address request, indirect request, (direct) request, suggestion
    - other directives, such as advice, proposal, permission, encouragement, urge,..., expressable by means of performative verbs

Table 1: Structure of the DIT^{++} taxonomy of general-purpose communicative functions.

10. Social Obligations Management: dialogue acts that take care of social conventions such as greetings, apologies, thanking, and saying goodbye.

One of the products of DIT is a multidimensional taxonomy of communicative functions, called the DIT^{++} taxonomy, designed for the purpose of dialogue act annotation and dialogue system design across a wide range of domains,[3] and which includes elements from various other annotation schema, such as the DAMSL, TRAINS, and Verbmobil taxonomies (Allen & Core, 1997; Allen et al., 1994; Alexandersson et al., 1998). Multidimensional taxonomies support dialogue utterances to be coded with multiple tags and have a relatively large tag set; such a tag set may benefit in several respects from having some internal structure.

First, clustering semantically related tags improves the transparency of the tag set for human users, as the clusters indicate the kind of semantic information that is addressed. Second, introducing a hierarchical or taxonomical structure which is based on semantic clustering may support the decision-making process of human annotators: an initial step in such a process can be the decision to consider a particular cluster, and subsequently more fine-grained distinctions may be tested in order to decide on a specific tag within the cluster. Third, a hierarchical organisation in the tag set may also be advantageous for automatic annotation and for achieving annotations which are compatible though not identical with those of human annotators (namely, the automatic annotation may use less specific tags than the human annotation). In general, a structured tag set can be searched more systematically (and more 'semantically') than an unstructured one, and this can clearly have advantages for dialogue annotation, interpretation, and generation.

Bunt (2005; 2006) suggests that the structure of a multidimensional annotation schema should be based not just on a clustering of intuitively similar functions, but on a well-founded notion of *dimension*, and proposes to define a *set of dimensions* as follows.

(2) Each member of a set of dimensions is a cluster of communicative functions which all address a certain aspect of participating in dialogue, such that:

1. dialogue participants can address this aspect through linguistic and/or nonverbal behaviour which has this specific purpose;
2. this aspect of participating in a dialogue can be addressed independently of the aspects corresponding to other members of the set of dimensions, i.e., an utterance can have a communicative function in one dimension, independent of its functions in other dimensions.

The first condition means that only aspects of communication are considered that are observed in actual communicative behaviour; the second that dimensions should be independent. A set of dimensions that satisfies these requirements can be useful for structuring an annotation schema, especially if the set of functions within each dimension is defined in such a way that any two functions are either mutually exclusive or have an entailment relation. In that case a functional unit can be annotated with (maximally) as many tags as there are dimensions, one function (at most, namely the most specific function for which there is evidence that it should be marked) for each dimension.

[3] See `http://dit.uvt.nl`.

Dimension	*Dimension-specific functions*	*Representative expressions*
Task/Activity	OpenMeeting, CloseMeeting; Appoint, Hire, Fire	domain-specific fixed expressions
Auto-Feedback	PerceptionNegative	*Huh?*
	EvaluationPositive	*True.*
	OverallPositive	*OK.*
Allo-Feedback	InterpretationNegative	*THIS Thursday.*
	EvaluationElicitation	*OK?*
Turn Management	TurnKeeping	final intonational rise
	TurnGrabbing	hold gesture with hand
	TurhGiving	*Yes.*
Time Management	Stalling	slowing down speech; fillers
Contact Management	ContactChecking	*Hello?*
Own Communication Man.	SelfCorrection	*I mean...*
Partner Communication Man.	PartnerCompletion	completion of partner utterance
Discourse Structure Man.t	DialogueActAnnouncement	*Question.*
	TopicShiftAnnouncement	*Something else.*
Social Obligations Man.	Apology	*I'm sorry.*
	Greeting	*Hello!, Good morning.*
	Thanking	*Thanks.*

Table 2: Examples of dimension-specific communicative functions and representative expressions for each dimension.

When we view a dimension in dialogue analysis in accordance with (2) as a particular aspect of interacting, like the 10 dimensions mentioned above, we see that dialogue acts like question and answer do not belong to any dimension. This is because one can ask a question about something in the task, or a about agreeing to close a topic, or about whose turn it is to say something, or about any other aspect of interacting, so questions can belong to *all* these dimensions. Every *occurrence* of a question function, as the function of a dialogue act that is performed, falls within one of the dimensions; which dimension is determined by the type of semantic content. Similarly for answers, statements, requests, offers, agreements, (dis-)confirmation, and so on. Clusters of such general types of dialogue acts therefore do not form a dimension, but can be used in any dimension; they are called *general-purpose functions*. This in contrast with communicative functions that are specific for a particular dimension, such as Turn Keep, Turn Release, Introduce Topic, Change Topic, Apology and Thanking. The DIT^{++} taxonomy therefore consists of two parts: (1) a taxonomy of *general-purpose functions*; (2) a taxonomy of *dimension-specific functions*. Table 1 shows the structure of the taxonomy of general-purpose functions; Table 2 lists examples of dimension-specific communicative functions in each of the DIT^{++} dimensions.

In order to define a context-change semantics for all the types of dialogue acts in the DIT^{++} taxonomy, the context models on which the semantics is based should contain all the types of information addressed by these dialogue acts.

Table 3 lists these types, and illustrates their use by dialogue utterances whose update semantics involves these types of information.

3 Multifunctionality and segmentation

Allwood (1992) distinguished two forms of multifunctionality, called sequential and *simultaneous*, using the following example:

(3) A: Yes! Come tomorrow. Go to the church!
Bill will be there, OK?
B: The church, OK.

Allwood observes: "A's utterance in the example contains sequentially the functions *feedback giving, request, request, statement and response elicitation.* Furthermore, the statement 'Bill will be there' could simultaneously be a promise and thus illustrates simultaneous multifunctionality." It should be noted that the term 'utterance' is used here in the sense of *"unit in spoken dialogue which corresponds to a stretch of speech from one speaker, bounded by lack of activity or another communicator's activity."* Utterances in this sense, which are more commonly called *turns* are often quite complex, and it is no wonder that they are often sequentially multifunctional. It is therefore more common to consider smaller functional units within turns, and refer to these units as 'utterances', as we shall also do in the rest of this paper.

Utterances in the latter sense are defined as contiguous stretches of linguistic behaviour which form grammatical units that have a communicative function. Segmenting a dialogue into utterances has the advantage of being more fine-grained than a segmentation into turns, and thus allowing a more precise functional markup; on the other hand, the determination of utterance boundaries (as opposed to turn boundaries) is a highly nontrivial task. Syntactic and prosodic features are often used as indicators of utterance endings (e.g. Shriberg et al., 1998; Stolcke et al., 2000; Nöth et al., 2002), but are in general not very reliable. In the case of nonverbal or multimodal communication, the notion of an utterance as a linguistically defined unit is even less clear.

Segmenting a dialogue into utterances has the effect of eliminating sequential multifunctionality. There are however other, segmentation-related forms of multifunctionality that remain, namely *discontinuous, overlapping*, and *interleaved multifunctionality*. The first of these occurs when an utterance embeds a smaller utterance which has a different communicative function. The following example illustrates this.

(4) 1. C: What time is the first train to the airport on Sunday?
2. I. The first train to the airport on Sunday is at... *let me see...* 5.32.

Here we see a discontinuous answer *The first train to the airport on Sunday is at [......] 5.32* to the preceding question. Example (4) also illustrates the phenomenon of overlapping multifunctionality, which occurs when part of an utterance with a certain function forms a sub-utterance with another function. In the example, the sub-utterance *The first train to the airport on Sunday* has the function of providing positive feedback on the understanding of the question, while the utterance as a whole answers the question.

Interleaved multifunctionality occurs when two utterances with different functions are interleaved to form a complex utterances, and is illustrated by the following example.

(5) I think twenty five euros for a remote... *is that locally something like fifteen pounds?...* is too much money to buy an extra remote or a replacement one .. *or is it even more in pounds?*

Here we see the discontinuous statement *I think twenty five euros for a remote [...] is too much money to buy an extra remote or a replacement one* interleaved with the discontinuous question *is that locally something like fifteen pounds [...] or is it even more in pounds?* These examples show that the segmentation of dialogue into utterances in the usual sense does not lead to distinguishing the stretches of behaviour that form functional units. Instead, such units should be allowed to be discontinuous, to overlap, and to be interleaved. To avoid terminological confusion, we use the term *functional segment* for this purpose (see further Geertzen et al., 2007).[4]

4 Types of multifunctionality

The multifunctionality of dialogue utterances not only takes several forms, as noted above (sequential, simultaneous, interleaved), but also comes in semantically different varieties. The following four types can be distinguished:

independent: a functional segment has more than one communicative function, due to having features expressing each of these functions;

entailed: a functional segment has two (or more) communicative functions because one function logically entails another;

implicated: a functional segment has two (or more) communicative functions because one function is conversationally implicated by another function;

indirect: the segment constitutes an indirect dialogue act, i.e. it has another communicative function than it would appear at first sight, which can be inferred from its 'literal' function in the context in which it occurs.

We discuss each of these types of multifunctionality in turn.

[4] A functional segment may also spread over multiple turns, as the following example shows:
A; Could you tell me what departure times there are for flights to Frankfurt on Saturday?
B: Certainly. There's a Lufthansa flight leaving at 08:15,
A: yes,
B: and a KLM flight at 08:50,
A: yes,
B: then there's a flight by Philippine airlines,...
In this example the A's question to consists of a list of items which B communicates one by one in separate turns in order not to overload A.

example utterance	*dialogue act type*	*information category*
Can I change the contrast now?	Task-related propositional question	task information
Please press reset first	Task-related request	task information
Did you say Thursday?	Feedback check question	own processing success
Okay?	Feedback elicitation	partner processing success
Let me see,...	Stalling	processing time estimates
Just a minute	Pause	processing time estimates
Well,...	Turn Accept	turn allocation
Tom?	Turn Assign	turn allocation
Let's first discuss the agenda	Dialogue structure suggestion	dialogue plan
Can I help you?	Dialogue structure offer	dialogue plan
On june first I mean second	Self-correction	own speech production
.... you mean second	Partner correction	partner speech production
Hello?	Contact check	presence and attention
You're welcome	Thanking downplayer	social pressure

Table 3: Semantic information categories as related to dialogue act types, and example utterances.

4.1 Independent multifunctionality

A functional segment may have several independent communicative functions, in different dimensions. Examples are:

1. ”Thank you”, spoken with markedly high pitch and cheerful intonation (like goodbyes often have), to signal goodbye in addition to gratitude;
2. “Yes”, said with in intonation that first falls and subsequently rises, expressing postive feedback (successful understanding etc.) and giving the turn back to the previous speaker;
3. Turn-initial Stalling and Turn Take (or Turn Accept);
4. Excessive turn-internal Stalling and elicitation of support (i.e., eliciting an utterance completion act in the Partner Communication Management dimension).

Semantically, the interpretation of an utterance which displays independent multifunctionality comes down to two (or more) independent update operations on different dimensions of an addressee's information state, one for each communicative function.

4.2 Implied communicative functions

4.2.1 Entailed functions

It was noted in Section 1 that one of the reasons why utterances may have multiple functions, is that one function may imply another. The two implication relations that we see in example (1) above are of a different nature. The turn-taking act that is implied by the first part of utterance 2 follows from the fact that there is a stalling act in turn-initial position; the feedback act implied by the answer in the second part of 2 follows from the fact that giving an answer presupposes understanding the corresponding question. The latter case corresponds to a logical entailment relation between answers and positive feedback acts, whereas the former is context-dependent, and more like a conversational implicature.

In the case of an entailment relation, a functional segment has a communicative function, F_1 expressed by utterance features, which is characterized by a set of preconditions which logically imply those of a dialogue act with the same semantic content and with the communicative function F_2.

Some examples of entailment relations between dialogue acts are:

1. Justification, Exemplification, Warning all entailing Inform; Agreement, Disagreement, Correction entailing Inform; Confirmation and Disconfirmation both entailing Propositional Answer; Check Question entailing Propositional Question;
2. Answer, Accept Offer, Reject Offer, Accept Suggestion, Reject Suggestion entailing positive feedback;
3. Responsive dialogue acts for social obligations management, such as Return Greeting and Accept Apology entailing positive feedback on the corresponding intiating acts (such as Init Greeting and Apology);
4. Evaluative feedback entailing positive feedback on perception and understanding; Negative feedback on perception entailing negative feedback on understanding (see below, Section 4.4).

Entailment relations typically occur between dialogue acts within the same dimension, and which have the same semantic content but communicative functions that differ in their level of specificity. More specific dialogue acts entail less specific ones with the same semantic content. Dialogue acts in different dimensions are concerned with different aspects of the interaction; therefore with different types of information, and hence there is usually no relation of entailment or other semantic relation between them.

Entailed functions within the same dimension correspond to the context update operation representing the entailed interpretation being subsumed by the update operation of the entailing one. They are thus semantically vacuous, and it therefore does not seem to make much sense to consider such cases as multiple functions that can be assigned to a functional segment.

Entailments may also occur also between an act in a non-feedback dimension and a feedback act. An answer, for example, is semantically related to a question, which has been expressed in a preceding utterance or sequence of utterances contributed by the dialogue partner. Relations such as the one between an occurrence of an answer and the corresponding question, are called *functional dependency relations*[5], and are part of the annotations in the corpora that we will consider in Section 5. This type of relation is relevant for answers, responses to directive dialogue acts (such as Accept Request and Reject Offer), and more generally to those dialogue acts that have a 'backward-looking function' (Allwood, 2000; Allen & Core, 1997), for which the functional dependency relation indicates the dialogue act that is responded to. This relation is of obvious importance for determining the semantic content of the responding act. Moreover, the fact that a speaker responds to a previous dialogue act implies that the speaker has (or at least believes to have) successfully processed the utterance(s) expressing the dialogue act that he responds to, and so the occurrence of a responsive dialogue act entails a positive (auto-)feedback act.

Entailed feedback acts corresponds to context-changing effects in the component of the context model that contains the speaker's assumptions about his own and his partner's processing of previous utterances. These context-changing effects are additional to those that express the semantics of the entailing responsive act, and should therefore be considered as adding an extra communicative function to the corresponding utterance.

4.3 Implicated functions

Implicated multifunctionality occurs when a functional segment has a certain communicative function by virtue of its observable features (in the given dialogue context), and also another communicative function due to the occurrence of a conversationally implicature. Like all conversational implicatures, this phenomenon is context-dependent, and the implicatures are intentional. Examples are:

1. an expression of thanks implicating positive feedback at all levels of the previous utterance(s) of the addressee;
2. positive feedback implied by shifting to a new topic, related to the previous one; more generally, by any relevant continuation of the dialogue;
3. negative feedback, implied by shifting to an unrelated topic; more generally, by any 'irrelevant' continuation of the dialogue.

Implicated functions are not expressed explicitly through the features of expressions, but can be inferred as being likely from the interpretation of the utterance features (as indicating a type of certain dialogue act) in a given context. Implicated functions are intended to be recognized, and correspond semantically to an additional context update operation, hence they are a true source of multifunctionality.

4.4 Entailed and implicated feedback functions

A speaker who provides feedback about his perception, understanding, or evaluation of previous utterances, or, in the terminology introduced above, performs an auto-feedback act, may be specific about the level of processing that his feedback refers to. For instance, a literal repetition of what was said with a questioning intonation is typically a signal that the speaker is not sure he heard well, whereas a rephrasing of what was said is not concerned with perception but with understanding. A signal of positive understanding implies that the speaker also perceived well; on the other hand, a signal of imperfect understanding implies good

[5] See also ISO (2009) for a discussion of these and other relations.

perception (or at least, the speaker whose feedback addresses the level of understanding does so with the assumption that there was no problem at the perceptual level).

In DIT, five levels of processing are distinguished which have logical relationships that turn up as implications between feedback acts at different levels:

(6) attention $<$ perception $<$ understanding $<$ evaluation $<$ execution

'Evaluation' should be understood here in relation to the information-state update approach followed in DIT, and the requirement that information states at all times be internally consistent, also when update operations are applied to them. For example, the recipient of an inform act with a semantic content p knows, upon understanding the behaviour expressing this act, that the speaker wants him to insert the information p in his information state. Before doing this, the recipient has to check whether p is consistent with his current state; if not; the update would be unacceptable. Evaluation leads to a positive result if the intended update operation is acceptable, and may be signaled by a positive feedback act referring to this level; a negative result will typically lead to a negative feedback signal. If the evaluation has a positive outcome, then the recipient can move on to the stage of execution, which is the highest level of processing of an input. For the example of the informing act with content p, execution would mean that the recipient inserts p in his information state.

When the input is a question, then the evaluation comes down to deciding whether the input can be accepted as such, e.g. does not conflict with the belief that this particular question has already been answered. Its 'execution' is then the gathering or computation of the information needed to answer the question. If execution fails, this typically leeds to a response like *I don't know*, which is viewed as a negative feedback act at execution level.

The implication relations between feedback at different levels are either *entailments* or *implicatures*. In the case of positive feedback, an act at level L_i entails positive feedback at all levels L_j where $i > j$; positive feedback at execution level therefore entails positive feedback at all other levels. By contrast, positive feedback at level L_i *implicates* negative feedback at all levels L_j where $i < j$; for instance, a signal of good perception implicates that there is a problem with understanding, for why not signal good understanding if that were the case? This is, however, not a logical necessity, but rather a pragmatic matter, hence an implicature rather than an entailment.

For negative feedback the entailment and implicature relations work in the opposite direction from positive feedback. For allo-feedback the same relations hold as for auto-feedback.

Implied feedback functions do not really constitute a separate kind of implied functions, but we distinguish them here and in the annotation strategies considered below because of there virtually ubiquitous character.

4.5 Indirect speech acts

The phenomenon known as 'indirect speech acts' is another potential source of multifunctionality. An utterance such as *Can you pass me the salt?* has been analysed as expressing both a question about the addressee's abilities and, indirectly, a request to pass the salt. Using DIT or another semantic, ISU-based approach, such an analysis does not make much sense, however, since a request to do X is normally understood to carry the assumption (on the part of the speaker, S) that the addressee (A) is able to do X; hence the interpretation of the utterance as a request would lead to an update of the context to the effect that A believes that S believes that A is able to pass the salt, while the interpretation as a question about the addressee's abilities would lead to an update including that A believes that S wants to know whether A is able to pass the salt. These two updates would be in logical conflict with each other, resulting in an inconsistent information state.

The DIT analysis of such cases is as follows. S has a goal G that could be achieved by successful performance of a dialogue act with function F_1; however, F_1 has a precondition p_1 of which S does not know whether it is satisfied, and which S believes A knows whether it is satisfied (for instance, a property of A). S therefore asks A whether p_1. A understands this situation (in fact, S and A mutually believe this situation to obtain), and understands that S wants to perform the dialogue act with function F_1 if the condition p_1 is satisfied. In other words, S's utterance is understood as a conditional request: *If you are able to pass me the salt, please do so*. Similarly, an utterance like *Do you know what time it is?* is understood as *Please tell* *me what time it is, if you know*, and *Are there any flights to Toronto this evening?* as *Which flights to Toronto are there this evening, if any?* So this type of 'indirect speech act' is viewed not as expressing multiple acts, but as expressing a single conditional dialogue act.

Another kind of indirect speech act is exemplified by *I would like to have some coffee.* This might be analysed as an inform act, and indirectly a request. The DIT analysis of such cases is as follows. Speaker S has a goal G which could be achieved by successful performance of a dialogue act with communicative function F_2 (such as Request). The utterance is interpreted as the request to A to perform the F_2 act if A is able and willing to do so. Hence again, the utterance is viewed not as expressing two dialogue acts, but rather as a single, conditional one.

Whether all types of indirect speech act can be analysed in a similar way, as corresponding to a single conditional dialogue act rather than to multiple acts, is an issue for further research. If the answer is positive, then indirect speech acts are in fact not a source of multiple functionality. If the answer is negative, or if the DIT analysis is not adopted, then it is.

5 Empirical determination of multifunctionality

The multifunctionality of utterances in dialogue can be empirically investigated given a corpus of dialogues annotated with communicative functions. We investigated the multifunctionality that is observed in a corpus of dialogues annotated with the DIT^{++} scheme, taking two variables into account:

(i) the segmentation method that is used, i.e.d, the choice of units in dialogue to which annotations are assigned; and

(ii) the annotation strategy that is used, reflecting alternative views on what counts as multifunctionality.

5.1 Experiment

Two expert annotators marked up 17 dialogues in Dutch (around 725 utterances) using the DIT^{++} scheme as part of an assessment of the usability of the annotation scheme. Several types of dialogue were included:

(1) dialogues over a microphone and head set with a WOZ-simulated helpdesk, providing assistance in the use of a fax machine (from the DIAMOND corpus[6]);

(2) human-human telephone dialogues with an information service at Amsterdam Airport;

(3) human-computer telephone dialogues about train schedules (from the OVIS corpus); [7]

(4) Dutch Map Task dialogues.

We compared three alternative segmentation methods:

a. turn-based: the turn is taken as the unit which is annotated with communicative functions;

b. utterance-based: every turn is chopped up into contiguous, non-overlapping grammatical units which have one or more communicative function;

c. functional-segment based: functional segments are distinguished for each (possibly discontinuous) stretch of behaviour which has one or more communicative function, where functional segments may be discontinuous, overlapping, and interleaved, and may spread over more than one turn.

The dialogues were segmented into functional segments and annotated accordingly; from this segmentation and annotation we reconstructed the annotation that would correspond to the coarser other two segmentation methods.

The following strategies were compared for dealing with the various possible sources of (simultaneous) multifunctionality:

a. strictly feature-based: only communicative functions are marked which are recognizable from utterance features (lexical, syntactic, prosodic), given the context of the preceding dialogue. Only explicit feedback functions are marked, and Turn Management functions are marked only if they are explicitly indicated through lexical and/or prosodic features;

b. + implicated functions: implicated functions are are marked as well;

[6] See `http://ls0143.uvt.nl/diamond`

[7] `http://www.let.rug.nl/~vannoord/Ovis`

c. + turn taking: a turn-initial segment (i.e., a functional segment occurring at the start of a turn) is marked by default as having a Turn Take function if it does not already have a Turn Grab function (i.e., it forms an interruption) or a Turn Accept function (i.e., the speaker accepts the turn that was assigned to him by the previous speaker). In other words, starting to speak is by default annotated as an indication of the Turn Take function;

d. + turn releasing: similarly, ceasing to speak is by default annotated as a Turn Release act;

e. + entailed feedback functions: entailed feedback functions are also marked, such as the positive feedback on understanding that is entailed by answering a question or accepting an offer;

f. + inherited functions: entailed functions within a dimension, due to degrees of specificity are also marked, such as a Check Question also being a Propositional Question, and a Warning also being an Inform;

g. + entailed feedback levels: signals of positive feedback at some level of processing are also marked as positive feedback at lower levels, and negative feedback at a certain level is also marked as negative feedback at higher levels;

g. + implicated feedback levels: signals of positive feedback at some level of processing are also marked as (implicated) negative feedback at higher levels; signals of negative feedback at a certain level are also marked as positive feedback at lower levels;

i. + indirect functions: in the case of indirect speech acts, both the function of the direct interpretation and the one(s) of the intended indirect interpretation(s) are marked.

The dialogues were annotated using strategy b; the annotations according to the strategies c-i were reconstructed by adding the relevant implied, indirect or default functions.

5.2 Results

The results are summarized in Table 2. The absolute figures in this table are not of great interest, given the small sample of annotated dialogue material on which they are based; relevant are especially the differences that we see depending on the segmentation method that is used and on what is considered to count as multifunctionality.

Table 4: Cumulative multifunctionality for various annotation strategies and segmentation methods.

segmentation method:	*turn*	*utter-*	*funct'l.*
annotation strategy:		*ance*	*segment*
a. strictly feature-based	2.5	1.7	1.3
b. + implicated functions	**3.1**	**2.1**	**1.6**
c. + turn taking	4.0	2.7	2.1
d. + turn releasing	4.8	3.3	2.6
e. + entailed feedback	5.2	3.6	2.8
f. + inherited functions	5.6	3.9	3.0
g. + implic. feedb. levels	6.3	4.2	3.2
h + entailed feedb. levels	6.6	4.5	3.4
i. + indirect functions	6.7	4.6	3.5

5.3 Discussion

As noted above, the annotated dialogue corpus used in the present study was marked up according to strategy b, i.e. it includes besides the communicative functions derived from utterance features also the implicated ones, except implicated functions at various feedback levels (which are taken into account in strategy g). The entailed and default functions that are additionally annotated when strategies c-f and h are applied, can all be derived automatically from the annotations resulting from strategy b.

The positive and negative feedback functions at certain levels of processing that are implicated by a feedback function at another level, and that are taken into account in strategy g, cannot be deduced from the strategy-b annotations, but these implicated functions can be assumed to occur by default, as they seem to always occur except in some very unusual dialogue situations.[8]

Indirect communicative functions, which are additionally taken into account in strategy i, cannot be deduced from strategy-b annotations in a straightforward way, but require a good understanding of the dialogue context (or a large corpus of examples in context, from which the indirect understanding might be learnable). However, we have argued above that in an ISU-based semantic framework it is highly questionable whether indirect speech acts should be treated as the occurrence of *both* a direct and an indirect act, and therefore that it can be argued that indirect speech acts do not add to the multifunctionality that is found in dialogue.

[8] Such an unusual situation may for example be that one is received by the king of a very traditional country with an extremely strict hierarchical political system, where the king is never to be contradicted or to be asked to clarify or repeat what he said.

All in all, the figures in the second row in Table (5.2) represent the *minimal degree of multifunctionality* that is found.

When the most fine-grained segmentation is applied, using functional segments, then all sequential multifunctionality is eliminated and only purely simultaneous multifunctionality remains. Using annotation strategy a, where all kinds of implicated, entailed, indirect, and default functions are left out of consideration, the annotations reflect purely the *independent multifunctionality* of functional segments. Table (5.2) shows that our data indicate that on average one in every three segments has two independent communicative functions. The minimal multifunctionality of functional segments, as just argued, is found when annotation strategy b is followed, and turns out to be 1.6 in our data. This means that on average two in every three segments have two independent communicative functions.

When utterance-based segmentation is used, we find that on average each utterance has two communicative functions. The difference with the multifunctionality of functional segments is caused by the fact that functional segments are often discontinuous. The main cause of this is the occurrence of Own Communication Management acts, where the speaker edits his contribution on the fly, interrupting his utterance by stallings, retractions, restarts, and so on.

The multifunctionality of a turn is simply the sum of the simultaneous multifunctionalities of its constituent utterances. It follows, from the figures in Table (5.2) for unsegmented turns, that in our corpus a turn on average contains one and a half contiguous utterances and nearly two functional segments. These figures may vary depending on the type of dialogue. For instance, in a meeting conversation where one participant is very dominant and produces long turns, alternated by occasional short turns from other participants, the number of utterances per turn will on average per greater. In general, the figures in the column for utterance-based segmentation have to be taken with a big grain of salt, as they depend a lot on the complexity of the turns in the dialogues that are considered.

6 Conclusions and future work

Returning to the three questions formulated at the start of this paper, we have in fact arrived at the following answers.

In response to the question whether dialogue utterances tend to have multiple functions, the answer is yes, definitely! Utterances in the usual sense, of contiguous stretches of linguistic behaviour with a grammatical status, have on average at least two functions. And if we take the most-fine-grained segmentation of dialogue into functional units and a minimal approach to the notion of multifunctionality, we still find that on average two out of every three units have more than one communicative funciton. These quantitative findings answer the first part of question 3: how many functions does an utterance typically have?

Question 2, why dialogue utterrances are multifunctional, has been answered in a theoretical sense by considering participation in a dialogue as involving multiple activities at the same time, such as making progress in a given task or activity; monitoring attention and correct understanding; taking turns; managing time, and so on. This approach has been backed up by empirical data, which show that functional segments display both what we called independent multifunctionality, having two functions in different dimensions, as well as implicated multifunctionality where the implicated function belongs to the feedback dimension(s). Entailment relations between dialogue act and default and indirect functions add further to the mulltifunctionality that can be observed.

Question 3 asks which factors influence the amount of multifunctionality that is found. The answer to this question is: first, the choice of units in dialogue which are considered as having communicative functions matters a lot. If turns are taken as units, then there is not much that can sensibly be said, due to the fact that turns may be quite complex, and therefore display sequential multifunctionality. Regardless of the choice of functional units, we have seen that the observed amount of multifunctionality depends strongly on the view that is taken on what counts as having multiple functions, and on the role that is given to implied, default, and indirect functions.

Finally, what are the consequences of the findings, reported and discussed in this paper, for the semantic interpretation of dialogue utterances? Any adequate account of the meaning of dialogue utterances will have to take their multifunctionality into consideration. Our findings confirm that the multifunctionality of functional segments can be viewed as arising only due to their meaning in different dimensions: a segment never has more than one function in any given dimension. (See the arguments above about entailed functions within a dimension being semantically vacuous.) This supports the view that an update semantics which interprets communicative functions as recipes for updating a part of the information state can be developed which uses separate updates for each dimension, which, due to the independence of dimensions, can be performed by autonomous software agents, one for each dimension.[9]

Acknowledgements

I would like to thank the members of the Tilburg Dialogue Club who have over the years been involved in the development of DIT, in particular Volha Petukhova, Jeroen Geertzen, Hans van Dam, Ielka van der Sluis, Simon Keizer, Roser Morante, and Amanda Schiffrin.

References

Alexandersson, J., B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz & M. Siegel (1998). *Dialogue acts in VERBMOBIL-2 (second edition).* Verbmobil Report 226. Saarbrücken: DFKI.

Allen, J., L. Schubert, G. Ferguson, P. Heeman, C.H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, D. Traum (1994) The TRAINS project: A case study in defining a conversational planning agent. Technical Report 532, Computer Science Department, University of Rochester.

Allen, J. & M. Core (1997) DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1). Technical Report, Multiparty Discourse Group, Discourse Resource Initiative, September/October 1997.

Allwood, J. (1992). *On dialogue cohesion.* Gothenburg Papers in Theoretical Linguistics 65. Gothenburg Unviersity, Department of Linguistics.

Allwood, J. (1984). Obligations and options in dialogue. *THINK Quarterly* 3 (1), 9–18.

Allwood, J. 2000. An activity-based approach to pragmatics. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue*, John Benjamins, Amsterdam, pp. 47–80.

Bunt, H. (2000). *Dialogue pragmatics and context specification*. H. Bunt and W. Black (eds) Abduction, Belief and Context in Dialogue. Amsterdam: Benjamins, 81-150.

Bunt, H. (2005). *A Framework for Dialogue Act Specification*. ISO-SIGSEM workshop, Tilburg. `http://let.uvt.nl/research/TI/sigsem/wg/meeting4/fdas-orig.pdf`

Bunt, H. 2006. Dimensions in dialogue annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

Bunt, H. (2009). *The DIT^{++} taxonomy for functional dialogue markup*. D. Heylen, C. Pelachaud, R. Catizone and D. Traum (eds.) *Proc. AMAAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts'*, Budapest, May 2009.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46.

Geertzen, J. (2009) The automatic recognition and prediction of dialogue acts. Ph.D. Thesis, Tilburg Universsity, February 2009.

Geertzen, J., V. Petukhova, and H. Bunt. 2007. A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification. In: *Proc. SIGDIAL 2007*, Antwerp, pp. 140–149.

ISO (2009) *Semantic annotation framework, Part 2: Dialogue acts.* ISO TC 37/SC 4/N442 rev03, ISO, Geneva, April 2009.

Keizer, S. and H. Bunt (2006) Multidimensional dialogue management. In Proc. *SIGDIAL 2006*, Sydney.

Keizer, S. and H. Bunt (2007) Evaluating combinations of dialogue acts. In *Proc. SIGDIAL 2007*, Antwerp.

Nöth, E., A. Batliner, V. Warnke, J.-P. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. (2002) On the use of prosody in automatic dialogue understanding. *Speech Communication*, 36(1-2):45–62.

Petukhova, V. and H. Bunt (2009a) *Dimensions of communication* Technical Report 2009-003, Tilburg centre for Creative Computing, Tilburg University.

Petukhova, V. and H. Bunt (2009b) The independence of dimensions in multidimensional dialogue act annotation. In *Proceedings NAACL HLT Conference*, Boulder, Colorado, June 2009.

Shriberg, E., R. Bates, A. Stolcke, P., Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):439–487.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Traum, D., Larsson, S. (2003). The information state approach to dialogue management. In J. van Kuppevelt & R. Smith 9eds.) *Current and New Directions in Discourse and Dialogue*, pp. 325-345. Kluwer, Dordrecht.

[9] This view also underlies the PARADIME dialogue manager (Keizer & Bunt, 2006; 2007).

Animal communication - bluffing, lying, impressing, and sometimes even information

Sverre Sjölander
Linköping University
Linköping, Sweden
svesj@ifm.liu.se

Abstract

The purpose of transmitting information in the animal world is to gain some kind of advantage for the sender, or to evade unpleasant consequences. Evolution has led to a kind of arms race, where the sender tries to give as favourable an effect as possible, whereas the receiver tries to see through the bluffing. It is only in birds and mammals that we see an awareness of the meaning of the message - it is mostly produced by innate mechanisms - but in the great apes we find intentional lies and bluffings. The similarities to human non-verbal communication are obvious.

The expanding role of prosody in speech communication technology

Nick Campbell
TCD
Dublin, Ireland
nick@tcd.ie

Abstract

Speech communication is a uniquely human attribute that plays a multi-faceted role in human social interaction. At its core from one point of view lies language and linguistic structure, yet from a more fundamental point of view we find 'prosody' underlying many levels of speech communication, serving to signal not just linguistic but also interpersonal and social information.

Early humans would have had recourse primarily to tone-of-voice for basic communication but as language use became more sophisticated over evolutionary time this medium of human interaction became subsidiary to more sophisticated elements of communication, though its use did not disappear entirely.

In the development of technology for processing human speech, the linguistic element has long been considered prime. This talk will focus, however, on the 'tone-of-voice' aspects of prosody in social interaction, tracing their development in technological research from a carrier of linguistic information, signalling semantic and syntactic structure, to that of a social indicator, signalling affective and interpersonal cues that are equally essential to effective communication in a social situation.

By thus unravelling the role of prosody in speech, we will trace its uses from higher to lower levels of sophistication, and suggest some aspects of prosodic interpretation that might enable a technology for the processing of interpersonal states and attitudes in addition to and alongside the processing of propositional content in the speech signal.

Who’s next? Speaker-selection mechanisms in multiparty dialogue

Volha Petukhova
Tilburg Center for Creative Computing
Tilburg, the Netherlands
v.petukhova@uvt.nl

Harry Bunt
Tilburg Center for Creative Computing
Tilburg, the Netherlands
harry.bunt@uvt.nl

Abstract

Participants in conversations have a wide range of verbal and nonverbal expressions at their disposal to signal their intention to occupy the speaker role. This paper addresses two main questions: (1) How do dialogue participants signal their intention to have the next turn, and (2) What aspects of a participant’s behaviour are perceived as signals to determine who should be the next speaker? Our observations show that verbal signals, gaze redirection, lips movements, and posture shifts can be reliably used to signal turn behaviour. Other cues, e.g. head movements, should be used in combination with other signs in order to be successfully interpreted as turn-obtaining acts.

1 Introduction

Turn management is an essential aspect of any interactive conversation and involves highly complex mechanisms and phenomena. Allwood (2000) defines turn management as the distribution of the right to occupy the sender role. People do not start or stop talking just anywhere, and not without a reason. The decision to take the next turn or to offer the next turn to the partner(s) depends on the speaker’s needs, motivations and beliefs, and on the rights and obligations in a conversational situation.

In the widely quoted study of Sacks, Schegloff and Jefferson (Sacks et al., 1974) a model for the organisation of turn-taking in informal conversations has been proposed. The authors observed that conversations most often proceed fluently, that mostly one conversational partner talked at a time, that occurrences of more than one speaker at a time were brief, and that transitions from one turn to the next without a gap or overlap were very common. They reasoned that there must be an underlying system of turn-taking involved in conversations. They posited that during a conversation there are natural moments to end a turn and initiate a new one, called Transition Relevance Places (TRPs), and formulated the following rules:

- If the current speaker (S) selects the next speaker (N) in the current turn, S is expected to stop speaking, and N to speak next.
- If S’s behaviour does not select the next speaker, then any other participant may self-select. Whoever speaks first gets the floor.
- If no speaker self-selects, S may continue.

The generality of these rules makes them explanatory and applicable in many situations, but prevents them from being specific about the characteristics of speaker-selection techniques. At least two questions remain: (1) Which perceived behavioural aspects are used by people to estimate the locations of TRPs, and (2) Which aspects of communicative behaviour serve as signals to determine who is a potential or intended speaker of the next turn.

With respect to the first question, recent years have seen a number of solid qualitative and quantitative findings. It was observed that many turn transitions happen without temporal delays because a potential next speaker knows when a turn ends. People are able to predict turn endings with high accuracy using semantic, syntactic, pragmatic, prosodic and visual features (Ford & Thompson, 1996; Grosjean & Hirt,1996; De Ruiter et al., 2006; Barkhuysen et al., 2008, among others).

While end-of-turn prediction has been studied extensively, little research has been done on the prediction who is a potential next speaker, and on next speaker self-selection behaviour. This is in particular important if we deal with more than two participants in dialogue. Dialogue participants may just start speaking if they want to say something, but they often signal their willingness or readiness to say something. In other words, they perform certain actions to take the turn over. Speakers may signal that they want to have the turn when it is available (*turn taking*); that they want and are ready to have the turn when it is given to them by the previous speaker (*turn accepting*); and that they want to have the turn despite the fact it is not available (*turn grabbing*).

In this study we focus on the properties of a speaker's utterances that correlate with his turn-obtaining efforts in multi-party dialogue. Correlation indicates that two variables are related, but does not measure cause. It does not mean that signs which are correlated with turn-obtaining efforts are interpreted as such by communicative partners. To investigate this issue, we also looked if speaker changes really occur shortly after certain signals have been sent. We should also take into account, however, that a participant's wish to have the turn may be overlooked or ignored by others for some reason, and that he does not get the opportunity to speak. Therefore, to obtain more certainty about utterance properties related to turn taking, we performed perception experiments where subjects judged the participant's turn-taking efforts.

Before discussing our analysis and findings we first introduce a few concepts and terms for the rest of this paper. The term 'turn' is used in the literature in two senses: (1) as in 'to have the turn', i.e., to occupy the speaker role; and (2) to refer to a stretch of communicative behaviour produced by one speaker, bounded by periods of inactivity of that speaker or by activity of another speaker. Turns in this sense are sometimes called 'utterances' (cf. Allwood, 2000). We will use the term 'turn' in this paper in both senses, in such a way that no confusion is likely to arise. A turn in the latter sense may contain several smaller meaningful parts, most often called 'utterances'; these units are linguistically defined stretches of communicative behaviour. In natural spoken dialogue, the stretches of communicative behaviour that have a communicative function do not always coincide with turns or utterances, since they may be discontinuous due to the occurrence of filled and unfilled pauses, self-corrections, restarts, and so on; and they may spread over multiple turns, when the speaker provides complex information which he divides into parts in order not to overload the addressee. The notion of *functional segment* was therefore introduced, defined as the smallest (possibly discontinuous) stretch of communicative behaviour that has a communicative function (and possibly more than one) (Geertzen et al., 2007). The notion of functional segment is especially useful when analysing the turn-taking behaviour of participants in dialogue because it allows multiple functional segments that are associated with a specific utterance or turn to be identified more accurately.

The rest of this paper is organized as follows. After introducing the corpus and its annotation in Section 2, we discuss our observations concerning the turn-taking behaviour of dialogue participants. Section 3 describes perception experiments, and reports on the recognition of a participant's behaviour as a turn-management signal. Conclusions are drawn in Section 4.

2 Observation study

2.1 Corpus material and annotations

In this study we used human-human multi-party interactions in English (AMI-meetings).[1] The *AMI corpus* contains manually produced orthographic transcriptions for each individual speaker, including word-level timings. Two scenario-based[2] meetings were selected with a total duration of 51 minutes, constituting a corpus of 2,396 functional segments which contain either verbal components, nonverbal components, or both. All four participants were English native speakers.

The nonverbal behaviour of the dialogue participants was transcribed using video recordings for each individual participant, running them without sound to eliminate the influence of what was said. This transcription includes gaze direction; head movements; hand and arm gestures; eyebrow, eyes and lips movements; and posture shifts. Transcribers were asked to annotate low-level features such as form of movement (e.g. head: nod, shake, jerk); hands: pointing, shoulder-shrug, etc.[3]; eyes:

[1] Augmented Multi-party Interaction (`http://www.amiproject.org/`).

[2] Meeting participants play different roles in a fictitious design team that takes a new project from kick-off to completion over the course of a day.

[3] Hand gesture transcription was performed according to Gut,U., Looks, K., Thies, A., and Gibbon, D. (2003). CoGesT: Conversational Gesture Transcription System. Version 1.0. Technical report. Bielefeld University `http://www.spectrum.uni-bielefeld.de/modelex/publication/techdoc/cogest/`

Speaker		Observed communicative behaviour						
D	words	What's	teletext					
	gaze	averted(table)	personA	personB				
	eyes		narrow					
	posture	working position						
annotation	Feedback	neg. understanding						
	TurnM.	Turn assign to A						
B	words				um	It's	British	thing
	gaze	averted(table)	personD	personA	personD			
	eyes			widen				
	lips			random movements				
	posture	bowing	working position					
annotation	Feedback		pos. attention					
	TurnM.			turn take	turn keep			

Figure 1: *Transcription and annotation example.*

narrow, widen; lips: pout, compress, purse, flatten, (half)open, random moves); direction (up, down, left, right, backward, forward); trajectory (e.g. line, circle, arch); size (e.g. large, small, medium, extra large); speed (slow, medium, fast); and repetitions (up to 20 times). The floor transfer offset (FTO: the difference between the time that a turn starts and the moment the previous turn ends) and duration of a movement (in milliseconds) were computed. At this stage no meaning was assigned to movements.

For each token in verbal segments prosodic features were computed. Prosodic features that are included are pause before the token, minimum, maximum, mean, and standard deviation of pitch (F0 in Hz), energy (RMS), voicing (fraction of locally unvoiced frames and number of voice breaks), speaking rate (number of syllables per second) and duration of the token. We examined both raw and normalized versions of these features[4]. For each verbal segment FTO, duration and word occurrence[5] features were computed.

Speech and nonverbal signs were annotated with the DIT^{++} tagset[6] using the ANVIL tool[7]. Utterances were segmented per dimension according to the approach presented in (Geertzen et al., 2007). For turn management DIT^{++} distinguishes between turn-obtaining acts (turn-initial acts) and acts for keeping the turn or giving it away (utterance-final acts). A turn-initial function indicates whether the speaker of this turn obtains the speaker role by grabbing it (*turn grab*), by taking it when it is available, (*turn take*) or by accepting the addressee's assignment of the speaker role to him (*turn accept*). A turn ends either because the current speaker assigns the speaker role to the addressee (*turn assign*), or because he offers the speaker role without putting any pressure on the addressee to take the turn (*turn release*). A turn may also have smaller units with boundaries where a reallocation of the speaker role might have occurred, but does not occur because the speaker indicates that he wants to keep the turn. Such a segment has a *turn keep* function. A segment was labelled as having a turn-management function only if the speaker performed actions for the purpose of managing the allocation of the speaker role. For example, a segment was annotated as having the function Turn Take only if the speaker performs a separate act to that effect. If the speaker just goes ahead and makes a contribution to the dialogue, without first signalling his intention to do so, then the segment was not marked with a Turn Management function. 412 segments were identified having a turn-initial function (17.2%) and 370 segments as having one of the turn final functions (15.4%). Figure 1 provides an example from the annotated corpus.

We examined agreement between annotators in identifying and labelling turn management segments using Cohen's kappa measure (Cohen, 1960)[8]. Two annotators who were experienced in

[4] Speaker-normalized features were obtained by computing z-scores (z = (X-mean)/standard deviation) for the feature, where mean and standard deviation were calculated from all functional segments produced by the same speaker in the dialogues. We also used normalizations by the first speaker turn and by prior speaker turn.

[5] Word occurrence is represented by a bag-of-words vector (1,640 entries) indicating the presence or absence of words in the segment.

[6] For more information about the tagset, please visit: `http://dit.uvt.nl/`

[7] For more information about the tool visit: `http://www.dfki.de/~kipp/anvil`

[8] This measure of agreement takes expected agreement into account and is often interpreted as follows: 0=none; 0-0.2=small; 0.2-0.4=fair; 0.4-0.6=moderate; 0.6-0.8=substantial; and 0.8-1.0=almost perfect.

annotating dialogue and were thoroughly familiar with the tagset reached substantial agreement (kappa = .76) in identifying turn segments and assigning turn-management functions.

2.2 Results

It was observed from the annotated data that meeting participants often indicate explicitly when they wish to occupy a sender role. More than half of all speaker turns were preceded by attempts to gain the turn, either verbally or nonverbally (59%). 17.2% of all functional segments were found to have one of the turn-initial functions: 12% are turn-taking segments, 4.4% have a turn-grabbing function and 0.8% are turn accepts. Consider the following examples:

(1) *B: What* ***you guys*** *received?* (Turn Release)
A1: 0.54 ***Um****(0.65)* (Turn Take) [9]
A2: I just got the project announcement

(2) *B1: yeah brightness and contrast*
D1: - 0.35 ***Well****0.19* (Turn Grab)
D2: 0.11 what we're doing is we're characterizing

(3) *B1: That something we'd want to include*
B2: do ***you****(participant D is gazed) think?* (Turn Assign)
D1: 1.82 ***Uh****(1.39)* (Turn Accept)
D2: Sure

The reasons to take the turn may be various. First, a participant may have reasons to believe that he was selected for the next turn by the previous speaker. This puts a certain pressure on him to either accept the turn or signal its refusal. Second, a dialogue participant may want to make a contribution to the dialogue and believe that the turn is available. Finally, a dialogue participant may wish to have the turn while believing that it is not available, because (1) he has a desire to express his opinion urgently; or (2) he wants to gain control over the situation, e.g. when the meeting chairman needs to get a grip on the interactive process; or (3) he notices that the current speaker is experiencing difficulties in expressing himself, and e.g. assists in completing the utterance; or (4) he wants to express his appreciation of an idea or suggestion put forward by another participant; or (5) he failed to process the previous utterance of another participant and needs immediate clarification; or (6) he expects the current speaker to finish his utterance, and wishes to be the next speaker before the partner completes his turn.

Verbally, turn-taking intentions were mainly expressed by the following tokens: *um* and its combinations such as *um okay*, *um alright*, *um well* and *um yeah* (11.5% of all turn-initial segments); *so* (5%); *and* and combinations like *and so, well and*, also by *um and, uh and, and um, and uh* (7.9%); *well* (5.8%); *right* and combinations like *right so* and *right well* (7%); *uh* (5.6%); *okay* and *mm-hmm/uh-uhu* (5%); *alright* (2.8%); *yeah* or its repetition (15.7%); *but* (2%); *just* (1.2%); and repetitive expressions (e.g. *I.. I.. I.. would like*) (1.5%).

The majority of these tokens may serve several communicative functions is dialogue. For example, *'um'* and *'uh'* are known to be used as fillers to stall for time and keep a turn. Moreover, these tokens also occur in segments which are not related to turn management. For example, *'okay'* can be used as positive feedback or to express agreement. They also can be multifunctional expressing, for example, positive feedback and turn taking simultaneously. Previous studies, e.g. (Hockey, 1993) and (Gravano et al., 2007), confirmed that the use of these cue phrases can be disambiguated in terms of position in the intonation phrase and analysis of pitch contour.

We observed significant mean differences between turn-initial use and non-turn-initial use of these tokens in terms of duration (turn-initial tokens being more than 115 ms longer); mean pitch (turn takings have $>$ 12Hz); standard deviation in pitch ($>$ 5Hz); and voicing (5% more voiced). As for temporal properties of verbal turn-initial functional segments, it was observed that the floor transfer offset (FTO) is between -699 and 1030 ms, where negative value means overlap and positive a gap between successive turns. Turn-grabbing acts have an FTO from -699 to -166ms; turn-accepting acts may also slightly overlap the previous segment and have FTO from -80ms to 136ms; turn-taking acts the longest FTO have (between 582 to 1030ms).

To assess the importance of nonverbal signs for identifying turn-initial segments, we conducted a series of correlation tests using the phi-coefficient. The phi measure is used to test the relatedness of categorical variables, and is similar to the correlation coefficient in its interpretation. Table 1 shows the correlation between segments annotated

[9] Here and elsewhere in the text figures given between brackets in examples indicate token duration in seconds; figures without brackets indicate silences between tokens in seconds.

(Non-)verbal signal	ϕ
wording (presence of tokens listed above)	.47*
any gaze redirection direct-averted direct(>1 person)-averted	.79* .42* .61*
head movement	.05
hand/arm movement	.01
eye shape change + eyebrow movement	.15
any lips movement half-open mouth random lips movements	.59* .39* .28*
posture shift working position-leaning backward/forward	.87* .29*

Table 1: Nonverbal signals correlated to turn-initial segmets (* significant according to two-sided t-test, $< .05$)

as having a turn-initial function and accompanying nonverbal signals.

Strong positive correlations were observed for gaze aversion, lip movements and posture shifts. Especially in multi-party conversations gaze plays a significant role in managing fluent turn transitions than in two-person dialogues, because of the increased uncertainty about who will be the next speaker. As for gaze patterns that accompany turn-initial segments, in 29.4% of the cases the participant has direct eye contact with his addressee. In 11.8% of the cases the participants who want to have the next turn gazes at more than one of the partners, most probably verifying their intention concerning the next turn. A dialogue participant who aims for the next turn first gazes at one or more partners, and averts his gaze shortly before starting to speak (44.1%).

Comparable patterns were observed in previous studies. A speaker usually breaks mutual gaze while speaking and returns gaze to the addressee upon turn completion (Kendon, 1967). Goodwin in (1981) claims that the speaker looks away at the beginning of turns and looks towards the listeners at the end of the turn. More recently, Novick (1996) found that 42% of the turn exchanges follows a pattern in which the speaker looks toward the listener while completing the turn. After a short moment of mutual gaze the listener averts his gaze and begins the next turn.

Independent from the possible meanings of specific types of head movements, and from their feedback functions, head movements are used for turn management purposes. It was noticed in (Hadar et al., 1984) that speakers use head movements to mark syntactic boundaries and to regulate the turn-taking process. In our data the intention to have the next turn was successfully signalled by repetitive short head movements (34.3%). In 11.8% of the cases turn-initial efforts were signalled by waggles (head movement back and forth and left to right) and often indicated negative feedback or uncertainty. In 3.9% of the cases headshakes as signals of disagreement were observed. Interestingly, however, head movements do not correlate significantly with turn-initial acts. By contrast, a combination of spoken signals like ‘okay’ or repetition of ‘yeah’ and multiple head nods are good signals of a participant’s turn-obtaining intention (ϕ=.41, p =.003). This is in accordance with Jefferson’s findings that people proceed from ‘mm-hmm’ to ‘yeah’ when they want to have the turn (Jefferson, 1985).

Hand and arm gestures that may be related to the participant’s intention to have the turn were not observed frequently. We identified some shoulder shrugs that signalled uncertainty (3.5%) accompanied by head waggles and hand movements when a participant listening to the speaking partner suddenly moves his hand/fist away from the mouth (2%) or makes an abrupt hand gesture for acquiring attention (3.9%).

To signal the intention to have the next turn, participants frequently made random silent lip movements, compressing, biting, licking, or pouting their lips (10.9%). They also often keep their mouth (half-) open (47.3%). In 16.4% they narrow (possible sign of negative feedback) or widen (indicating surprise) their eyes accompanied by lowering or raising eyebrows, respectively.

Various types of upper-body posture shifts were often used as turn-initial signals (25.5%). Participants would change their body orientation from working position (both hands on the table, leaning slightly forward, head turned to the speaker) to leaning forward, backward or aside (17.6%), producing random shifts (shifting one’s weight in a chair) in 2%, shifting from bowing position (bending, curling, or curving the upper body, usually while writing) (5.9%). Cassell et al. in (2001) looked at posture shifts at turn boundaries and discourse segment boundaries, and showed that both boundaries had an influence on posture shifts. Posture shifts with the upper body were found more frequently at the start of a turn than in the middle or end (48%, 36%, and 18% respectively).

Generally, dialogue participants recognize an intention to take the turn successfully. In 60.8% of all the cases turn-obtaining efforts were acknowledged and the partner's wish to have the turn was satisfied. Participants who used more than one turn-initial signal or two modalities (e.g. combining head movements and posture shifts, or verbal and nonverbal signs) were more successful in obtaining the next turn. As for the remaining 39.2% it is difficult to judge whether the turn-taking efforts were interpreted as such by partners and ignored, or whether the signals were overlooked. Looking closer at gaze behaviour of meeting participants, our intuition is that in the majority of cases (65.2%) the turn-gaining efforts were most probably overlooked, because the participant was not gazed at by other partners. In another 34.8% of the cases, the participant's turn-gaining efforts were most likely ignored, since the partners did have direct eye contact. Nonetheless, since our analysis is based on the interpretation of annotators, this intuition could be wrong. To deal with this problem, perception experiments were performed which are reported in the next section.

3 Perception study

3.1 Stimuli and procedure

Two series of perception experiments were designed to study whether naive subjects interpreted certain behaviour of meeting participants as signals to have the next turn. From the annotated data we randomly selected 167 video clips with 4 different speakers (2 male, 2 female). Two referees judged the clips assigning them to the following categories:

1. a turn-initiating act is performed when the next turn is available;
2. a turn-initiating act is performed when the next turn was assigned to this participant;
3. a turn-initiating act is performed when the turn is not available but the participant needs: (a) to signal negative feedback on processing the partner's utterance; (b) to elaborate the partner's utterance; (c) to address the partner's suggestion; (d) to clarify the partner's utterance; or (e) to shift the topic;
4. no turn-taking act is performed.

The judges reached a substantial agreement on this task (kappa scores of .67). 52 stimuli, on which the judges fully agreed, were selected for further experiments: 4 of category 1; 4 of category 2; 36 of category 3; and 8 of category 4. The duration of each clip was about 10 seconds, containing the full turn of the previous speaker, and the recordings of the participant's movements and pause after the turn (if any) till the next turn starts. The subjects had 10 seconds to react to each stimulus. They were given the task to answer the question whether they think that a participant in question is performing any turn-initial act or not.

15 subjects (4 male and 11 female, all between the ages of 20 and 40) participated in one of the two sets of experiments: 9 subjects were asked to evaluate the video fragments without sound and 6 subjects evaluated the same fragments which were provided with sound. They were allowed to watch every video as many times as they liked.

	without sound	**with sound**
turn take	.31	.65
turn accept	.20	.55
turn grab	.32	.43
no turn-initial act	.79	1.00
overall	.48	.64

Table 2: Cohen's kappa scores for each class label for two sets of rating experiments

3.2 Results

3.2.1 Subject rating

We examined inter-subject agreement using Cohen's kappa measure (Cohen, 1960). Table 2 shows kappa scores calculated for each individual condition, for two class labels and for two sets of ratings.

Subjects reached moderate agreement judging whether a meeting participant performed a turn-initial act or not if they could not hear what was said, relying only on their interpretation of the nonverbal information; they reached substantial agreement when they could hear what was said. Agreement is higher (.79 = substantial agreement when judging videos without sound, and 1.00 = perfect agreement when sound was available) when a participant does *not* display any turn-taking efforts. Among the turn-initial acts the turn grabbing which was performed to signal negative feedback on the previous speaker utterance (at the level of interpretation or of evaluation) has been evaluated with higher agreement than the others (.57, $t < .05$) under both condition, most probably because participants produce distinctive facial expressions characterized by changing an eye shape, eyebrow and lips movements, often accompanied by a head shake or waggle additionally to other signals. The lowest agreement was found rating the turn-accept efforts of dialogue participants. This can be explained by the fact that participants to whom the next turn is assigned do not necessary perform any extra (nonverbal) action to indicate that they wish to be the next speaker, so that the raters often judge the participant's behaviour as having no turn-management function if they cannot hear that the turn was actually assigned by the previous speaker. Raters who could hear what the other participants say reached higher agreement than judges to whom speech transcription was not available. Thus, context information, such as the previous speaker's turn, seems to be important for the perception of turn-taking behaviour, perhaps also because dialogue participants actually anticipate TRPs (Ruiter et al., 2006), which makes it easier to perceive speaker-selection actions and to interpret turn-obtaining intentions.

3.2.2 Recognition of turn-initial acts

In this section we describe nonverbal features which we think may be helpful for explaining why subjects interpreted a participant's behaviour as having a turn-obtaining function (or not). We examined the following features: (1) gaze (directed, averted and combination of those); (2) head movement, if any; (3) hand gesture, if any; (4) eyebrow movement, if any; (5) eye shape change, if any; (6) lips movement, if any; (7) posture shift, if any; and (8) some combinations of these features.

We conducted a series of statistical tests, similar to those described in Section 2.2, and measured for each class label the correlations between the proportion of subjects that chose each label and the features described above. Table 3 presents correlations for the conditions with and without sound.

We can conclude that nonverbal signals are important for recognizing speaker-selection intentions. A gaze pattern such as 'gazing at more than one person and then averting the gaze', and various types of lips movements and (half-)open mouth in particular, correlate positively with a turn-initial act and have strong negative correlation with non-turn-initial acts). Head nods, on the other hand, turn out not to be significant for turn-taking purposes, because they may be used to signal active listening without the intention to take the turn (e.g. so-called backchannels). A combination of head movements and other signals, by contrast, was perceived by judges as a turn-initial signal, e.g. a head movement accompanied by lips movements, or posture shifts and certain gaze pattern such as 'mutual gaze - averted' (the combination of all three has a strong positive correlation with turn-initial acts: .55, t $<$ .05). Thus, dialogue participants who use multiple signals or modalities are more successful in gaining the next turn. Conversational partners are also more likely to perceive and understand the partner's turn behaviour when relying on multiple information sources.

	ϕ **(without sound)**	ϕ **(with sound)**
turn-initial act		
gaze 'averted'	-.34*	-.44*
gaze 'direct(more persons)-averted'	.54*	.52*
head movement	.49	.25
head nods	.40	.28
hand gesture	.49	.21
eye shape change + eyebrow movements	.54*	.46*
(half-) mouth	.58*	.35*
lips movement	.44	.34*
posture shift	.41	.30*
'posture shift + head movement'	.34	.35*
'lips + head movements'	.57*	.39*
'eye shape change + head movements'	.47	.27
'eyebrow + head movements'	.46	.25
'gesture + head movements'	.44	.15
gaze 'direct-averted' + posture shift	.37	.34*
gaze 'direct-averted' + head movement	.55*	.40*
gaze 'direct-averted' + lips movements	.60*	.59*

Table 3: Features correlated with the proportion of votes for each class label (without/with sound ratings). * differs significantly from zero according to two-sided t-test, t $<$ .05

4 Conclusions and future work

In this study we were interested in identifying speaker-selection mechanisms in multiparty dialogue. The main aim was to determine which aspects of a participant's behaviour serve to signal the intention to have the next turn.

A range of verbal expressions may be used to signal the intention to have the next turn, including several types of fillers, discourse markers, repetitive expressions, and other vocal sounds.

We have found that gaze redirection is the most important nonverbal indicator of turn management in multiparty dialogue, although turn organisation cannot be explained completely by gaze behaviour. In general, a participant who wants to claim the next turn first looks at the other participants and averts his gaze shortly before starting to speak.

As for head movements, multiple head nods were found to be significantly correlated with turn-initial acts. The results of the perceptual study showed, however, that head nods are not interpreted as having a turn-initial function. By contrast, some combinations of head movements and other signals, either verbal ('okay' or 'yeah') or nonverbal (e.g. lips movements) are associated with turn-initial functional segments.

Concerning hand and arm gestures, no statistically significant results can be reported due to the low frequency of their occurrence in our data.

According to our data, facial expressions are used not only to express emotions, attitudes and states of cognitive processing, but also the intention to occupy the speaker role. Our observational and perceptual analyses show that lips movements and changes in eye shape correlate positively with turn-initial acts.

Posture shifts, finally, were frequently found at the start of a turn, and strongly correlate with turn-initial acts; they were perceived as a strong turn-initial cue on their own and in combination with other signals.

From our observational and perceptual studies it may be concluded that the combination of nonverbal signs and signals from several modalities (speech and movements) forms a reliable indicator of the intention to take the turn, and the dialogue participants who used these complex signals for the purpose to claim the next turn were successful in getting it.

This paper reports results from a limited number of dialogues and small-scale perceptual experiments, but the findings are promising. Future research will look into the performance of perceptual experiments with richer sets of stimuli, and use the results also for further observational analysis, since it is still very hard to obtain high-quality annotated data of nonverbal behaviour.

References

Jens Allwood. 2000. An activity-based approach to pragmatics. In: Harry Bunt and William Black, editors, *Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics*, John Benjamins, Amsterdam, The Netherlands, pp. 47–80.

Pashiera N. Barkhuysen, Emiel J. Krahmer, and Mark G.J. Swerts. 2008. The interplay between auditory and visual cues for end-of-utterance detection. *The Journal of the Acoustical Society of America*, 123(1):354–365.

Harry Bunt. 2006. Dimensions in dialogue annotation. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Justine Cassell, Yukiko I. Nakano, Timothy W Bickmore, Candace L. Sidner, and Charles Rich. 2001. Non-Verbal Cues for Discourse Structure. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20: 37–46.

Cecilia E. Ford and Sandra A. Thompson. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In: Emanuel A. Schegloff and Sandra A. Thompson, editors, *Interaction and grammar*, Cambridge: Cambridge University Press, pp. 135– 184.

Jeroen Geertzen, Volha Petukhova and Harry Bunt. 2007. A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification. In: *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, pp. 140–149.

Francois Grosjean and Cendrine Hirt. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes*, 11:107–134.

Charles Goodwin. 1981. Conversational Organization: Interaction between hearers and speakers. New York: Academic Press.

Agustin Gravano, Sefan Benus, Hector Chavez, Julia Hirschberg and Lauren Wilcox. 2007. On the role of context and prosody in the interpretation of 'okay'. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Uri Hadar, Timothy J.Steiner, Ewan C. Grant, and F. Clifford Rose. 1984. The timing of shifts of head postures during conversations. *Human Movement Science*, 3:237–245.

Beth Ann Hockey. 1993. Prosody and the role of okay and uh-huh in discourse. In: *Proceedings of the Eastern States Conference on Linguistics*, pp. 128-136.

Gail Jefferson. 1985. Notes on a systematic Deployment of the Acknowledgement tokens 'Yeah' and 'Mmhm'. *Papers in Linguistics*, 17(2): 197–216.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologia*, 26: 22–63.

David G. Novick, Brian Hansen, and Karen Ward. 1996. Coordinating Turn-taking with Gaze. In: *Proceedings of the International Symposium on Spoken Dialogue*, Philadelphia, PA, pp.53–56.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4): 696–735.

Jan Peter de Ruiter, Holger Mitterer, and Nick J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82: 515–535.

On cue — additive effects of turn-regulating phenomena in dialogue

Anna Hjalmarsson
Centre for Speech Technology, KTH
Stockholm, Sweden
annah@speech.kth.se

Abstract

One line of work on turn-taking in dialogue suggests that speakers react to “cues” or “signals” in the behaviour of the preceding speaker. This paper describes a perception experiment that investigates if such potential turn-taking cues affect the judgments made by non-participating listeners. The experiment was designed as a game where the task was to listen to dialogues and guess the outcome, whether there will be a speaker change or not, whenever the recording was halted. Human-human dialogues as well as dialogues where one of the human voices was replaced by a synthetic voice were used. The results show that simultaneous turn-regulating cues have a reinforcing effect on the listeners’ judgements. The more turn-holding cues, the faster the reaction time, suggesting that the subjects were more confident in their judgments. Moreover, the more cues, regardless if turn-holding or turn-yielding, the higher the agreement among subjects on the predicted outcome. For the re-synthesized voice, responses were made significantly slower; however, the judgments show that the turn-taking cues were interpreted as having similar functions as for the original human voice.

1 Introduction

This paper describes a perception experiment that investigates the probabilities of who will be the next speaker based on potential “cues“ in the behaviour of the previous speaker. The experiment was designed as a game where the subjects were asked to listen to two-party dialogues and, whenever the recording halted, guess who would be the next speaker. The aim is to investigate if combinations of simultaneous cues affect the confidence of listeners’ judgments. The results also have implications for spoken dialogue system (SDS) research; If SDS can signal turn completion or non-completion in a way that can be easily discriminated by humans, turn-transitions in such systems could be made more intuitive. Thus, a secondary aim of this study is to test if the cues can be reproduced in a synthetic voice. Both human-human dialogues and dialogues where one of the human voices was replaced with a synthesized voice were tested.

1.1 Incremental language processing

Spoken dialogue systems that opt for human-likeness (Edlund et al., 2008) should be flexible and allow their users to hesitate and revise their speech in a way that is similar to interacting with a human dialogue partner. However, turn management in current SDS is in general not very sophisticated. One frequent strategy is to interpret long silences, above a certain threshold (Ferrer et al., 2002), as end of user turn. Thus, the system still risks barging in over its users because of the large variance in silence duration for spontaneous speech (Campione & Veronis, 2002). Faster processing of input only partly solves the problem, since the response delay due to end of turn detection is still not targeted. In fact, perceiving, planning and producing speech is time consuming for humans too, but we have strategies to avoid long ambiguous silences. First, we start to plan new contributions before the other person has stopped speaking. When starting to speak, we typically do not have a complete plan of what to say but yet we mange to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions.

If behavioural cues related to these human strategies can be identified, we can employ similar methods in SDS. The objective is to indicate to the user that the system plans to continue speaking and by doing this avoid user confusion regarding whether the ongoing system utterance is complete or not. In a similar fashion, the system also needs strategies to efficiently signal and detect end of turns.

1.2 Turn taking in spoken dialogue

Humans are expected to produce new dialogue contributions within a certain time. Then again, speech is not generated in regular constant pace of vocalized segments, but in streams of fragments in varying sizes (Butterworth, 1975). In addition, spontaneous dialogue involves unexpected interruptions or disfluencies such as pauses, corrections and repetitions that we use to refine, alter, and revise our plans as we speak (Clark & Wasow, 1998). Despite its irregularities, we only talk simultaneously for brief periods of time (Schegloff, 2000). Sacks et al. (1974) suggest that this is viable because humans have a mutual understanding of transition relevance places (TRPs). A frequent assumption is that humans can predict these TRPs almost exactly and that a majority of speaker shifts are directly adjoining without any overlap or silence. If this is true, interlocutors are able to predict approaching end of turns in advance very precisely. These TRPs are claimed to be detected in terms of expected end points of semantic or lexical units (e.g. de Ruiter et al., 2006). Yet, analysis of turn transitions in American English, German and Japanese have shown that pauses and overlaps are normally (Gaussian) distributed (Weilhammer & Rabold, 2003), suggesting that perfectly adjoining transitions are rare.

1.3 Turn management signals

An early series of works on turn-taking (cf. Duncan, 1972; Duncan & Fiske, 1977) suggest that interlocutors react to a set of signals employed by the previous speaker to indicate approaching turn endings. According to Duncan (1972 p.283): “The proposed turn-taking mechanism is mediated through signals composed of clear-cut behavioural cues, considered to be perceived as discrete”. Analysis of dialogues showed that the number of available turn-yielding cues was linearly correlated with listeners’ turn taking attempts. However, if speakers’ employed signals to suppress such attempts the number of turn-taking attempts radically decreased, regardless of the number of turn-yielding signals.

Cues relevant for turn-taking

Turn-holding cues are those referred to as attempt-suppressing signals by Duncan. This type of cue indicates that the speaker intends to hold the turn. Turn-holding cues reported by Duncan include drawl on the final syllable (phrase-final lengthening), an intermediate pitch level and sociocentric sequences (stereotyped lexical expressions or cue phrases). Turn-yielding cues reported include rising or falling pitch, the termination of a hand gesture, a drop in loudness and completion of grammatical pauses. Recent work by Gravano (2009) presents a number of phenomena found to take place at significantly higher frequencies before speaker switches. These cues include a falling or high-rising intonation, a reduced lengthening, a lower intensity level, a lower pitch level, points of textual completion, a higher frequency of jitter, shimmer and noise-to-harmonics ratio and longer inter-pausal unit duration. Moreover, in line with Duncan’s findings, Gravano’s show strong support for a linear relationship (positive correlation) between the number of simultaneously available turn-yielding cues and the number of turn-taking attempts. In line with Gravano and Duncan, this work further investigates how discrete cues form a complex signal that guides interlocutors’ turn-taking behaviour in dialogue.

Duncan and followers have mainly focused on describing correlates of actual turn-taking behaviour. Nonetheless, there is a range of acceptable behaviours; some may be perceived as impolite, yet effective if speakers get their points across. Consequently, interlocutors have the choice to act “hazardously” and defy the “principles” of turn-taking. For example: speakers can choose to barge in at less suitable places and avoid taking the floor when expected to. Bearing this in mind, in this experiment we explore the probabilities of different outcomes regardless of the outcome of the original dialogue. Schaffer (1983) and Oliveira & Freitas (2008) approached turn-taking issues from a similar perspective, i.e. analyzing the judgments of non-participating listeners in perceptual experiments. However, while their approach was to isolate the signals presented to the subject, our approach is to label the cues separately and subsequently study their combined effect on listeners’ judgments in a context that is similar to listening in on someone else’s conversation. As pointed out by Oliveira & Freitas (2008), analyzing dialogues outside of their contexts is problematic; yet, by allowing the subjects in our study to follow dialogues incrementally in chronological order rather than listening to disconnected phrases, we hope to overcome some of these problems.

It should also be mentioned that the outcome of turn-yielding signals is difficult to predict. Even if the previous speaker directs questions with the intention of eliciting a specific response or feedback from a listener, this participant may choose not to take the floor. Then again, often turn-yielding cues merely signal completion of a turn and leave the floor open. This means anyone may take the floor, including the previous speaker, whereas if a speaker signals turn holding intentions, the outcome is more predictable. From a SDS perspective, being able to suppress turn-taking attempts and discriminate users' internal pauses from turn completions is useful knowledge.

Duncan has been criticised for not reporting any inter-annotator agreement or formal description of his "signals" (Beattie et al., 1982). Whether the phenomena that Duncan refers to as "signals" should be considered as conscious or not is problematic. There are for example acoustic cues, e.g. drop in energy or inhalations that guide interlocutors in their turn-taking. However, a likely origin of these "signals" is the anatomy of our speech organs. If we plan to continue speaking, we keep the speech organs prepared and if we plan to finish, we release them (Local & Kelly, 1986). In this paper, all perceivable phenomena relevant for turn-taking are referred to as cues, regardless if they are conscious or not.

2 Dialogue data

The dialogues used as stimuli in this experiment were collected in order to obtain data in the DEAL domain. DEAL is a spoken dialogue system for conversation training for second language learners of Swedish under development at KTH. The scene of DEAL is set at a flea market where a talking animated agent is the owner of a shop selling used objects. The objectives are to build a system which is fun, human-like, and engaging to talk to, and which gives language learners conversation training (Hjalmarsson et al., 2007). The recorded dialogues are informal, human-human, face-to-face conversation in Swedish. The task and the recording environment were set up to mimic the DEAL domain and role-play. The corpus includes eight dialogues with six different speakers. All together about two hours of speech were collected. The dialogues were transcribed orthographically including non-lexical entities such as laughter, repetitions, filled pauses, lip-smacks, breathing and hawks. Two annotators labelled the data for cue phrases (CP) with high inter-annotator agreement (kappa 0.82) (Hjalmarsson (2008)). Cue phrases (also frequently referred to as discourse markers) are linguistic devices used to signal relations between different segments of speech. The cue phrases used here were phrases labelled to have either *response eliciting* or *additive* discourse pragmatic functions. Examples of response eliciting CPs are "eller hur" (right) and "då" (then) and examples of additive CPs are "och" (and), "eller" (or) and "men" (but). Though commonly not categorized as such, we also included filled pauses in this category. Response eliciting CPs were expected to have turn-yielding functions, while the additive CPs and the filled pauses were expected to have turn-holding functions.

The transcripts from four dialogues were also time-aligned with the speech signal. This was done using forced alignment with subsequent manual verification of the timings.

2.1 Manual labelling of cues

The perception experiment was designed to elicit probabilities of a speaker change versus a hold regardless of the outcome of dialogue, that is, without considering its actual continuation. The four dialogues in the corpus that had been time-aligned were automatically segmented into inter-pausal units (IPUs), a sequence of words surrounded by silence longer than 200 milliseconds (ms). According to Izdebski & Shipp (1978) humans need just under 200 ms to verbally react to some stimulus, which suggests that the speakers in the original dialogues had enough time to react to any potential cues in the end of previous IPU. For shorter silences or in overlapping speech it was impossible to halt the recordings without revealing to the subjects who the next speaker was. The four dialogues contained 2011 such silences, of which 85% were internal pauses and 15% were silences between speakers. Henceforth, silences within speaker turns will be referred to as *pauses* while silence between speakers will be referred to as *gaps*. The terminology is adopted from Heldner, M., & Edlund, J. (submitted).

To distinguish and explore all cues claimed to be relevant for turn-taking is beyond the scope of this paper. Since the focus is on the contributive effect of simultaneously occurring cues, the number of cues was restricted to five categories. The five categories were pitch contour, semantic completeness, phrase-final lengthening, non-lexical elements such as perceivable breathing and lip-smacks and some frequently occurring cue phrases (see Table 1). The dialogues recorded were face-to-face interactions that most likely contain visual turn-management cues such as hand and facial gestures. However, the visual gestures were not considered here and the labellers and subjects only had access to the audio recordings. The reason for this was to focus on the lexical and acoustic cues that can potentially be reproduced in a synthetic voice. Reported differences between face-to-face and telephone conversation are longer duration of silences in face-to-face interaction (Bosch et al., 2004). However, if Duncan's observations are correct, the more cues available, regardless of modality, the more predictable is the outcome.

Category	Turn-yielding cues	Turn-holding cues
Pitch contour	fall	flat
Final lengthening	no	long
Non-lexical	Audible expirations	Audible inhalations, lip-smacks
Cue phrases	response eliciting CPs	Additive CPs, filled pauses
Semantic completeness	complete	incomplete

Table 1 : Cue categories

Deciding what is a cue is problematic. To consider a parameter as a cue implies that its receiver perceives it or at least that it is perceivable by some other human in the same context. To tackle this problem we used two annotators for all parameters and only parameters that both annotators agreed upon were considered as cues. As follows, the absence of a cue does not necessarily entail its opposite, it simply means the labellers did not perceive the cue or that they did not agree on which category it belonged to. However, the cues are exhaustive and cannot contain yielding and holding functions in the same dimension. As discussed in Ward (2006), knowing where to look and how other prosodic features interact with the relevant cue is problematic. To focus on signals that are perceivable by humans in a dialogue context the labellers did not have any visual representations of the sound. Each labelling task included only the target parameter and no turn-taking issues were considered during labelling. The cues were labelled independently, one by one, in an attempt to avoid influences from other cues. Still, for the prosodic cues, other auditory cues could not be excluded from the recordings used for labelling.

Pitch slope

For pitch slope, the task was to label flat, rising or falling pitch contour. This roughly corresponds to ToBi labelling H-L% (plateau), H-H% (high-rise) and L-L% (falling pitch)[1]. The labellers were provided with only the last 500 ms of the IPU to avoid influences of the lexical context. Inter-annotator agreement for pitch slope was rather poor (kappa 0.36). However, a confusion matrix revealed that the majority of the confusions were between falling and rising slope. After listening to the data, a possible explanation is that a frequently occurring contour in the data was a rising curve with a minor slope at the end that labellers may have judged differently. This suggests that a more fine-grained labelling scheme could have been used. Still, as already mentioned, only stimuli where labellers agreed were considered to contain cues. Since the literature provides no clear-cut results of the effects of a rising pitch, which appears to contain both turn-yielding and turn-holding functions (Edlund & Heldner, 2005), this was not considered a cue.

Phrase-final lengthening

The labelling procedure for phrase-final lengthening was almost identical to the one of pitch slope except for the target labels, which were long, short and no phrase-final lengthening. Inter-annotator agreement for this task was also poor (kappa 0.37), however, the confusion matrix suggests that the annotators' boundaries were skewed, since almost all confusions were between neighbouring categories. Minor lengthening was not considered a cue.

Semantic completeness

Semantic completeness represents the lexical context of the dialogues. To extract syntactically complete phrases using part of speech tagging is not feasible since utterances in dialogue often violate syntactic rules and since dialogue relies much on context that is not captured by syntax. As an alternative, labellers were asked to decide whether the last utterance was pragmatically complete or not considering the previous context. The labelling was done incrementally from the orthographic transcriptions of the dialogues without listening to the recordings. Non-lexical elements such as filled pauses and breathing had been removed from the transcripts, since they are considered to represent acoustic information — information that is already represented in other cues. The label tool only displayed the left context of the dialogue up to the silence just after the target IPU. After each judgment, the dialogue

[1] ToBi is a standard for labelling English prosody (Silverman et al., 1992)

segment up to the next target pause was provided incrementally. Inter-annotator agreement was high for this task (kappa 0.73). The labelling procedure for semantic completeness is very similar to the procedure used by Gravano (2009).

2.2 Stimuli selection

The task in the experiment was to guess who the next speaker was whenever the dialogue play-back halted. To allow the subjects to get familiar with the dialogue context, i.e. getting a fair understanding of the left context, the dialogue segments could not be too short. At the same time, the test should include segments from more than one dialogue, with different speakers and still not be exhaustingly long. In the final test, segments from four different dialogues ranging from 116 to 166 seconds were used based on their richness in variety of cue types and variety in cue quantity. The four dialogues included three different speakers, one male and two female. The male speaker participated in all four dialogues. In a first pilot experiment, target IPUs, i.e. stimuli in the experiment, were randomly selected, which resulted in a stimuli set that were weighted neither for the number of cues nor for the distribution of gaps and pauses in the overall dialogue. For the final experiment, all IPUs were labelled with cues in advance. Target IPUs were then selected from a list with cue labels without listening to the recordings. The selections were made to get IPUs that represent a weighted distribution of gaps and pauses over speakers and a variety of cues. However, it was difficult to find segments in the data that fulfilled all requirements and a perfect weighted range was impossible to obtain because some combinations did not occur in the data and it is questionable whether these are very frequent in any type of dialogues. In the end, 128 IPUs were used as stimuli (see Table 2**).**

	Turn-yielding cues				
Turn-holding cues	**0**	**1**	**2**	**3**	**4**
0	8	18	17	4	1
1	15	10	2	1	
2	22	8			
3	14	3			
4	4				
5	1				

Table 2 : Cue distribution over stimuli IPUs

2.3 Re-synthesis of dialogues

One motivation for this work was to investigate whether the cues could be reproduced in a synthetic voice and perceived as having similar functions. In order to test this, one party in the dialogues was replaced with a diphone synthesis. The synthetic voice was reproduced with timings from the manually verified forced alignments and fundamental frequency automatically extracted from the human voice using Expros, a tool for experimentation with prosody in diphone voices (Gustafson & Edlund, 2008). Only the male party in the dialogues was re-synthesized, since we only had access to a male diphone synthesis. Since breathing and lip-smacks could not be re-synthesized, we kept the original human realizations from the recordings.

3 Method

The GUI of the test (see Figure 1) included two buttons with “pacmans” and a button where the subjects could pause the test. The pacmans represented the speakers in the dialogues and, when the corresponding interlocutor spoke, the pacman opened and closed its mouth repeatedly. The subjects’ task was to listen to recordings and, at each time when the recording halted, guess who the next speaker was by pressing the corresponding pacman button. The speakers in the dialogues were recorded on different channels and the movements of the face with the left position on the screen corresponded to the sound in the subject’s left ear, and vice versa. To make the subjects aware that the play-back had halted, both faces turned yellow. The subjects had 3 seconds to make the response or else the dialogue would continue. Each time the recording halted, the mouse pointer was reset to its original position, in the middle of the pause button. This was done to control the conditions before each judgment to enable comparisons between the trajectories of the subjects’ movements and their reaction times. The motivation was to track users’ mouse events and use these as a confidence measure similar to Zevin & Farmer (2008).

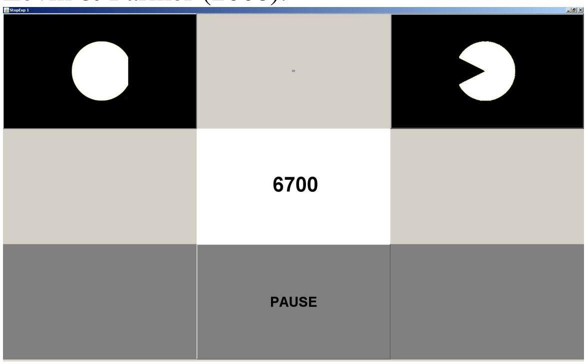


Figure 1 : Experiment GUI

The experimental setup was designed as a game where the subject received points based on whether they could guess the actual continuation of the dialogue. To elicit judgements based on first intuition rather than afterthought, speed was rewarded. The faster subjects responded, the fewer minus points they incurred when they were wrong and the more bonus points they received if they were right. Whether they made the right choice or not was not important, but it was used as an objective rewarding system to motivate the users. Who was considered the next speaker was based on which interlocutor vocalized first, regardless of whether this was a turn-yielding attempt or only short feedback responses (backchannels). Two movie tickets were awarded to the “best” player.

3.1 Pilot experiment

A pilot experiment with ten subjects was conducted to test the experimental setup and features of the GUI. The reset of mouse pointer before each response did not seem to affect the subjects noticeably. In fact, some of them even claimed that they had not noticed that the pointer moved. There were, however, obvious training effects; i.e. the response times were significantly faster at the end of the test. In the final experiment, training effects were controlled for by changing the order of the dialogues. There was also a 210 seconds long training session to allow the subjects to become familiar with the task.

3.2 Experiment

The final experiment included 16 subjects, 9 male and 7 female, between the ages of 27 and 49. All were native Swedish speakers except for two who had been in Sweden for more than 20 years. Five of the subjects were working at the department of Speech Music and Hearing, but the majority had no experience in speech processing or speech technology. Each subject listened to two human-human dialogues and two dialogues where one party was replaced with the diphone synthesis. The re-synthesized dialogues differed between subjects.

4 Results

It was difficult to find dialogue segments with an equal distribution of cue types and cue type combinations. All cues were considered as having equal weight and the relative contribution of the different cues was not considered. Some cue combinations were rare (Table 2) and since small variances in the data will affect the results for these cues, cue combinations represented in less than five IPUs were excluded. Moreover, since the results from the human–human condition and the human-synthesis condition appeared to be very similar, both conditions are included in the overall results presentation.

First, IPUs with a majority of turn-holding cues were judged significantly faster than IPUs with a majority of turn-yielding cues (t-test p<.05). However, as already discussed in Section 1.3, the outcome of turn-holding cues is more predictable. This is also confirmed by the overall distribution of pauses versus gaps (85% respectively 15%) and the extent to which the subjects agreed on the expected outcome for the different cue categories.

4.1 Reaction times

Reaction times can never be negative and the maximum value (3 seconds) was set generously, well above the time needed for most judgments (the geometric mean was 1166 ms). The distribution of reaction times is therefore skewed to the left. As suggested by Campione & Veronis (2002) the log-normal law is a better fit to duration data. Reaction times were therefore transformed into a logarithmic scale (base 10). Moreover, the average reaction times differed considerably between subjects (from 933 ms to 1510 ms). The reaction times were therefore also z-normalized for each subject. The reaction times for the judgments are a likely indication of how confident the subjects were in their decision. This was supported by the fact that stimuli with high agreement, regardless of cues, were judged significantly faster by subjects (Tukey’s test p<.05) (see Figure 2).

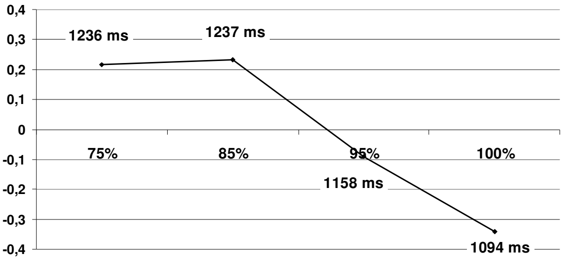


Figure 2 : Average reaction time $\log_{10}$ z-normalized over IPUs with % agreement.

For completeness, each point is labelled with its average $\log_{10}$ value (un-normalized) in milliseconds. All differences are significant, except for between 75% and 85% agreement.

The reaction times for stimuli with more turn-holding cues were significantly shorter (ANOVA p<.05, df=3). The differences are displayed in Table 3 (Tukey’s test p<.05). IPUs with contradictory cues, i.e. both turn-yielding and turn-holding cues, are not included here. Although not all steps differ significantly, there is a strong trend; the more turn-holding cues, the faster the reaction time.

Turn-holding cues		Difference in mean response time, *i* – *j* $\log_{10}$ z-value ($\log_{10}$ in ms)	Standard error	p-value
i	*j*			
0	1	0.141 (32.0 ms)	0.07	.382
	2	**0.363 (89.3 ms)**	**0.06**	**.000**
	3	**0.562 (138.5 ms)**	**0.08**	**.000**
1	2	0.222 (57.3 ms)	0.08	.079
	3	**0.420 (106.5 ms)**	**0.09**	**.000**
2	3	0.198 (49.2 ms)	0.09	.183

Table 3 : Differences in average response time between 0-1, 0-2, 0-3, 1-2, 1-3, 2-3 turn-holding cues (Tukey's p<.05, df=3). Significant differences in bold.

4.2 Synthesis versus natural

To present all cue combinations, including IPUs with both turn-yielding and turn-holding cues visually, three dimensional bubble charts will be used from now on. The charts display the number of turn-yielding cues on the x-axis and turn-holding cues on the y-axis.

Overall, reaction times for the synthetic voice are significantly longer (t-test p<.05). However, the reaction times decrease with an increased number of turn-holding cues in a very similar fashion as for the natural voice. This is illustrated in Figure 3. The width of the bubbles represents the z-normalized reaction times on a logarithmic scale. Unfilled bubbles represent the synthetic voice and black bubbles the human voice (the bubbles lay on top of each other). As in the overall data set (see Table 3), the reaction times for IPUs with more turn-holding cues were also significantly shorter for the synthetic voice (Tukey's p<.05).

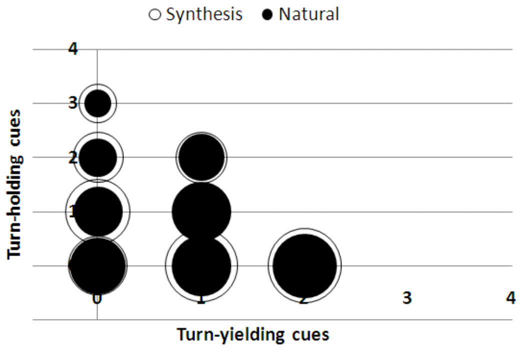


Figure 3 : Average reaction time $\log_{10}$ z-normalized for natural and synthetic voice

4.3 Agreement

The experiment can be viewed as a series of Bernoulli trials with dichotomous response, SWICH or HOLD. To study the effects of simultaneous cues on the actual judgments, binary stepwise logistic regression was used. The results show that there are significant relationships between turn-yielding cues and SWITCH and turn-holding cues and HOLD (p<.05). The diameters in the bubble charts in Figure 4 and Figure 5 represent % judgments for SWITCH versus HOLD for human and synthetic voice. The results show that cues are perceived as hypothesized.

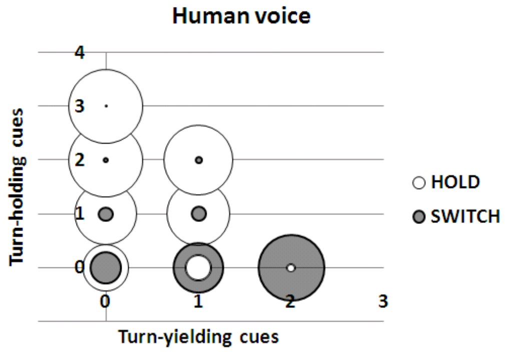


Figure 4 : The distribution of judgments for SWITCH versus HOLD for Human voice

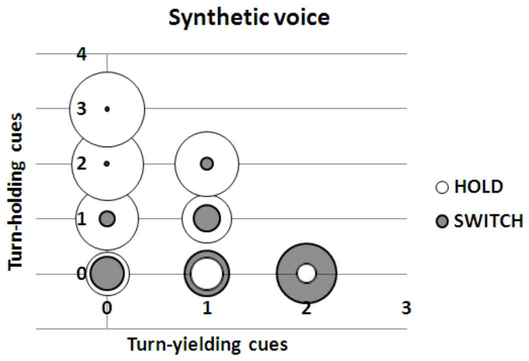


Figure 5 : The distribution of judgments for SWITCH versus HOLD for Synthetic voice

5 Final remarks

The results show that the turn-regulating cues are perceived as expected and in line with previous work. The novel contributions in this work include the reported reinforced effect of simultaneous lexical and non-lexical turn-regulating cues on non-participating listeners. Moreover, whereas previous research has focused on turn-yielding cues, we have also been able to present results that support a combined effect of turn-holding cues. Another important contribution is the results from re-synthesizing the human voice which suggests that these behavioural cues can be reproduced in a synthetic voice and perceived accordingly.

6 Acknowledgements

This research was carried out at Centre for Speech Technology, KTH. The research is also supported by the Swedish research council project #2007-6431, GENDIAL. Many thanks to Rolf Carlson, Jens Edlund, Joakim Gustafson, Mattias Heldner, Julia Hirschberg and Gabriel Skantze for help with labelling and valuable comments.

References

Beattie, G. W., Cutler, A., & Pearson, M. (1982). Why is Mrs. Thatcher interrupted so often?. *Nature, 300*(23), 744-747.

Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Proc. of the 7th International Conference TSD 2004* (pp. 563-570). Heidelberg: Springer-Verlag.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research, Volume 4* (Number 1).

Campione, E., & Veronis, J. (2002). A large-scale multilingual study of silent pause duration. In *ESCA-workshop on speech prosody* (pp. 199-202). Aix-en-Provence.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37*(3), 201-242.

Duncan, S., & Fiske, D. (1977). *Face-to-face interaction: Research, methods and theory.* Hillsdale, New Jersey, US: Lawrence Erlbaum Associates.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica, 62*(2-4), 215-226.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

Ferrer, L., Shriberg, E., & Stolcke, A. (2002). Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody. In *Proc. of ICSLP* (pp. 2061-2064).

Gravano, A. (2009). *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. Doctoral dissertation, Columbia University.

Gustafson, J., & Edlund, J. (2008). expros: a toolkit for exploratory experimentation with prosody in customized diphone voices. In *Proc. of PIT 2008, Kloster Irsee, Germany,* (pp. 293-296). Berlin/Heidelberg: Springer.

Heldner, M., & Edlund, J. (2008). *Pauses, gaps and overlaps in conversations.* Manuscript submitted for publication.

Hjalmarsson, A., Wik, P., & Brusk, J. (2007). Dealing with DEAL: a dialogue system for conversation training. In *Proc. of SigDial* (pp. 132-135). Antwerp, Belgium.

Hjalmarsson, A. (2008). Speaking without knowing what to say... or when to end. In *Proc. of SIGDial 2008*. Columbus, Ohio, USA.

Izdebski, K., & Shipp, T. (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research, 21*, 638-651.

Local, J., & Kelly, J. (1986). Projection and "silences": Notes on phonetic and conversational structure. *Human studies, 9*(2-3), 185-204.

Oliveira, M., & Freitas, T. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. In *Speech Prosody 2008* (pp. 485). Campinas, Brazil.

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

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics, 11*, 243-257.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29*(1), 1-63.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. In *ICSLP'92*. Banff, Canada.

Ward, N. (2006). A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic, In *Proc. of Interspeech*, Pittsburgh, Pennsylvania, USA.

Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. In *ICPhS 2003*. Barcelona, Spain.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-578). Chicago.

Zevin, J., & Farmer, T. (2008). Similarity Between Vowels Influences Response Execution in Word Identification. In *Proc. of Interspeech*. Brisbane, Australia.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language, 82*(3), 515-535.

Anaphora and Direct Reference: Empirical Evidence from Pointing

Massimo Poesio
Università di Trento

Hannes Rieser
University of Bielefeld

Abstract

Empirical evidence from body measurements suggests that the referent of a demonstration is not directly specified, but obtained by applying a default inference rule to the region specified by the pointing cone. Building on this evidence we propose a unified theory of anaphoric and demonstrative uses in which accessibility is obtained via resource situations.

1 Introduction

The traditional semantics of demonstrative expressions is based on a sharp distinction between anaphoric reference and direct reference derived from Kaplan. Kaplan proposed that

> '[] each demonstrative, d, will be accompanied by a *demonstration*, δ, thus: d[δ]. The character of a complete demonstrative is given by the semantic rule: In any context c, d[δ] is a directly referential term that designates *the demonstratum*, if any, of δ in c, and that otherwise designates nothing. Obvious adjustments are to be made to take into account any common noun phrase which accompanies or is built into the demonstrative.' (Kaplan 1978, pp. 771-772).

Thus, for instance, demonstrative *This chair* in *This chair was hand-made by an artisan* accompanied by a pointing gesture to the chair (the demonstration) is interpreted as direct reference to the chair. By contrast, *This chair* in the text *Hannes bought a chair in the centre of Rovereto. This chair was hand-made by an artisan* is anaphoric. The two expressions have radically different interpretations.

This distinction has been challenged by semanticists such as Roberts (2002), as well as in corpus linguistics (Gundel et al, 1993); we will argue in this paper that it is also seriously challenged by empirical evidence about pointing. Modern body tracking methods make it possible to measure with precision what a subject is pointing at. In a study combining experiments, statistical investigation, computer simulation and theoretical modelling techniques, Lücking, Pfeiffer and Rieser (2009) investigated the semantics and pragmatics of co-verbal pointing in dialogue. Lücking, Pfeiffer and Rieser established a semantic and two pragmatic hypotheses concerning the role of pointing in multi-modal expressions, and tested these with an annotated and rated corpus of Object Identification Games. The corpus was set up in experiments in which body movement tracking techniques were used to generate a space of pointing measurements. Statistical investigation and simulations showed that especially pointing to distal areas is not consistent with the semantic hypothesis. On the other end, the results can be predicted with high accuracy by hypothesizing a simple default inference extracting from the pointing gesture information sufficient to identify a referent uniquely. These results cast serious doubt on classical theories of the semantics-pragmatics interfaces insofar as they indicate that compositionality often presupposes pragmatically computed values.

In the paper we summarize the results of the Lücking *et al* study and formulate a unified hypothesis about the interpretation of demonstratives in terms of PTT (Poesio & Traum, 1997; Poesio & Rieser, submitted), a theory of the semantics and pragmatics of dialogue in which all actions in the discourse situation are explicitly represented and in which default inferences leading to their connection can be formulated.

2 A Brief Introduction to PTT

PTT (Poesio and Traum, 1997; Poesio & Muskens, 1997; Poesio & Rieser, submitted) is a theory of dialogue semantics and dialogue interpretation developed to explain how utterances are incrementally interpreted in dialogue, considering both their semantic impact and their impact on aspects of dialogue interaction traditionally considered as outside the scope of semantic theory, building on the work of Clark (1996) and on ideas from Situation Semantics (Barwise and

Perry, 1983; Cooper, 1996, Ginzburg, to appear). In this section we briefly discuss the two aspects of the theory that are relevant for the formulation of our unified hypothesis about demonstratives; for more details on PTT, including a complete fragment for German, see (Poesio & Rieser, submitted).

2.1 The common ground as a record of the discourse situation

PTT is an INFORMATION STATE theory of dialogue (Larsson & Traum, 2000; Stone, 2004; Ginzburg, to appear) in which the participants in a conversation maintain an information state about the conversation consisting of private information together with a conversational score including 'grounded' (Clark, 1996) and semi-public information. One respect in which PTT derives from Situation Semantics is hypothesis that the conversational score consists of a record of all actions performed during the conversation, i.e., what in Situation Semantics is called the DISCOURSE SITUATION (Barwise and Perry, 1983; Ginzburg, to appear). An ordinary conversation does not consist only of actions performed to assert or query a proposition, but also of actions whose function is to acquire, keep, or release a turn, to signal how the current utterance relates to what has been said before, or to acknowledge what has just been uttered. The discourse situation also contains information about non-verbal actions such as pointing.

Poesio and Traum (1997) argued that this view of the conversational score could be formalized using the tools already introduced in DRT (Kamp and Reyle, 1993)—specifically, in Muskens's Compositional DRT (1996), because speech acts-- CONVERSATIONAL EVENTS, in PTT terms—and non verbal actions are in many respects just like any other events, and because conversational events and their propositional contents can serve as the antecedents of anaphoric expressions. For instance, Poesio & Rieser (submitted) hypothesize that the two directives in (1) (an edited version of two turns from the Bielefeld ToyPlane Corpus) result in the update to the common ground in (2).[1]

(1) Inst: So jetzt nimmst Du eine orangene
Schraube mit einem Schlitz
So now you take a orange screw with a slit
Cnst: Ja
OK
Inst: Und steckst Sie dadurch, von oben, daß
also die drei festgeschraubt werden dann
and you put it through from above so that
the three get fixed

(2) [K1.1, up1.1, ce1.1, K2.1, up2.1, ce2.1 |
K1.1 **is** [e,x,x3| **screw**(x), **orange**(x),
slit(x3), **has**(x,x3),
e:**grasp**(Cnst, x)],
up1.1: **utter**(Inst, "So jetzt nimmst Du … "),
sem(up1.1) **is** K1.1,
ce1.1:**directive**(Inst&Cnst,Cnst,K1.1),
generate(up1.1, ce1.1),
K2.1 **is** [x6,e',s,w,y| x6 **is** x,
e':**put-through**(Cnst,x6,hole1),
w **is** wing1, y **is** fuselage1,
s: **fastened**(w,y)],
up2.1:**utter**(Inst, "und steckst Sie ... "),
sem(up2.1) **is** K2.1,
ce2.1:**directive**(Inst,Cnst, K2.1) ,
generate(up2.1, ce2.1)]

(2) records the occurrence of two conversational events, *ce1.1* and *ce1.2*, both of type **directive** (Matheson et al, 2000) whose propositional contents are separate DRSs specifying the interpretation of the two utterances in (1). The contents of conversational events are associated with propositional discourse referents K1.1 and K2.1 (discourse referents whose values are DRSs) as proposed in (Poesio and Muskens, 1997) and done, e.g., in SDRT (Asher and Lascarides, 2003). It is further assumed in PTT that dialogue acts are **generated** (Pollack, 1986) by locutionary acts (Austin, 1962), which we represent here as events of type **utter**.

Non-verbal actions are also viewed in PTT as conversational events albeit of a different type. So for instance an act of pointing by agent DG would lead to the following update of both agents' information state:

(3) [pe1.1| pe1.1:**point**(DG, α)]

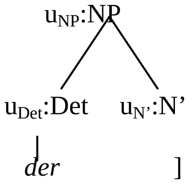
where α is what DG is pointing at—determining experimentally what is α was the main question addressed by (Lücking, Pfeiffer and Rieser, 2009), as we will see.

It is assumed in PTT (Poesio, 1995) that the conversational score is incrementally updated whenever a verbal or non-verbal event is perceived. In particular, each word incrementally updates the discourse situation with a locutionary act of type **utter** and with syntactic expectations about the occurrence of more complex utterances as hypothesized in LTAG (Schabes et al, 1988). Thus, for instance, an utterance of definite article *der* results in the conversational score being updated with the occurrence of an utterance u_{der} of type Det (a *micro conversational event* (MCE) (Poe-

[1] We use the syntax from Muskens (1996) and his **is** operator. We use terms with the prefix *ce* (ce1, ce2, etc) for discourse referents denoting conversational events; terms with the prefix *u* for utterances; terms denoting other events will be indicated by the prefix *e*. We will indicate terms denoting states by the prefix *s*; all other terms will have prefixes *x*, *w*, *y*, and *z*.

sio 1995a)) and with the expectation that this utterance will be part of an utterance a of a NP which will also include an utterance $u_{N'}$ of an N'. We will depict this update as follows:

(4) *Der* ⇒ [u_{der}, u_{NP}, $u_{N'}$|



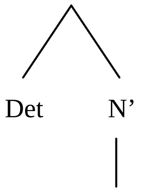
We further assume that MCEs have a (conventional) semantics associated to them, and that this semantics is the value of a **sem** function (in fact, a family of functions $\mathbf{sem}_{[]}$, $\mathbf{sem}_{[\pi]}$, etc.). We assume that the lexical semantics of words that update the discourse model and of anaphoric expressions is as proposed in Compositional DRT (Muskens, 1996), as discussed below, and that the semantics of phrasal utterances is obtained compositionally via defeasible inference rules that by default assign, for instance, to an utterance of an NP like u_{NP} above the conventional semantics **sem(**u_{NP}**)** resulting from the application of **sem**(u_{der}) to **sem**($u_{N'}$), but that can be overridden e.g., in the case of metonymy or as in the case of anaphoric expressions, as we will see below (Poesio & Traum 1997, Poesio to appear, Poesio & Rieser submitted).

2.2 Anaphora in PTT

The current treatment of definites and anaphoric expressions in PTT (Poesio, to appear) is based on the 'functional' interpretation of definite NPs due to Loebner (1987) but has many points in common with the treatment proposed e.g., in (Chierchia, 1995). According to Loebner, what all definites have in common is that they are terms – i.e., *functions* that may take a different number of arguments, but all have a value of type e. Thus, for example, proper name *Jack* would have as translation the (0-argument) function ι x. (x = j), whereas the definite description *the pope* would have as translation the 1-argument function λs. ι x. (x = **pope**(s)(x)), taking a situational or temporal argument *s*.

The Loebnerian treatment of definite descriptions is translated in the PTT framework by assigning to the definite article (e.g., German *der*) an elementary tree with the CDRT semantics below.

(5) *der*: NP



der: λP. λP'. ([y|y = ι x. P(x)]; P'(y);

According to Loebner, a definite is licenced either because P is semantically functional, as in classical examples like *the king of France*, or because P is turned into a function by a modifier, as in *the first point to make is that*.., or because P is pragmatically *coerced* into a function by resolving it.

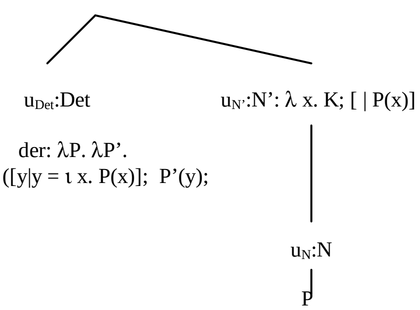
Standard DRT accessibility would predict that anaphoric interpretation in discourse situations is not possible: e.g., it would predict that the antecedent of *Sie* in (1), the screw, is not accessible. But Poesio (1993) proposed that what makes antecedents accessible in discourse situations is that definites uniformly receive their interpretation through a **resource situation** (Barwise and Perry, 1982; Cooper, 1996, Ginzburg, to appear). The resource situation hypothesis was recast in DRT terms in (Poesio 1994, Poesio & Muskens, 1997) by proposing that resource situations are contexts—DRSs—and that all anaphoric expressions contain an implicit variable over contexts, and it is this variable that supplies the value for the discourse referent. So for instance the NP *der Kreis* interpreted anaphorically would receive the following presuppositional interpretation:

(6) *der Kreis* ⇒
λP'. ([y|y = ι x. K; **ring** (x)]; P'(y))

Where K is a resource situation where an object of type **ring** is particularly salient. (Note that K is used *presuppositionally*.)

The anaphoric interpretation in (6) is obtained through a *coercion process* –a defeasible semantic composition rule—that assigns to the N' in a definite construction as an interpretation a predicate λ x. K; [| P(x)] that is pragmatically functional wrt a resource situation K, as in (5'):

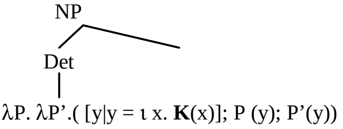
(5')



These coercion rules were called **Principles for Anchoring Resource Situations** (PARS) in (Poesio, 1993; Poesio, 1994). One such principle ruled anaphora, licensing the coercion above when the content K of a speech act is globally salient and contains an object of the right type. (Full specification of the principle omitted for reasons of space.) A second principle made parts of the visual scene salient as results of instructions that directed the attention to those parts of the scene. We will argue here that the evidence from Lücking *et al* suggests that pointing is another way for anchoring resource situations, thus providing a unified account of all types of definite reference, as already proposed by e.g., Roberts (2002) whose account, however, differs from ours in crucial respects.

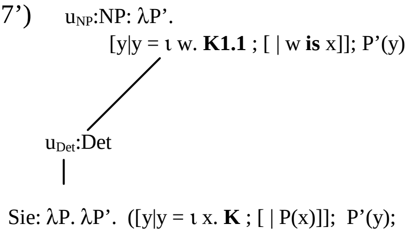
It has often been argued that, syntactically, pronouns in English are like determiners. The translation proposed for pronouns such as *Sie* in (7) makes pronouns behave semantically like determiners, as well.

(7)



This translation is based on the idea that whereas the definite article may be licenced by a semantically functional, but non anaphoric, predicate, pronouns must always be licenced pragmatically—i.e., there must be some highly salient resource situation K containing a highly salient object. Furthermore, pronouns require a contextual property restricting the interpretation of the referent y: resolving a pronoun amounts to identifying such restriction. One obvious candidate is an identity property—i.e. a property of the form λw ([| w **is** z]) for z a discourse entity. According to the treatment just sketched, resolving *Sie* in (1) involves identifying K1.1 in (2) as resource situation and x as antecedent (i.e., applying the result to the identity property λw ([| w **is** z])), obtaining the following interpretation.

(7')


u_NP:NP: λP'.
[y|y = ι w. K1.1 ; [| w is x]]; P'(y)
u_Det:Det
Sie: λP. λP'. ([y|y = ι x. K ; [| P(x)]]; P'(y);

3 Experimental Evidence on Pointing

3.1 Semantic and Pragmatic Hypotheses on Pointing

Putting together assumptions by the early Wittgenstein, Davidson and Kaplan (1978), we can formulate the **Semantic Hypothesis** about pointing as follows:

(Sem) A demonstration [pointing] going together with a simple or a complex demonstrative in context c designates exactly one object, the object referred to in c.

The experimental literature in experimental pragmatics (Bangerter, 2004;Bangerter& Oppenheimer, 2006; Clark, 2003; Clark & Bangerter, 2004), however, leads to two rivalling hypotheses. The first one shifts the emphasis to inference to an object (*Strong Prag*); the second one deals with the focus of attention (*Weak Prag*) doing away with the notion of an object referred to altogether.

(Strong Prag) A demonstration triggers a perceptually based inference wrt a context c from the pointing device to the object referred to in c.

(Weak Prag) Demonstration shifts its addressee's attention towards a specific domain in a context c.

If one can show that (*Sem*) characterizes pointing behaviour in general, one does not need the pragmatics hypotheses, since pointing acts behave like constants. If one finds out that pointing success depends on contextual parameters, one has to resort to pragmatic hypotheses. Furthermore, if one finds evidence for (*Strong Prag*), one obviously has proved (*Weak Prag*), granted that one ties (*Strong Prag*) to intention and attention. Anyway, (*Weak Prag*) alone is not of much help, since it is too weak to distinguish pointing from focusing or emphasizing. For this reason we only concentrate on StrongPrag here.

3.2 Experimental Methods

In order to test the semantic and pragmatic hypotheses Lücking, Pfeiffer and Rieser (2009) conducted an empirical study using a so-called **Object identification game setting**. In this setting there are two participants, called *Description Giver* (DG) and *Object Identifier* (OI). DG and OI are set within the operational area of a marker-based optical tracking system with nine cameras (6DOF tracker, ART GmbH). The information delivered by the cameras provides positions and orientations of optical markers in an absolute coordinate system. Only the DG is tracked by markers on arms, index fingers, hands, and head. Both OI and DG are located around a table (77.5 × 155.5 cm) with 32 parts of a *Lorentz Baufix* toy air-plane, the experimental domain.

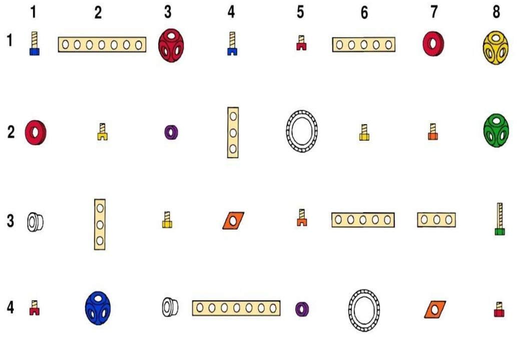


Figure 1: The experimental domain.

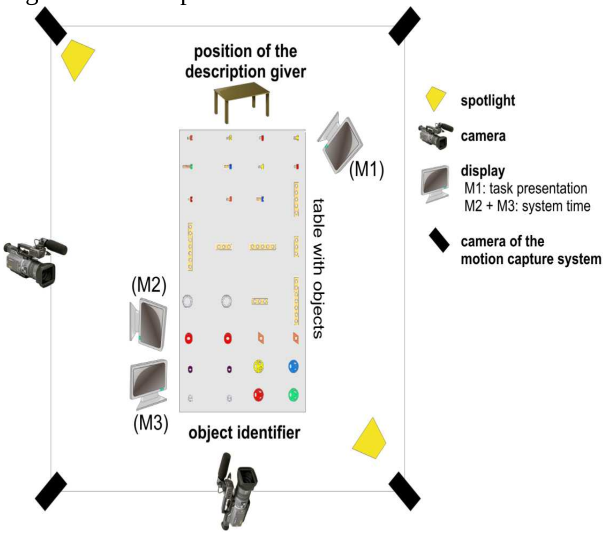


Figure 2: Setup of the setting within the interaction space of the motion capturing system. The interaction is observed by two video cameras and nine cameras of a motion capturing system.

The interaction between DG and OI was restricted to avoid negotiation processes. It consists of three steps: 1. Demonstration by DG (bimodal or only gestural); 2. Interpretation and identification by OI with a pointer (the referent remains in its place); 3.Verbal feedback by DG. The dialogues in the object identification games were of the following sort (original data):

(8) DG:Der weiße Kreis da bei mir direkt auf
der Linie, der weiße Kreis, der Reifen
da.
*The white circle near to me directly on
the line, the white circle, the ring here.*
OI: [pointing].
DG: Ja.
OK.

3.3 Operationalization of the Hypotheses, Results and Analysis

The precise measurements of the motion capturing system provide us with the means to closely investigate pointing, reconstructing position and orientation of the index finger during each stroke. We also know the positions of the objects on the table. Thus Lücking, Pfeiffer and Rieser were able to project for each demonstration the beam from the index finger at the time of the stroke and compute whether the ray hits an object. It can be determined by the orientation of the index finger (index-finger-pointing, IFP) or, alternatively, by the direction of gaze, aiming at the target over the tip of the finger (gaze-finger-pointing, GFP).

Testing the *(Sem)* hypothesis on the pointing gesture means translating it in terms of predictions that can be measured using these methods. Lücking *et al* proposed the following:

Strict Operationalisation of the *(Sem)* hypothesis: *A pointing gesture refers to the object which is hit by a pointing-ray extending from the index-finger*.

If we calculate for each variant a pointing-ray originating in the index finger, oriented along the specific direction and intersect it with the table surface, we get a distribution of points around the object showing precision and accuracy of the pointing gesture (see Fig. 3).

Looking at Fig. 3, we see that pointing is fuzzy. In most of the demonstrations the projected ray fails the target. Reconsidering the semantic hypothesis in the context of the results shown in the bagplots of Fig. 3, a more relaxed conceptualization comes to mind which could compensate for the low precision of pointing but still allows us to sustain the *(Sem)* hypothesis. This leads to a relaxed operationalization of the *(Sem)* hypothesis using a pointing-cone to model the low precision of pointing:

Relaxed Operationalisation of the *(Sem)* hypothesis: *A pointing gesture refers to the object which is hit by a pointing-cone extending from the index-finger.*

However, the success rates were too low to provide a foundation for the weaker *(Sem)* hypothesis, leading finally Lücking *et al* to conclude that pointing is not a semantic referring device.

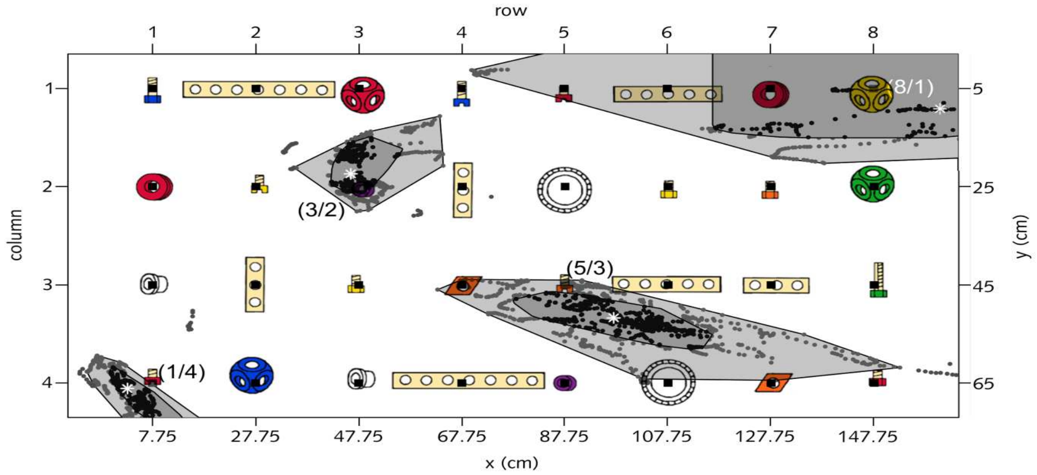


Fig. 3: The distribution of the objects on the table is overlaid with bagplots visualizing precision and accuracy of the pointing gestures for four selected objects (indicated by the pair of coordinates). The dots mark the intersections of a pointing ray with the table surface; a star indicates the mean position. Darker shading covers 50 percent and lighter shading 75 percent of the points. Obviously, most of the rays fail to hit the target object.

As stated, for semantics we would need a test providing a definite single object for every demonstration. This is different in pragmatics. Here we can use inference to choose among a set of possible referents selecting the most likely one intended by DG. Examining the *(Strong Prag)* hypothesis we only used motion capturing data. An example of inference process identifying one object among the objects in the pointing cone could be one that ranks the delimited objects according to their distance from the central axis of the pointing-cone. Lücking *et al* called this heuristics (INF):

(INF) *An object is referred to by pointing only if*

a) *the object is intersected by the pointing cone and*

b) *the distance of this object from the central axis of the cone is less than any other object's distance within this cone.*

Lücking *et al* further weakened their relaxed operationalisation for the *(Sem)* hypothesis and allowed several objects to lie within the pointing-cone as long as the intended target object can be singled out from the set of objects delimited *via* a subsequent inference. So they arrived at the following:

Operationalisation of the *(Strong Prag)* hypothesis: *A pointing gesture refers to the one object selected by an appropriate inference from the set of objects covered by a pointing-cone extending from the index-finger.*

(In other words, the object demonstrated is the one nearest to the axis of the pointing cone where a) and b) are considered to be necessary conditions.)
This weighting heuristics succeeds in 96 percent of the cases when using Index-Finger-Pointing and in 92 percent of the cases when using Gaze-Finger-Pointing. These results are mainly due to the weighting heuristics and not to a clear-cut cone intersection. We take these figures as strong evidence that *(Strong Prag)* holds, i.e., that the referent in demonstrative uses is arrived at via a pragmatic inference process which, however, is not infallible (i.e., it is a defeasible inference)

4 A Unified Account of Anaphoric And Demonstrative Uses

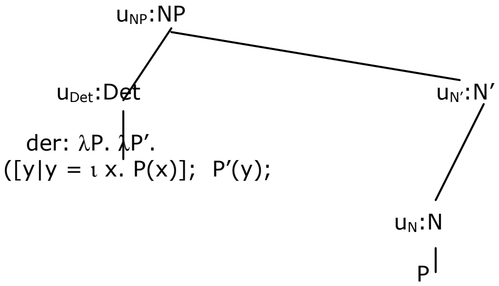
If it is true that the referents of demonstratives are obtained through an inference like (INF), then there is no need to stipulate that demonstrative phrases like *this chair* are ambiguous between an anaphoric and a direct reference use: the translation of definites proposed in 2.2 can serve as the lexical translation for definites like *der weisse Kreis* both when used anaphorically and when used demonstratively in (8).

Assuming that the visual scene is a resource situation K_{visual} as proposed in (Poesio, 1993), then the results by Lücking, Pfeiffer and Rieser (2009) suggest that an act of pointing identifies a subset of this situation $K_{pointing}$- the set of objects in the pointing cone.

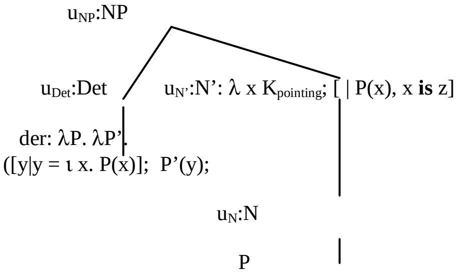
(3') [pe1.1| pe1.1:**point**(DG, $K_{pointing}$)]

INF is a defeasible inference rule analogous to the Principles for Anchoring Resource Situations proposed in (Poesio, 1993), except that it coerces the interpretation of the nominal predicate to be identical with the object z in $K_{pointing}$ which is the closest object to the central axis of the cone:

Principle for Anchoring Resource Situations via Pointing
If u_{NP} is a micro-conversational event with structure



u_{NP} is cotemporal with pointing gesture pe1.1:**point**(DG, $K_{pointing}$), and z ∈ $K_{pointing}$ is the closest object to the pointing axis of the cone, then defeasibly coerce the interpretation of N’ to λ x $K_{pointing}$; [| P(x), x **is** z]:


u_{NP}:NP
u_{Det}:Det
$u_{N'}$:N’: λ x $K_{pointing}$; [| P(x), x **is** z]
der: λP. λP’.
([y|y = ι x. P(x)]; P’(y);
u_N:N
P

5 Related Work

Roberts (2002) aims at a unified account of several types of demonstratives, pronominal and descriptive, accompanied by canonical demonstrations or textual deixis and discourse deixis. The following quotation sums up her approach: *The heart of this proposal is the claim that a demonstrative NP conventionally presupposes that a familiar discourse referent for the demonstratum of its associated demonstration is the same as the discourse referent which satisfies the NP’s familiarity presupposition.* (p. 315) There are similarities between the PTT-account of demonstrative definites and the one presented in Roberts: for example, the hypothesis that the definiteness information is presuppositional (p. 312). The difference between her account and the one presented here is that here notions like demonstration, demonstratum, pointing, direction pointed at, context, salience, proximity and non-proximity are part of the *explicandum* for which the experimental situation, including body tracking devices serves as an *explicans*. So little is taken for granted and notions are backed up by rigid measurements. Similarities and differences would merit a more thorough discussion.

6 Conclusions

Modern experimental techniques are beginning to make it possible to empirically test fine-grained semantic hypotheses. We argued in this paper that in particular evidence from body measurements can be used to evaluate the extent to which demonstrations uniquely identify the referent of a demonstrative. The evidence is that the identification requires a pragmatic inference process. The next step will be to revisit other linguistic evidence for direct reference at the light of these data.

Acknowledgments

We are grateful to Andy Lücking and Thies Pfeiffer for letting us use material from an unpublished paper. This research was in part supported by the University of Trento, Centre for Mind / Brain Sciences, and by the University of Bielefeld.

References

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. CUP

Austin, J. L. (1962). How to Do Things with Words. Oxford, OUP

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419.

Bangerter, A. and Oppenheimer, D. M. (2006). Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture*, 6(1):85–102.

Barwise, Jon & Perry, John (1983). *Situations and attitudes*. MIT Press.

Butterworth, G. and Itakura, S. (2000). How the eyes, head and hand serve definite reference. *British Journal of Developmental Psychology*, 18:25–50.

Chierchia, G. (1995). *Dynamics of Meaning*. Univ. of Chicago Press.

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

Clark, H. H. (2003). Pointing and placing. In Kita, S., editor, *Pointing*, chapter 10, pages 243–269. Lawrence Erlbaum.

Clark, H. H. and Bangerter, A. (2004). Changing ideas about reference. In Noveck, I. A. and Sperber, D., editors, *Experimental Pragmatics*, pages 25–49. Palgrave Macmillan, New York.

Cooper, R. (1996). The role of situations in generalized quantifiers. In Lappin, S. (ed), *Handbook of Contemporary Semantic Theory*, Blackwell.

Ginzburg, J.. To Appear. *Semantics for Conversation*. CSLI Publications

Kamp, H. and U. Reyle (1993). *From Discourse to Logic*. Dordrecht: Kluwer.

Kaplan, D. (1978). On the logic of demonstratives. *J. of Philosophical Logic*, 8:81–98.

Larsson, S. and Traum, D. (2000). Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, v.6, pp. 323-340.

Loebner, S. (1987) Definites. *Journal of Semantics*, 4, 279-326

Lücking, A., Pfeiffer, Th., and Rieser, H. (2009). Pointing and Reference Reconsidered (submitted)

Matheson, C., M. Poesio, and D. Traum, (2000). Modeling Grounding and discourse obligations using update rules, *Proc. Of the NAACL*.

Muskens, R. (1996). Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy 19*, 143-186

Poesio, M. (1993). A Situation-Theoretic Formalization of Definite Description Interpretation. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, (editors), *Situation Theory and its Applications, vol. 3,* p.339--374.

Poesio, M. (1994). *Discourse Interpretation and the Scope of Operators.* PhD dissertation, University of Rochester.

Poesio, M. (1995). "A Model of Conversation Processing Based on Micro Conversational Events," In *Proceedings of the Cognitive Science Society*.

Poesio, M. to appear. *Incrementality and underspecification in semantic interpretation*. CSLI Publications.

Poesio, M. and Muskens, R. (1997). “The Dynamics of Discourse Situations”, *Proc. 11th Amsterdam Colloquium*, 247-252.

Poesio, M. and Rieser, H., submitted. *Completions, coordination and alignment in dialogue*.

Poesio, M. and Traum, D. (1997). Conversational Actions and Discourse Situations, *Computational Intelligence*, v. 13, n.3, pp.1- 45

Pollack, M. (1986), *Inferring Domain Plans in Question-Answering.* PhD, Univ. of Pennsylvania.

Roberts, C. (2002). Demonstratives as definites. In K. van Deemter & R. Kibble (eds.) *Information Sharing*. CSLI Press, Stanford, 89-196.

Schabes, Y., Abeillé, A., & Joshi, A.K. (1988). New parsing strategies for tree adjoining grammars. In *Proceedings of COLING*, pp. 578– 583. Budapest, August.

Stone, M. (2004) Intention, interpretation and the computational structure of language. *Cognitive Science* 28(5):781-809.

Viability of a Simple Dialogue Act Scheme for a Tactical Questioning Dialogue System

Ron Artstein **Sudeep Gandhe** **Michael Rushforth** **David Traum**
Institute for Creative Technologies, University of Southern California
13274 Fiji way, Marina del Rey, CA 90292, USA
<lastname>@ict.usc.edu

Abstract

User utterances in a spoken dialogue system for tactical questioning simulation were matched to a set of dialogue acts generated automatically from a representation of facts as ⟨object, attribute, value⟩ triples and actions as ⟨character, action⟩ pairs. The representation currently covers about 50% of user utterances, and we show that a few extensions can increase coverage to 80% or more. This demonstrates the viability of simple schemes for representing question-answering dialogues in implemented systems.

1 Introduction

Dialogue acts are often used as representations of the meaning of utterances in dialogue, both for detailed analyses of the semantics of human dialogue (e.g., Sinclair and Coulthard, 1975; Allwood, 1980; Bunt, 1999) and for the inputs and outputs of dialogue reasoning in dialogue systems (e.g., Traum and Larsson, 2003; Walker et al., 2001). There are many different taxonomies of dialogue acts, representing different requirements of the taxonomizer, both the kinds of meaning that is represented and used, as well as specifics of the dialogues and domain of interest (Traum, 2000). There are often trade-offs made between detailed coverage and completeness, simplicity for design of domains, and reliability for both manual annotation and automated recognition.

In this paper, we examine the adequacy for use in tactical questioning characters of a fairly simple dialogue act scheme in which the set of possible dialogue acts is automatically created by applying illocutionary force constructor rules to a set of possible semantic contents generated by an ontology of a domain. The advantage of this kind of scheme is that a dialogue system is fairly easily authored by domain experts who work on the level of a simple ontology, without detailed knowledge of dialogue act semantics and transitions. The disadvantage is that it (intentionally) has limited expressibility in that some dialogue functions are not directly expressible, and it is not so easy to represent multiple meanings of an utterance.

We evaluated the scheme as follows: first we created an initial version of the character by authoring the ontology and using this to automatically generate the set of dialogue acts that fit into designed protocols for tactical questioning dialogues. Initial Natural Language Understanding and Generation capabilities were also authored using a classification approach (Leuski and Traum, 2008). The complete system was then used to generate a corpus of man-machine dialogues by having people interact with the character. Finally, the user utterances in this corpus were annotated by multiple annotators according to the dialogue act taxonomy. We evaluated both the coverage of the dialogue act taxonomy and the reliability of the annotations. The reliability of the matching was 49% above chance and full agreement was reached for only 30% of the utterances, but a detailed analysis shows that coverage of the current representation is closer to 50%, and that a few extensions can bring it to 80% or more.

The rest of the paper is structured as follows. Section 2 describes the tactical questioning genre of dialogue, and the dialogue system architectures that have been used to create specific domains and characters for this genre, as well as the development process for creating characters. The domain specification and dialogue representation is described in section 3. Section 4 presents the specific experiments, with the results presented in section 5, and a detailed analysis of the coverage of the dialogue act representation in section 6.

2 The Tactical Questioning Domain

Tactical Questioning is an activity carried out by small-unit military personnel, defined as "the expedient, initial questioning of individuals to obtain information of immediate value" (U.S. Army, 2006). A tactical questioning dialogue system is a simulation training environment where virtual characters play the role of a person being questioned. Unlike typical question-answering systems, tactical questioning characters are designed to be non-cooperative at times. The character may answer some of the interviewer's questions in a cooperative manner, but may refuse to answer other questions, or intentionally provide incorrect answers (lie). Some of the strategies that an interviewer may use in order to induce cooperation include building rapport with the character, addressing their concerns, making promises and offers, as well as threatening or intimidating the character; the purpose of the dialogue system is to allow trainees to practice these strategies in a realistic setting.

Building tactical questioning dialogue systems is an on-going project at Institute for Creative Technologies, which has evolved through a number of different architectures; see Traum et al. (2008) for a detailed overview. The third and current architecture introduces an intermediate representation for dialogue acts, a finite-state representation of local dialogue segments, a set of polices for engaging in the network, and a rule-based dialogue manager to update the context and choose dialogue acts to perform (Gandhe et al., 2008). This functionality allows for short subdialogues where the character can ask for and receive certain assurances (such as protection or confidentiality) and still remember the original question asked by the trainee.

With earlier tactical questioning systems, based on text-to-text classifiers, character development typically proceeds in a bottom-up fashion: we start by collecting a corpus of in-domain human-human dialogues through roleplays or Wizard-of-Oz sessions, and use this as a starting point for the implementation of a question-to-response mapping. This mapping is refined as the system goes through iterative test cycles: additional user questions are gathered and mapped to appropriate responses, and the character's domain is expanded by authoring new responses. The use of an intermediate representation for dialogue acts requires top-down authoring: the first step is specifying the domain, that is the set of facts that the character can be questioned about; dialogue acts are created automatically from the domain specification, and these represent what the character can understand. When iterative testing with users reveals deficiencies or gaps in the character's understanding capabilities, expansion cannot take place at the textual level but must go back to the domain specification or the rules for creating dialogue acts.

Our tactical questioning system is designed for rapid prototyping and creation of multiple characters with shared knowledge about a specific domain (Gandhe et al., 2009). The representation language for dialogue acts is therefore fairly simple, unlike that of more complex systems (Traum and Hinkelman, 1992; Traum and Rickel, 2002; Keizer and Bunt, 2006). The core of the representation language rests on facts represented as ⟨object, attribute, value⟩ triples, and which constitute the material for questioning by the user. For the system to succeed, this impoverished representation must capture enough information about the users' actual utterances.

3 Domain specification and dialogue acts

In the scenario for the experiment, the user plays the role of a commander of a small military unit in Iraq whose unit had been attacked by sniper fire. The user interviews a character named Amani who was a witness to the incident and is thought to have some information about the identity of the attackers (Figure 1). Amani's knowledge about the incident is represented as facts which are ⟨object, attribute, value⟩ triples; each fact is either true or false – false facts are used by Amani when she wants to tell a lie. Table 1 gives some facts about the incident. For example, Amani knows that the name of the suspected sniper is Saif, and that he lives in the store. She can lie and say that she doesn't know the suspect's name. She does not

Table 1: Some facts about the incident

Object	Attribute	Value	T/F
strange-man	name	saif	true
strange-man	name	unknown	false
strange-man	location	store	true
brother	name	mohammed	true

Figure 1: Amani – A virtual human for Tactical Questioning. The figure sitting in the chair represents Amani's brother, Mohammed, who is not an interactive character.

have an available lie about the suspect's location, though she can always refuse to answer a question.

In addition to facts about the incident, the domain specifies certain attributes that are unique to the characters (both Amani and the user). Characters may have attitudes towards objects; they can perform actions such as offers, threats, admissions and suggestions; and they have a set of compliments and insults that they can use for building rapport with their interlocutors. All of these, together with the facts, are specified in an XML format that defines the domain of interaction (Gandhe et al., 2008; Gandhe et al., 2009).

The domain represents the character's knowledge. It defines a space of dialogue acts which are the interpretations of language utterances; this is the level at which the character reasons about the conversation. Dialogue acts are automatically generated from the domain specification, by applying an illocutionary force (or *dialogue act type*) to a semantic content containing the relevant portion of the domain specification. Each fact generates 3 dialogue acts – an assertion of the fact by the character, a yes-no question by the user, and a wh-question by the user which is formed by abstracting over the value. For example, the fact ⟨strange-man, name, saif⟩ defines a dialogue act by Amani with a meaning equivalent to "the suspect is named Saif", and two questions by the user, equivalent in meaning to "is the suspect named Saif?" and "what is the suspect's name?" (note that distinct facts may give rise to identical question dialogue acts). Each user action generates a corresponding dialogue act, as well as forward-function (elicitation) and backward-function (response) dialogue acts by the character (Allwood, 1995; Core and Allen, 1997). Currently, elicitations are only defined for offers (so Amani can ask for a particular offer); responses of various kinds are defined for all of the user's illocutionary acts (offers, threats, compliments, insults). Additionally, some generic dialogue acts are defined independently of the domain – these include greetings, closings, thanks, grounding acts (such as repeat-back or request-repair), and special dialogue acts that are designed to handle out-of-domain dialogue acts from the user. Table 2 shows the various dialogue act types used in the current tactical questioning architecture and the number of full acts of each type generated for the user and Amani, given Amani's ontology. The full algorithm for generating dialogue acts is presented in Gandhe et al. (2009).

The link between dialogue acts and actual utterances is done via Natural Language Understand-

Table 2: Dialogue acts in the Amani domain

Dialogue Act Type	Amani	User
accept	1	1
ack	1	1
apology	1	1
assert	36	
closing	3	3
compliment		3
elicit	6	
greeting	1	1
insult		2
offer		3
offtopic	1	1
pre_closing	3	3
refuse_answer	1	1
reject	1	1
repeat-back	10	10
request-repair-object	10	10
request_repair	1	1
response	54	3
thanks	1	1
unknown		1
whq		31
ynq		35

ing and Generation modules. The NLU uses a statistical language modeling text classification technique (Leuski and Traum, 2008), trained on pairings of user utterances to dialogue acts, to determine the appropriate dialogue act for novel text produced by the speech recognizer; if it cannot find a good match with high confidence, the classifier outputs a special “unknown” dialogue act which informs the dialogue manager that the user utterance has not been properly understood. A similar classifier, trained on mappings from character dialogue acts to text, is used for generation. A dialogue manager is responsible for the transition from user dialogue acts, provided by the NLU module, to character dialogue acts which are passed to the NLG module. The dialogue manager is based on the information state model (Traum and Larsson, 2003). It uses rules described in State Chart XML (Barnett et al., 2008) to keep track of obligations (Traum and Allen, 1994), questions under discussion, offers and threats; similar rules track the character’s emotional state (Roque and Traum, 2007) as well as grounding (Roque and Traum, 2009). The main responsibilities of the dialogue manager are to update the information state of the dialogue and use it to select the contents of the response.

The dialogue manager drives the character’s interaction and is responsible for all of its reasoning, and it works at the level of dialogue acts. But users have their own mental models of what can be said to the system, and are not aware of what distinctions the system can represent. We therefore need to determine whether the dialogue act representation – intentionally designed to be simple – is rich enough to capture the meaning in user utterances. To answer this question we carried out an experiment with actual user utterances.

4 Experiment

To test how well the automatically generated dialogue acts capture the meaning of actual user utterances, we performed a matching experiment. First, we collected a corpus of interactions of users with the initial version of Amani. The dialogue participants were all staff members at ICT; they had experience talking to virtual characters, including question-answering characters, but were not familiar with the Amani scenario prior to the dialogues, nor had any experience talking to a third-generation question-answering character. Dialogue participants were given an instruction sheet with some information about the incident, the character, and suggestions for interaction (e.g. the possibility of making offers) – similar to the instruction sheet a trainee would receive. The instructions did not include guidance about particular language to use with the character. We collected a total of 261 user utterances from 16 dialogues, which varied in length from 2 to 40 utterances.

User utterances from interactions with the system were transcribed, and then matched to the existing user dialogue acts by 3 experienced annotators. The annotators were all involved with the project: they included the first and third authors, and a student annotator. The purpose of the study was to find out how adequate the current domain representation was, what extensions it needed, and what systematic problems arose that might require not only changes to the domain specification but to the way dialogue acts are defined. Since this study was of an exploratory nature, the instructions were very simple and given in a single sentence: “Match each user utterance to the most appropriate player speech act; if none is appropriate, match to ‘unknown’.”

Annotators matched utterances to dialogue acts using the domain creation tool (Gandhe et al., 2009). We proceeded under the assumption that each utterance text is mapped to a single dialogue act, not taking into account context that would disambiguate different dialogue acts for the same text appearing at different times. This was not a major concern with our corpus, because the vast majority of utterance texts occur only once (224 distinct utterance texts), and of the 7 utterance texts with frequency of 3 or more, 6 are greetings or closings. The analysis below is therefore on utterance texts, ignoring how many times these utterances appeared.[1]

5 Reliability

As a means of checking that the annotators had a similar understanding of the task, we calculated inter-annotator reliability using Krippendorff’s α (Krippendorff, 2004).[2] Reliability cannot be taken

[1] A more extensive study would have to look at the frequency of utterance texts and at the classification of text-identical user utterances to distinct dialogue acts when they occur in different contexts.

[2] Krippendorff’s α is a chance-corrected agreement coefficient, similar to the more familiar K statistic (Siegel and

Table 3: Inter-annotator reliability

	α	$A_o{}^{(a)}$	$A_e{}^{(a)}$
Dialogue act	0.489	0.545	0.109
Dialogue act type	0.502	0.585	0.166
Matches domain	0.383	0.741	0.580

[a] Krippendorff's α is defined in terms of observed and expected disagreement: $\alpha = 1 - D_o/D_e$. For expository purposes we have converted these into values representing observed and expected agreement: $A_o = 1 - D_o$, $A_e = 1 - D_e$.

as a measure of the reproducibility of the annotation procedure, since the annotators were not working from detailed written guidelines, and any shared understanding must therefore come from their previous experience. Rather, reliability is indicative of how straightforward the task is before implementing corrective measures such as detailed guidelines and domain and dialogue act improvements. Table 3 shows the results of the agreement study on three sets of data: the top row is the annotators' mapping of utterances to individual dialogue acts; the middle row is derived from the actual annotation by replacing each dialogue act with its type; and the bottom row treats "unknown" as one category and collapses all the other dialogue acts into a second category, marking a decision of whether the utterance fits at all to any of the existing dialogue acts.

Reliability was substantially above chance, though not as high as typically accepted norms; it can definitely be improved with clearer annotation guidelines (see section 6 below). An important source of disagreement was whether an utterance was a good enough match for an existing dialogue act: while observed agreement on this distinction is necessarily higher than on the dialogue act or dialogue act type, reliability (or chance-corrected agreement) is substantially lower, due to the fact that much higher agreement is expected by chance. Choosing the threshold for matching an utterance to a dialogue act is a known problem for the classifier, which uses a single threshold that represents the optimal balance between false positives (inappropriate matches above threshold) and false negatives (appropriate matches below threshold); the study shows that this is a difficult task for human judges as well. One judge marked 89 utterances as "unknown", another marked 79, while the third judge marked only 33 utterances as "unknown".

The study also shows that when annotators agreed on the dialogue act type, they typically also agreed on the on the dialogue act itself: observed agreement on dialogue act types is not much higher than on dialogue acts, and reliability (or chance-corrected agreement) shows an even smaller difference. To make the analysis simpler, we proceed with the analysis of the individual utterances using the dialogue act type alone.

6 Utterance analysis

A total of 72 user utterances were marked with an identical dialogue act type (other than "unknown") by all the annotators. These included some straightforward greetings (such as *Hello Amani*), compliments (*You have a beautiful home*), thanks (*Thank you that helps a lot*), closings (*Goodbye madam*), offers – both explicit (*I promise to keep this discussion secret*) and implicit (*Everything you tell me is in confidence*), and questions (*What is the name of the man with the large gun*). While these account for just under 30% of the total utterance types, this shows that the existing dialogue act representation already provides for substantial coverage of what users say.

Some additional disagreements are fairly easily fixed. There are 24 disagreements on question type, of which 15 include the phrase *do you know* or *can you tell/describe*, for example *Do you know the name of the sniper?* These are formally yes/no questions but carry the impact of a wh-question, and a cooperative positive response would provide the sought-after information; the difference between asking a *can you tell/do you know* question and a direct wh-question is that the former allows a "no" response (or a non-cooperative "yes"), whereas the latter requires a phrase or sentence as a response. However, in order to make communications clearer, our tactical questioning characters are designed to always give fuller answers than a simple yes or no, so the distinction is immaterial. We could extend the dialogue act representation to represent *can you tell/do you know* questions,

Castellan, 1988). Like K, α ranges from -1 to 1, where 1 signifies perfect agreement, 0 obtains when agreement is at chance level, and negative values show systematic disagreement. The main difference between α and K is that α takes into account the magnitudes of the individual disagreements; in this study we treated all disagreements as equivalent, so α is essentially equivalent to K except that α employs a small correction for sample size. For additional background, definitions and discussion of agreement coefficients, see Artstein and Poesio (2008).

but even though this type of question is rather frequent, distinguishing it from direct wh-questions would have little impact on the system, so a better guideline would be to treat these as wh-questions.

Other disagreements between question types are related to the domain specification. For example, the question *Have you seen him around lately* is clearly a yes/no question, but it is not an exact match to an existing dialogue act. The domain does specify the fact ⟨strange-man, last-seen, yesterday⟩, which all annotators found to be a close enough match to the user utterance. However, one annotator matched it to the wh-question derived from this fact (equivalent in meaning to "when did you last see him?"), whereas the two others matched it with the corresponding yes/no question (equivalent to "did you last see him yesterday?"). It is not clear what sort of guidelines would bring uniformity to this type of disagreements, but like the previous type, this is not expected to affect system performance.

Certain greetings were also the cause of disagreement that can probably be reconciled with more explicit annotation guidelines. There was confusion as to how to mark formulaic greetings which are literally questions (e.g. *How are you?*) or statements (*it's nice to meet you*). This can be solved through an explicit guideline to mark them as greetings, or by adding corresponding facts to the domain specification and matching these utterances to the literal dialogue acts. The first solution would be more useful for affecting the character's emotion and rapport (since she will understand these as greetings), while the second would allow more specific responses.

Other disagreements that can probably be alleviated to some extent result from confusion among the annotators about the distinctions between certain pairs of dialogue acts – accept and acknowledge, closing and pre-closing, request-repair and repeat-back. These, together with the greetings and questions discussed above, constitute 55 utterances; together with the utterances on which there is full agreement there are 127 user utterances (57% of all utterance types) which can be classified properly into dialogue acts using the current domain specifications.

The remaining user utterances are not covered by the existing dialogue acts. However, simple extensions can account for many of them. The most common utterances in this class are questions about an object but without a specific attribute, such as *Can you tell me about the shooter?* Our corpus contains 26 such questions, that is almost 12% of all question types. To deal with these utterances we added a new type of dialogue act – a wh-question with just an object and no attribute. These dialogue acts are generated automatically for all objects in the domain, and corresponding policies have been added to the dialogue manager.

An additional 16 user utterances (7%) are simply not in the domain: for example, the question *Do you own a gun?* does not have a corresponding fact, but it would be very easy to add one, and an appropriate dialogue act would be generated automatically. A small number of user questions cannot be represented through existing dialogue acts even though the relevant facts exist in the domain specification. For example, the user utterance *Can you tell me who lives on top of Assad's shop?* is fully answered by the fact ⟨strange-man, location, store⟩ – but we do not generate dialogue acts that ask which object has a known attribute and value. Since such questions are relatively rare in our corpus (only 4), we decided against generating this type of dialogue act, opting instead to represent the questions that do arise as independent facts, so the above fact is now also represented as ⟨the-shop, occupant, strange-man⟩. This is a compromise solution, because the character is not aware that these two facts in the domain are essentially identical in content. The advantage of this duplication of facts is in keeping the domain simple, without generating an inflated space of dialogue acts which are rarely encountered in practice.

Overall, almost 50 user utterances fall into the above classes – utterances that can be represented using the ⟨object, attribute, value⟩ scheme by either adding facts to the domain or extending the dialogue acts generated from these facts. Together with the utterances discussed previously, these account for nearly 80% of the user utterances.

The remaining utterances are a mixed bag. Sometimes a user asks Amani to clarify an elicitation request, as in *Which promises do you want to hear?* or *Are you worried about your safety?* The system used in the experiment had no corresponding dialogue acts, but these have since been added. Several compound utterances correspond to more than one dialogue act – the utterance *Amani, if I offer you and your family protection can you lead me to the sniper?* contains a conditional offer and a question. These will be dealt with using a separate utterance segmenter which is under development. Some utterances are inherently vague (perhaps intentionally). For example, when the user says *Your safety is very important to us* in response to a request for a guarantee of safety, it is not clear whether an offer has been made (there are 10 such utterances in our corpus). Some utterances contain rather obscure references; for example, in response to Amani's assertion that many Iraqis have guns, the user says *Wanna see mine?* which should probably be understood as a threat. The question *Can you tell me something useful?* was taken to be an insult by one annotator. One utterance, *Hello Mohammed*, is addressed to Amani's brother who is not an interactive character. Each of these types of utterances would require a different strategy in order to allow the character to understand it. Developing such capabilities for all of these utterances would be beyond the scope of the tactical questioning system, but this is not really necessary: there will always be some utterances that the character cannot understand, and the dialogue manager is designed to deal with this situation by providing off-topic responses or allowing the character to take initiative. The study shows that the vast majority of user utterances can be understood using the simple dialogue act representation language, and this is sufficient for tactical questioning characters.

7 Conclusion

This study has shown that from a simple representation of facts as ⟨object, attribute, value⟩ triples and actions as ⟨character, action⟩ pairs we can automatically generate dialogue acts that provide substantial coverage for interpreting user utterances spoken to a tactical questioning dialogue character. We have identified a few deficiencies in the dialogue act generation process, most notably requiring additional types of questions, which have been corrected in subsequent development. An extended system with an expanded domain and additional dialogue act types has been recently tested in the field with a large number of new users, and we are currently working on analyzing the results. We expect this new study to give a more accurate estimate of the proportion of user utterances covered by the representation.

One limitation that emerges from the current study is the linking of only one dialogue act per utterance, which makes it more difficult to capture the multifunctionality of dialogue. For example, many utterances which have an illocutionary effect such as greetings, threats, and insults can be phrased in the form of a question which may also be relevant in the domain. Some functions can be computed automatically from the main dialogue act applied to the context, but some inferences are more challenging and would be better served by labelling multiple acts directly, which would complicate both the authoring and annotation tasks. Representing multiple facets of such an utterance without implementing to a full inference chain which calculates implicatures and illocutionary force from literal meanings remains a challenge for future research.

Acknowledgments

We wish to thank Nicolle Whitman and Sarrah Ali, who helped in developing the initial tools and authoring the initial Amani scenario. Thanks also to 3 anonymous reviewers for their thoughtful comments and suggestions. We alone are responsible for any errors and omissions.

The project or effort described here 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

Jens Allwood. 1980. On the analysis of communicative action. In M. Brenner, editor, *The Structure of Action*. Basil Blackwell. Also appears as Gothenburg Papers in Theoretical Linguistics 38, Dept of Linguistics, Göteborg University.

Jens Allwood. 1995. An activity based approach to pragmatics. Technical Report (GPTL) 75, Gothenburg Papers in Theoretical Linguistics, University of Göteborg.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jim Barnett, Rahul Akolkar, R. J. Auburn, Michael Bodell, Daniel C. Burnett, Jerry Carter, Scott McGlashan, Torbjörn Lager, Mark Helbing, Rafah Hosn, T. V. Raman, and Klaus Reifenrath. 2008. State Chart XML (SCXML): State machine notation for control abstraction. http://www.w3.org/TR/scxml/, May.

Harry C. Bunt. 1999. Dynamic interpretation and dialogue theory. In Michael M. Taylor, F. Néel, , and Don G. Bouwhuis, editors, *The Structure of Multimodal Dialogue, Volume 2*. John Benjamins, Amsterdam.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, Massachusetts, November. AAAI.

Sudeep Gandhe, David DeVault, Antonio Roque, Bilyana Martinovski, Ron Artstein, Anton Leuski, Jillian Gerten, and David Traum. 2008. From domain specification to virtual humans: An integrated approach to authoring tactical questioning characters. In *proceedings of Interspeech 2008*, Brisbane, Australia, September.

Sudeep Gandhe, Nicolle Whitman, David Traum, and Ron Artstein. 2009. An integrated authoring tool for tactical questioning dialogue systems. In *6th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Pasadena, California, July.

Simon Keizer and Harry Bunt. 2006. Multidimensional dialogue management. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 37–45, Sydney, Australia, July. Association for Computational Linguistics.

Klaus Krippendorff, 2004. *Content Analysis: An Introduction to Its Methodology*, chapter 11, pages 211–256. Sage, Thousand Oaks, California, second edition.

Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In *Proccedings of 26th Army Science Conference*, Orlando, Florida, December.

Antonio Roque and David Traum. 2007. A model of compliance and emotion for potentially adversarial dialogue agents. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 35–38, Antwerp, September.

Antonio Roque and David Traum. 2009. Improving a virtual human using a model of degrees of grounding. In *Proceedings of IJCAI 2009*, Pasadena, CA, July.

Sidney Siegel and N. John Castellan, Jr, 1988. *Nonparametric Statistics for the Behavioral Sciences*, chapter 9.8, pages 284–291. McGraw-Hill, New York, second edition.

John McHardy Sinclair and Malcolm Coulthard. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford University Press.

David R. Traum and James F. Allen. 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Las Cruces, New Mexico, June. Association for Computational Linguistics.

David R. Traum and Elizabeth A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599. Special Issue on Non-literal language.

David R. Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, chapter 15, pages 325–353. Kluwer, Dordrecht.

David R. Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first International Joint conference on Autonomous Agents and Multiagent systems*, pages 766–773.

David Traum, Anton Leuksi, Antonio Roque, Sudeep Gandhe, David DeVault, Jillian Gerten, Susan Robinson, and Bilyana Martinovski. 2008. Natural language dialogue architectures for tactical questioning characters. In *Proceedings of 26th Army Science Conference*, Orlando, Florida, December.

David Traum. 2000. 20 Questions on Dialogue Act Taxonomies. *J Semantics*, 17(1):7–30.

U.S. Army. 2006. Police intelligence operations. Field Manual FM 3-19.50, U.S. Army. Appendix D: Tactical Questioning.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 515–522, Morristown, NJ, USA. Association for Computational Linguistics.

What we can learn from Dialogue Systems that don't work
On Dialogue Systems as Cognitive Models

David Schlangen
Department of Linguistics
University of Potsdam, Germany
das@ling.uni-potsdam.de

Abstract

In the 'real world', dialogue systems typically are made to work long days in call-centres of airlines and banks, fielding customer queries (and often inviting customer rage). In academia, a strong line of research is aimed at making such systems better at such tasks (in the hope of reducing customer annoyance). Here, I want to explore potential uses of spoken dialogue systems not as members of the workforce but in the lab, as a tool for the cognitive sciences. I argue that dialogue systems can be employed as situated, implemented computational models of language-capable agents; models whose predictions can be evaluated in real-time in ecologically valid settings, by human conversant. I sketch a methodology for building such models, propose areas where they can best be employed, and discuss the relations between research in this direction and more applied research.

1 Introduction

(Pieraccini and Huerta, 2005) recently noted that "there are three different lines of research in the field of spoken dialogue", one focusing on "understanding human communication, the second on designing the interface for usable machines, and the third on building those usable machines". Collapsing the latter two classes into one, we may label these views the *tool-for-understanding* and the *getting-things-done* approaches.[1]

Interestingly, (Pieraccini and Huerta, 2005) don't give any references for whom they see as representing the first line of research. And on closer inspection of the literature, there indeed seems to be little work in the dialogue systems community that would identify itself as belonging solely to the *tools-for-understanding* camp (it's a different matter in the embodied agents community).[2] In this paper, I'd like to explore the problems and potential of the *tool-for-understanding* direction and its relation to the *getting-things-done* camp.

The paper is structured as follows: First, I briefly review what computational cognitive models are and discuss how dialogue systems can be seen as a special class thereof. Then, I discuss a methodology for employing SDSs to address cognitive questions, and areas that seem particularly amenable to this methodology, given the current state of the technology. I then discuss a number of possible objections against the proposed use of dialogue systems. I close with some thoughts on the relation between the different uses for dialogue systems, and a general discussion.

2 Dialogue Systems as Cognitive Models

How can dialogue systems, with all their well-known technical problems and clumsy dialogue behaviour possibly function as models of cognitive abilities, and of which ones in any case? Before I address these questions, let us backtrack a bit and briefly review what cognitive models actually are.

2.1 Levels of Analysis in Cognitive Models

In the most abstract sense, a model in the cognitive sciences can be seen as a function from an agent's inputs to its outputs—typically, but not necessar-

[1] This of course reflects a classic dichotomy within the field of artificial intelligence which goes by many names: engineering vs. "empirical science concerned with the computational modeling of human intelligence" (Jordan and Russell, 1999); or, wrt. dialogue systems, "simulation" vs. "interface" (Larsson, 2005), or just simply applied vs. pure research.

[2] Recent examples of systems that seem to fall more on the *tool-for-understanding* side (but that do not make clear whether they see themselves as such) are (Allen et al., 1995; Allen et al., 2000; DeVault and Stone, 2009; Skantze and Schlangen, 2009).

ily, percepts and behaviours, respectively. In non-trivial cases, this function will depend in some way to the input (i.e., not be constant), and so can be seen as specifying an *information processor*.

As Marr (1982) pointed out in his seminal work on vision, such a function can be specified in different ways, which address different analytical interests; his classification is shown here in Table 1. A computational model is one which focuses on the problem that is being solved by the processor, i.e. only on the function in a mathematical sense. A representational model adds concerns about the exact way the processor computes the function; an implementational model also worries about the physical details of the processor.

A popular and fruitful recent line of research puts a further constraint on models on the computational level. With the, often tacit, assumption that natural behaviours have evolved to be near-optimal, they assume that agents act *rational*, i.e. that they solve their computational problems in an optimal way (minimizing their cost, maximizing their gain), given the available information (Anderson (1991), see also Chater and Oaksford (2008) for a recent overview). This direction has the advantage of offering a clear mathematical basis for computational modeling (probability theory, and more specifically Bayesian belief updating); we will discuss below to what extent it can support dialogue modeling.

Because it offers a convenient vocabulary to talk about inputs, outputs, and everything in between, we introduce here some central notions. The task of the agent can in such a model be stated clearly: it is to find that action a_t, given the observations of the world o_{t-1}, that has the best chance of bringing the world to a desired state s_{t+1}.

I now try to situate dialogue systems within this view of cognitive modelling.

2.2 Dialogue Systems: Situated Computational Whole-Agent Models

First, a few words on what I mean by "dialogue system". Often, the term is used specifically for mono-modal, *voice-only* systems that do rather limited practical tasks, and is used somewhat in opposition to *conversational agent* (seen as more capable, but less oriented towards practical applications), *multi-modal system* (with more modalities available to it) or *embodied conversational agents* (with a simulated or real "body", and consequently also more modalities). I do not intend such an opposition here, and use *dialogue system* to cover all these kinds of systems; the defining property here is that it is an (artificial) system that can enter into and hold some, perhaps limited, but in any case sustained form of (in the prototypical case) language-based interaction in real-time with a human. I will argue that for our purposes there are more commonalities between these different kinds of systems than is usually assumed, and that even the humblest kind of system (voice-only, not embodied) has to answer challenges that, depending on how and with which focus they are answered, can turn it into a cognitive model of an interesting type.

Now, what kind of analysis can dialogue systems offer, and of what? Let's first look at the task environment in which a dialogue agent finds itself. The information-processing task it needs to address is the quite substantial one of understanding language, and possibly a part of the world the conversation is about, well-enough to come up with a reaction, possibly in language as well, that is appropriate. (Note the restriction on *well-enough*; I will come back to this later.) This is the first step where dialogue systems can be usefully employed in cognitive modelling: building such a system forces one to precisely specify the task environment for (a particular setting of a) dialogue and the phenomenon of interest.

Given a particular conversational competence of interest (e.g., fast reaction times in turn-taking; more on possible modeling targets below in Section 3.2), a dialogue system can make, by embodying a computational model of them, theories of this competence testable. This property of making the predictions of a theory testable is something that dialogue systems of course share with any kind of computational model (for that is what dialogue systems are, to finally relate the discussion here to the previous section) in the cognitive sciences. However, they do this in an unusual way, by exposing themselves on-line to the situation type they are meant to model. With respect to the task of language processing, dialogue systems are *whole-agent models*: they need to say something about all levels of language processing (however many one assumes), from perceiving through understanding to generating it. This contrasts with the way for example theories of reading time are evaluated, namely against pre-collected

Computational Theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and the output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1: The three levels of analysis of information processing tasks of (Marr, 1982)

corpus data (see e.g. (Lewis and Vasishth, 2005)); these are what could be called *sub-module models*.

For us, this property of being a *whole-agent model* is the 'unique selling proposition' of dialogue systems as *tools-for-understanding*. As complete models (w.r.t. a certain ability, and other constraints that will be discussed presently) of the agent-type they are meant to model, they have to produce a much wider range of behaviours than sub-module models, and have to be explicit about how these behaviours arise from that of the sub-modules (assuming that they do have discernible sub-modules). This is a challenge that can hardly be addressed otherwise, as (Marr, 1982) noted: "Almost never can a complex system of any kind be understood as a simple extrapolation from the properties of its elementary components".

It's not only the range of modelled behaviour where dialogue systems can have an advantage over off-line models, though. The kind of phenomena that seem to be promising goals for tackling in a dialogue system understood as cognitive model (see next section) also seem hard to model and evaluate otherwise. Decisions of an agent in a dialogue (the a_t from Section 2.1) typically have delayed rewards (how good was the conversation), and complete models of the world (that is, models of how the actions of the agent change the state of the world, $P(s_t|s_{t-1}, a_t)$, and of how the world is perceived, $P(o_t|s_t)$) are generally not available and, given the size of the state space, hard to learn from data—all of which suggests interactive evaluation as a strategy that is more promising than for example trying to reproduce a gold-standard from a corpus.[3]

The on-line nature of this interaction finally makes dialogue systems an ideal tool for exploring ideas from another recent approach within the cognitive sciences: *situated* or *embodied cognition*: "the theory of situated cognition [...] claims that every human thought and action is adapted to the environment, that is *situated*, because what people *perceive*, how they *conceive of their activity*, and what they *physically do* develop together." (Clancey, 1997, p.1). On-line interactions with dialogue systems inevitably happen in contexts, in situations, embedded at the very least in time, if not in space, and the systems need to address such situational features.

Let's wrap up the discussion of which of Marr's levels dialogue systems cover. As elaborated above, dialogue systems clearly represent a computational analysis: they contain a specification of what it is that is being computed, what the components of that computation are, and what the goals are. They are also by definition *implemented*—although most dialogue systems do not make any claims about the cognitive plausibility of the representations and algorithms they use. Lastly, most likely dialogue systems will not any time soon be able to tell us anything about the physical realisation of conversational skills, and hence aren't models on the physical level.

This then concludes this section: in the view proposed here, Dialogue Systems are situated, implemented *whole-agent models* of human language processing capabilities, and are as such computational cognitive models, perhaps with partial claims to representational and algorithmic realism as well.

3 Methodology and Domains

3.1 Of Robotic Bees and Conversational Agents

In (Michelsen et al., 1992), an experiment is described that represents the culmination of years of research on communication among honeybees: To test their understanding of the communication methods used by honeybees, the researchers built a mechanical model of a forager bee, put

[3] Interestingly, in the line of research that uses Bayesian methods like Reinforcement Learning to solve Partially-Observable Markov Decision Processes (see Lemon and Pietquin (2007) for a recent overview), a middle position is taken: the systems learn by interacting with user models which generate the observations, and which in turn are learnt from data. In effect, this is what could be called a "semi-interactive" setting, where two implemented models converse which each other.

it in a typical communication situation (inside a beehive), and let it perform various forms of dances, implementing variants of the models of bee-communication which the researchers had previously built from observation. The effectiveness of the dances (and hence the adequacy of the theories) was then evaluated by the number of bees that as a result flew to the predicted (communicated) locations.

We envision a quite similar place for dialogue systems in the study of human communication, and a similar methodology: artificial agents embody a theory of communication, whose adequacy is evaluated through the reactions it provokes in a naturalistic setting. However, compared to the honeybee, human communicative situations are somewhat more varied, and there are interesting interactions between technical limitations on what can be computationally modelled and choice of situation. The appropriate methodology then looks more like the following: a) start from theory that says something about phenomenon you want to study; b) devise communicative setting that keeps this phenomenon as unrestricted as possible while restricting other aspects as much as possible; c) record humans in this setting; d) derive from this a more fine-grained model, which is e) implemented in computational model; f) evaluate the model not only for how well it reproduces the phenomenon but also for the reactions it provokes.[4] (In practice, of course several iterations of c) to f) may be necessary.)

We go through the most important steps in the following.

3.2 Choice of Setting

The processing of human language poses quite formidable technical challenges, and the extant realisations even only of the sub-modules typically seen to be involved in it (e.g., parsing, "understanding", generation) are miles away from achieving human-like performance. This seems to pose a problem: if the components are that bad, how can we expect the result of their connection to be anywhere near a usable model of human behaviour (as in, one that helps answer interesting questions)?

The answer is, we shouldn't. Or at least we shouldn't be expecting to be able to model *unrestricted, intelligent conversation*. It is unrealistic to expect dialogue systems to be able to model "intelligent conversation" *per se*, that is, to expect them to be able to give "intelligent" replies to all kinds of utterances. Luckily, there are two (not mutually exclusive) ways around this problem. One is to restrict the setting in such a way as to require "intelligent" (or, better, appropriate) replies only in a narrow domain that *can* be modelled. The other is to shift the focus to other features of dialogue: Dialogue is not just about saying and meaning the right things. It's also about saying the right things at the right moment, and about giving the right kinds of other, not directly task-related signals.

It seems then that, at least in the short term, the most promising areas for modeling in dialogue systems are not those of the dynamics of meaning in dialogue, but that of the dynamics of interaction (where it is an interesting open question as to how much these can be disassociated). To give a laundry list of possible areas in control of interactivity that come to mind: turn-taking, timely feedback, emotional feedback, alignment between conversants. Also promising seems the study of emergent behaviours, created by interactions (planned / controlled or not) of parallel processes.

When the phenomenon of interest is selected and explanatory theories are consulted or constructed, the next step is to devise a setting in which the model can be evaluated. The challenge I see here is to choose a situation that reduces as much as possible the demands on the technical components, while still being as much as possible ecologically valid. The goal here is to externalise and expose the limits that the system has (insofar as they aren't part of what one wants to study) and to turn them into constraints posed by the situation (task, setting). E.g., a dialogue system will have understanding problems (ASR, NLU), so it's a good idea to restrict the situation in such a way that the space of expected interactions gets smaller, and the restrictions are intuitively clear to the human interactant.

To give an example for such a strategy (although the authors do not explicitly phrase it like this): in (Skantze and Schlangen, 2009), a system is presented that investigates how human-like levels of interactivity / turn-taking speeds can be reached. To investigate this, the authors restrict the situation into which the system is put to dictation of number

[4] Steps c) to f) follow the methodology proposed by (Cassell, 2007) for the construction of Embodied Conversational Agents.

sequences. This is a task that is intuitively understandable to human conversation partners, while making technical tasks that are not the direct goal of the investigation easier. (ASR can be expected to perform better on such a limited vocabulary.)

A lot of the ingenuity of using dialogue systems to answer questions about human language use will lie in the choice of restricted, but understandable settings.

3.3 Operationalisation, Model Construction

Once the setting has been determined and the general predictions of the theory have been mapped to it, the next step is to operationalise the theory so that it can be modelled computationally. Using the vocabulary introduced above, the task is to determine the range of actions that the system is meant to be able to take, the observations that are to be expected, and the state of the world that is to be tracked. (An additional detail is whether uncertainty about any of these elements should be modelled as well.)

Forcing explicitness at this step already is something that dialogue systems can contribute to the study of human language use. A functioning computational model of an ability (say, turn-taking) shows at least that the information given to it (say, word sequences and prosodic information) contains enough information to solve the computational problem.

In most current dialogue systems, the function from observations to actions is specified procedurally, as the outcome of the combination (in a pipeline, or partially parallel) of various processing modules. This reflects on the one hand what is seen as the structure of the problem—linguistics has traditionally separated the task of language understanding into the “modules” of syntax / parsing, semantics / interpretation and pragmatics / understanding—and on the other hand simply good software engineering practice. It also allows a more tentative approach, where less needs to be explicitly stated about the structure of the problem than what would be needed in a purely rational approach. (This of course can also be viewed as a downside of this approach.) Finally, as briefly mentioned above, it often is hard to get data from which free parameters of a rational model could be learned, and so analytical models with symbolic rules provide more control over the algorithm.[5]

It should be noted here that for the level of computational modeling, none of these differences matter. What matters here is a clear understanding of the problem; rational or probabilistic models perhaps have an advantage here because they enforce a clearer statement. If one puts weight on differences in processing mode, one starts to enter the algorithmic / representational level; for this to matter with respect to the modeling task, one would then need to claim realism for one way of processing or the other. Here again dialogue systems promise to be a useful tool, by making testable claims of advantages of different implementation methods.

The goal of studying human communication by means of computational modeling also gives the system designer the freedom to not fully implement those processing modules that aren’t meant to be part of the model. For example, if the aim of the model is the study of discourse structure, and logical forms are required as input of the sub-module which is being tested, one could try a setting where a human “wizard” (Wooffitt et al., 1997) is in the loop—as long as this doesn’t change the interactional dynamics one is interested in. Alternatively, an “oracle” could be employed: in a setting where what the human user will talk about is known in advance, for example because the user is asked to perform certain tasks, this information can be given to some modules of the system (unbeknown to the user) like reference resolver, speech recognition etc. Or, a system that is meant to model interaction features can use ELIZA-like techniques for content-management. (Cf. the discussion of “micro-domains” in (Edlund et al., 2008); more on this below.)

3.4 Evaluation

The final step is to evaluate the system for how well it does its job of modeling the phenomenon (and, more generally, of being ‘human-like’). Evaluating dialogue systems is a difficult business, as has often been discussed (Walker et al., 1998; Edlund et al., 2008). The behaviour of a dialogue system is the result of the combination of many modules, and it is often difficult to ascertain which module’s performance contributed what—asking the human users directly will often not give meaningful results.

[5] But see (Miller et al., 1996; Lemon and Pietquin, 2007; Schuler et al., 2009 in press) for some attempts at (partially) non-modular, probabilistic systems.

For using dialogue systems as *tools-for-understanding*, we see three basic ways for evaluation (which can be used together): First, if one has an objective measure of the modelled phenomenon available, one can treat the resulting interactions of human subjects and dialogue system as a corpus, and can compare the relevant measures in this corpus with measures of corpora of human–human interaction. Second, one can use subjective measures (user questionnaires) to evaluate the impression the system made. If one want to avoid asking directly for the feature one wants to evaluate, an indirect approach can be chosen where the evaluation question is held constant ("did you find the interaction similar to one with a human partner?"), but the system is varied along the interesting dimension (i.e., is intentionally 'disabled' wrt. the modelled phenomenon). Third, one can play or show the finished interactions to other experiment participants and let them evaluate the naturalness (a so-called "overhearer evaluation", (Whittaker and Walker, 2006)).[6]

4 Possible Objections

"Creating a human-like dialogue system means creating an Artificial Intelligence, and creating an Artificial Intelligence is impossible!"[7]

There are two parts to this objection. We'll deal with the last one first. Is creating an AI possible?

The criticism in (Larsson, 2005), if I understand it correctly, seems to turn on the assumption (following Dreyfus (1992)) that "the background [necessary for understanding human language] is not formalizable". The claim is that this applies both to attempts at explicitly formalising such background (e.g., using databases of facts and logical calculi to reason over them) as well as to learning approaches, and that from this observation it follows that "computers will never achieve human-level language understanding". While the position I've been advocating here does not require any claim about the possibility of human-level language understanding (more on this in a minute), I'd still like to note that I do not find the conclusion compelling.

The basis of the criticism seems to be the symbol grounding problem (see e.g. (Harnard, 1990)), i.e. the problem of providing abstract symbols with external, real-world meaning. In a quite sweeping manner, (Larsson, 2005) sets the bar for entry into the club of grounded beings high, and counts among the experiences that are required for understanding human language "being born by parents, going through childhood and adolescence and growing up and learning personal responsibility, social interaction". I do not see how a convincing in-principle argument can be formed along these lines. Ultimately, this seems to me an empirical question, and, *pace* (Wittgenstein, 1984 1953), I'd wait until I encounter a talking lion before I conclude whether I understand it or not.

Which brings me to the first part of the objection. Does the question whether building a (human-level) AI is possible even matter? Clearly, free conversation requires intelligence. Turing (1950) famously proposed a conversational deception test (am I talking to man or machine?) as a test for intelligence. But, as discussed above, human language use is not restricted to holding free conversation (and convincing the conversational partner one is human)—language is also used in other settings, and there are other competences that can be dissociated from this, and can be studied and modelled independently.[8,9]

Evaluation of these competences then amounts to running what could be called *Particularised Turing Tests*: Can the system convince the user that it is (like) a human operating under some, possibly relatively strict, constraints? An example could be a setting where the conversational partner is only allowed to ask questions. Do the utterances still come with a good timing? (The evaluation of course does not have to be Turing test-style, ie. as deception; see above for evaluation methodology.) (Edlund et al., 2008) call such settings "micro-domains", and specify as evaluation goal whether the system can be taken "for a human by *some person*, under *some set of circumstances*".

"Cognitive Science is about making predictions, not engineering systems. Building dialogue systems is an engineering task."

While the spoken dialogue systems technology is far away from providing standard environments

[6] (Cassell, 2007) provides interesting anecdotal evidence of the use of this technique.

[7] A version of this objection has recently been raised in this forum (Larsson, 2005), and so we discuss is a bit more extensively here.

[8] To Larsson's (2005) credit, this is acknowledged in his criticism.

[9] Cf. the Practical Dialogue Hypothesis in (Allen et al., 2000): "The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general conversational competence."

like SOAR (Laird et al., 1987), components for example for ASR (e.g., Sphinx4, (Walker et al., 2004)) and dialogue managers (TRINDIkit, (Larsson and Traum, 2000)) are freely available. It is however true that considerable effort has to be spent on forming out of such components running systems into which one can build the models that are the primary interest. This can only get better if research groups start to share resources on a larger scale. Efforts to achieve this are currently underway (e.g., resources registry organised by SIGdial).

"You end up with bad cognitive science (too many compromises to get it to work at all) and bad engineering (too simple / useless domain)"

This is a serious objection. Attempting to use dialogue systems technology, which is still quite immature, can lead to making many compromises to just getting some form of reliable behaviour at all out the system. There is a danger of landing in a no-mans land, building a system that is neither particularly helpful in understanding the problems faced by human language processors or advances the state of technology. It is my opinion however that this can be avoided, and the methodology sketched above can help towards doing so.

"You need at the very least eyes, arms and legs to be cognitively plausible."

This is a (slightly caricaturising) summary of the central tenet of embodied cognition (Anderson, 2003). As mentioned above, I see dialogue systems as in any case *situated*, as they function in the same temporal environment as their conversation partner. When it comes to dealing with content, I am sympathetic with the view that grounding of symbols in percepts is a useful approach; however, as detailed above, not all of cognition having to do with language use is about content.

"People interact differently with machines and with humans, so machines have different computational problem to solve."

While there is evidence for the first part of the objection (Fischer, 2006), this also seem to depend on the metaphor with which human users enter into the interaction (Edlund et al., 2008). Moreover, in any case it is unlikely that human language users are even flexible enough to produce a *fundamentally* different kind of behaviour towards artificial conversational agents. The objection does however point out that it is important to frame the situation in which the model is evaluated carefully.

5 Dialogue Systems as Cognitive Models and as Computer Interfaces

Both Pieraccini and Huerta (2005) and Larsson (2005) point out that what we've called the *tool-for-understanding* and the *getting things done* approaches are complementary. In what sense, though? First, the differences. The directions answer to different constraints, to differences in what the free variables are. For cognitive models, the goal has to be human-like performance (wrt. the phenomenon being modelled), for practical system, the *primary* goal has to be efficiency and effectiveness wrt. to the task—human-likeness may or may not be a useful secondary goal. Consequently, the modeler in the *tool-for-understanding* view is free to choose a domain that lends itself best to an as-isolated-as-possible study of a phenomenon (see Section 3), while a researcher or practitioner building an applied system is free to implement behaviours that do not appear at all human-like.

So much for the differences. A common interest of course is to build components that help with language processing. Good speech recognition for example is as much a precondition for convincing computational models of language use as it is one for good practical systems. The overlap goes further, though. The already briefly mentioned work on POMDPs (Lemon and Pietquin, 2007) for example is, although being pursued more from an applied perspective, highly interesting also from a cognitive modeling perspective, as it uses techniques that can guarantee optimal computations.[10]

To conclude this section, I'd like to propose, with (Larsson, 2005), that "it would be good practice to explicitly state what the goals of a certain piece of research are", namely whether one wants to investigate human language use, using dialogue systems as a tool, or whether one wants to improve human–computer interaction.

6 Conclusions

In this paper, I have discussed the potential and possible problems of using spoken dialogue systems (ecumenically understood as all kinds of ar-

[10] Interestingly, there is some reservation against such methods from a commercial perspective (Paek and Pieraccini, 2008), where the additional constraint of provability of dialogue strategies seems to be important for customers who employ such systems.

tificial systems that can interact via spoken natural language) as models of (certain aspects of) human cognition. I have sketched a methodology for doing so, proposing that the main use of dialogue systems for now lies in how they can help being more explicit about one structures the tasks.

The models that can be built at the moment are rather crude and limited, and necessarily containing many simplifications. The hope is that combined efforts on practical systems and on systems built as *tools-for-understanding* can improve both kinds of systems, and help advance our understanding of human language use.

Acknowledgments: The author gratefully acknowledges support by DFG (Emmy Noether Programme).

References

James F. Allen, Lenhart K. Shubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesio, and David R. Traum. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI*, 7:7–48.

James F. Allen, Donna Byron, M. Dzikovska, George Ferguson, L. Galescu, and A. Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6(3).

John R. Anderson. 1991. The place of cognitive architecture in a rational analysis. In K. van Lehn, editor, *Architectures for Intelligence*, chapter 1, pages 1–24. Lawrence Erlbaum Associates, Hillsdale, N.J., USA.

Michael L. Anderson. 2003. Embodied cognition: A field guide. *Artificial Intelligence*, 149:91–130.

Justine Cassell. 2007. Body language: Lessons from the near-human. In Jessica Riskin, editor, *Genesis Redux: Essays in the History and Philosophy of Artificial Life*. Chicago University Press, Chicago, USA.

Nick Chater and Mike Oaksford, editors. 2008. *The Probabilistic Mind: Prospects for a Bayesian cognitive science*. Oxford University Press, Oxford, UK.

William J. Clancey. 1997. *Situated Cognition: On Human Knowledge and Computer Representation*. Cambridge University Press, Cambridge, UK.

David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 184–192, Athens, Greece, March. Association for Computational Linguistics.

Hubert Dreyfus. 1992. *What computers still can't do*. MIT Press, Boston, Massachusetts, USA.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50:630–645.

Kerstin Fischer. 2006. *What Computer Talk Is and Is not: Human-Computer Conversation as Intercultural Communication*. Linguistics – Computational Linguistics. AQ-Verlag, Saarbrücken, Germany.

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

Michael I. Jordan and Stuart Russell. 1999. Computational intelligence. In *The MIT Encyclopedia of the Cognitive Sciences*, pages lxxvi–xc. MIT Press, Cambridge, Massachusetts, USA.

John Laird, Allan Newell, and P. Rosenbloom. 1987. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64.

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3–4).

Staffan Larsson. 2005. Dialogue systems: Simulations or interfaces. In *Proceedings of* DIALOR, *the 9th Workshop on the Semantics and Pragmatics of Dialogue*, Nancy, France.

Oliver Lemon and Olivier Pietquin. 2007. Machine learning for spoken dialogue systems. In *Proceedings of Interspeech 2007*, Antwerp, Belgium.

R.L. Lewis and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco, USA.

Axel Michelsen, B. B. Anderson, J. Storm, W. H. Kirchner, and M. Lindauer. 1992. How honeybees perceive communication dances, studied by means of a mechanical model. *Behav. Ecol. Sociobiol.*, 30:143–150.

Scott Miller, David Stallard, Robert Brobow, and Richard Schwartz. 1996. A fully statistical approach to natural language interfaces. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, USA.

Tim Paek and Roberto Pieraccini. 2008. Automating spoken dialogue management design using machine learning: An industry perspective. *Speech Communication*, 50:716–729.

Roberto Pieraccini and Juan Huerta. 2005. Where do we go from here? research and commercial spoken dialog systems. In *Proceedings of the 6th SIGdial workshop on Discourse and Dialogue*, Lisbon, Portugal, September.

William Schuler, Stephen Wu, and Lane Schwartz. 2009 in press. A framework for fast incremental interpretation during speech decoding. *Computational Linguistics*.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 745–753, Athens, Greece, March.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 12(3).

Willi Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition. Technical Report SMLI TR2004-0811, Sun Microsystems Inc.

Steve Whittaker and Marilyn Walker. 2006. Evaluating dialogue strategies in multimodal dialogue systems. In Wolfgang Minker, D. Bühler, and Laila Dybkjaer, editors, *Spoken Multimodal Human–Computer Dialogue in Mobile Environments*. Springer, Den Haag, The Netherlands.

Ludwig Wittgenstein. 1984 [1953]. *Tractatus Logicus Philosophicus und Philosophische Untersuchungen*, volume 1 of *Werkausgabe*. Suhrkamp, Frankfurt am Main.

R. Wooffitt, N.M. Fraser, N. Gilber, and S. McGlashan. 1997. *Humans, Computers and Wizards*. Routledge, London and New York.

Compositional and ontological semantics in learning from corrective feedback and explicit definition

Robin Cooper
University of Gothenburg
Sweden
cooper@ling.gu.se

Staffan Larsson
University of Gothenburg
Sweden
sl@ling.gu.se

Abstract

We present some examples of dialogues from the literature on first language acquisition where children appear to be learning word meanings from corrective feedback and argue that in order to be able to account for them all in a formal theory of semantic change and coordination, we need to make a distinction between compositional and ontological semantics. We suggest how TTR (Type Theory with Records) can be used in making this distinction and relating the two kinds of semantics.

1 Introduction

This paper concerns the semantics and pragmatics of semantic coordination in dialogues between adults and children. The overall goal of this research is to attempt a formal account of language coordination in dialogue, and semantic coordination in particular.

In Larsson and Cooper (2009), we provide a dialogue move analysis of some examples from the literature on corrective feedback. We also provide a fairly detailed discussion of one example using TTR (Cooper, 2005; Cooper, 2008) to formalize concepts. In this paper we argue that in order to be able to account for these examples in a formal theory of semantic change and coordination, we need to make a distinction between compositional and ontological semantics. Both these aspects of meaning need to be represented in the linguistic resources available to an agent. We suggest how TTR can be used in making this distinction and relating the two kinds of semantics.

We take the following view on first language acquisition: children learn the meanings of expressions by observing and interacting with others. We regard language acquisition as a special case of a more general phenomenon of language coordination, that is, the process of coordinating on a language sufficiently to enable information sharing and coordinated action. One thing which is special about language acquistion is that there can be a clear assymmetry between the agents involved with respect to expertise in the language being acquired when a child and an adult interact. However, we want to propose that the mechanisms for semantic coordination used in these situations are similar to those which are used when competent adult language users coordinate their language.

Two agents do not need to share exactly the same linguistic resources (grammar, lexicon etc.) in order to be able to communicate, and an agent's linguistic resources can change during the course of a dialogue when she is confronted with a (for her) innovative use. For example, research on alignment shows that agents negotiate domain-specific microlanguages for the purposes of discussing the particular domain at hand (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Pickering and Garrod, 2004; Brennan and Clark, 1996; Healey, 1997; Larsson, 2007). We use the term *semantic coordination* to refer to the process of interactively coordinating the meanings of linguistic expressions.

We want a formal semantics allowing for meanings that can change dynamically during the course of a dialogue as a result of meaning updates triggered by dialogue moves. In particular, innovative uses of linguistic expressions may trigger updates to lexical entries. To account for this we need to account for how agents detect expressions which are innovative with respect to the agent's current linguistic resources, either because the expression is entirely new to the agent or because it is a known expression which is used with a new meaning.

We also need an account of how agents assign meanings to innovative expressions relative to the context of use. It is important here to distinguish local coordination on situated meanings, which is part of conversational grounding (Clark and Brennan, 1990; Traum, 1994) from coordination on meanings which affects agent resources such as lexical entries. It is the latter that we are interested in here.

Finally, we need to account for how the lexicalised meaning of a non-innovative expression can be updated based on its previously assumed meaning and the meaning of an innovative use which contrasts with it. For example, if we learn that an object is not an A but rather a B (where B is innovative for us) then we need not only to learn B but also to refine the meaning of A so that it does not apply to the object.

In the rest of this paper we will first present a view of how agents adjust their linguistic resources on the basis of dialogue interaction (section 2). We will then discuss how compositional semantics can be derived from corrective feedback (section 3) and then give a brief background to TTR (section 4). Finally, we will show how ontological semantics can be added to compositional semantics derived from corrective feedback and explicit definition.

2 Agents that coordinate linguistic resources

As in the information state update approach in general (Larsson and Traum, 2000), dialogue moves are associated with information state updates. For semantic coordination, the kind of update is rather different from the one associated with dialogue moves for coordinating on task-related information, and involves updating the linguistic resources available to the agent (grammar, lexicon, semantic interpretation rules etc.), rather than e.g. the conversational scoreboard as such. Our view is that agents do not just have monolithic linguistic resources as is standardly assumed. Rather they have generic resources which they modify to construct local resources for sublanguages for use in specific situations. Thus an agent A may associate a linguistic expression e with a particular concept (or collection of concepts if e is ambiguous) $[\mathrm{e}]^A$ in its generic resource. In a particular domain α e may be associated with a modified version of $[\mathrm{e}]^A$, $[\mathrm{e}]^A_\alpha$ (Larsson, 2007). In some cases $[\mathrm{e}]^A_\alpha$ may contain a smaller number of concepts than $[\mathrm{e}]^A$, representing a decrease in ambiguity.

Particular concepts in $[\mathrm{e}]^A_\alpha$ may be a *refinement* of one in $[\mathrm{e}]^A$, that is, the domain related concepts have an extension which is a proper subset of the extension of the corresponding generic concept. This will, however, not be the case in general. For example, a black hole in the physics domain is not normally regarded as an object described by the generic or standard meaning of *black hole* provided by our linguistic resources outside the physical domain. Similarly a variable in the domain of logic is a syntactic expression whereas a variable in experimental psychology is not and quite possibly the word *variable* is not even a noun in generic linguistic resources.

Our idea is that the motor for generating new local resources in an agent lies in coordinating resources with another agent in a particular communicative situation s. The event s might be a turn in a dialogue, as in the examples we are discussing in this paper, or, might, for example, be a reading event. In a communicative situation s, an agent A may be confronted with an *innovative* utterance e, that is, an utterance which either uses linguistic expressions not already present in A's resources or linguistic expressions known by A but associated with an interpretation distinct from that provided by A's resources. At this point, A has to accommodate an interpretation for e which is specific to s, $[\mathrm{e}]^A_s$, and which may be anchored to the specific objects under discussion in s.

Whereas in a view of semantics inherited from formal logic there is a pairing between a linguistic expression e and an interpretation e' (or a set of several interpretations if e is ambiguous), we want to see e as related to several interpretations: $[\mathrm{e}]^A_s$ for communicative situations s, $[\mathrm{e}]^A_\alpha$ for domains α (where we imagine that the domains are collected into a complex hierarchy or more and less general domains) and ultimately a general linguistic resource which is domain independent, $[\mathrm{e}]^A$. We think of the acquisition of a pairing of an expression e with an interpretation e' as a progression from an instance where e' is $[\mathrm{e}]^A_s$ for some particular communicative situation s, through potentially a series of increasingly general domains α where e' is regarded as being one of the interpretations in $[\mathrm{e}]^A_\alpha$ and finally arriving at a state where e' is associated with e as part of a domain independent generic resource, that is, e' is in $[\mathrm{e}]^A$.

There is no guarantee that any expression-interpretation pair will survive even beyond the particular communicative situation in which A first encountered it. For example, the kind of *ad hoc* coinages described in Garrod and Anderson (1987) using words like *leg* to describe part of an oddly shaped maze in the maze game probably do not survive beyond the particular dialogue in which they occur. The factors involved in determining how a particular expression-interpretation pair progresses we see as inherently stochastic with parameters including the degree to which A regards their interlocutor as an expert, how many times the pairing has been observed in other communicative situations and with different interlocutors, the utility of the interpretation in different communicative situations, and positive or negative feedback obtained when using the pairing in a communicative situation. For example, an agent may only allow a pairing to progress when it has been observed in at least n different communicative situations at least m of which were with an interlocutor considered to be an expert, and so on. We do not yet have a precise proposal for a theory of these stochastic aspects but rather are seeking to lay the groundwork of a semantic treatment on which such a theory could be built.

3 Learning compositional semantics from corrective feedback

Recent research on first language acquisition (Clark, 2007; Clark and Wong, 2002; Saxton, 1997; Saxton, 2000) argues that the learning process crucially relies on negative input, including corrective feedback. This research is often presented in the context of the discussion of negative evidence, which we believe plays an important role in language. However, we want to relate corrective feedback to the discussion of alignment. We see corrective feedback as part of the process of negotiation of a language between two agents. Here are the examples of corrective feedback that we discuss in connection with our argument for this position in Larsson and Cooper (2009):

"Gloves" example (Clark, 2007):

- Naomi: mittens
- Father: **gloves**.
- Naomi: gloves.
- Father: when they have fingers in them they are called gloves and when the fingers are all put together they are called mittens.

Panda example (constructed)

- A: That's a nice bear.
- B: Yes, it's a nice **panda**.

"Turn over" example (Clark and Wong, 2002):

- Abe: I'm trying to tip this over, can you tip it over? Can you tip it over?
- Mother: Okay I'll **turn** it over for you.

A frequent pattern in corrective feedback is the following:

original utterance A says something

innovative utterance B says something parallel to A's utterance, containing a use which is innovative for A

learning step A learns from the innovative use

The learning step can be further broken down as follows:

1. Syntactically align innovative utterance with original utterance
2. Use alignment to predict syntactic and semantic properties of innovative use
3. Integrate innovative element into *local grammar/lexicon* and *local ontology*.
4. Gradually refine syntactic and semantic properties of innovative use and incorporate into more general *linguistic resources* and more general *ontologies*.

We think that an important component in corrective feedback of this kind is syntactic alignment, that is, alignment of the correcting utterance with the utterance which is being corrected. This is a rather different sense of alignment than that associated with the negotiation of a common language, although the two senses are closely linked. By "syntactic alignment" here we mean something related to the kind of alignment that is used in parallel corpora. It provides a way of computing parallelism between the two utterances. Syntactic alignment may not be available in all cases but when it is, it seems to provide an efficient way of identifying what the target of the correction is.

Syntactic alignment in the gloves example can be visually represented thus:

Naomi: mittens
|
Father: **gloves**

For the "panda" example, the corresponding representation is

A: That's a nice bear
| | |
B: Yes, it's a nice **panda**

Finally, in the the "turn over" example:

Abe: Can you tip it over
| | |
Mother: Okay I'll **turn** it over for you

We assume that in the "gloves" example, syntactic properties can be predicted from syntactic alignment:

Naomi: [$_N$ mittens]
↓
Father: [$_N$ **gloves**]

In the "panda" example, the syntactic category of the innovative expression *panda* can be predicted from alignment (*panda* is aligned with non-innovative *bear* which is known to be a noun). This conclusion could be confirmed by an active chart edge spanning the substring *a nice* analyzed as an NP needing a noun. More confirming information can be extracted from the parse chart by noting that the assumption that *panda* is a noun allows us to complete an NP-structure parallel to the analysis of *a nice bear* with which it is aligned.

$$\begin{array}{lllll} \text{A: That's} & [_{\text{NP}} & [_{\text{Det}}\ \text{a}] & [_{\text{A}}\ \text{nice}] & [_{\text{N}}\ \text{bear}]] \\ & | & | & | & \downarrow \\ \text{B: Yes, it's} & [_{\text{NP}} & [_{\text{Det}}\ \text{a}] & [_{\text{A}}\ \text{nice}] & [_{\text{N}}\ \textbf{panda}]] \end{array}$$

Active edge: NP → $[_{\text{Det}}$ a] $[_{\text{A}}$ nice] • N

In the "turn over" example, evidence comes from alignment and the resulting passive edge (together with alignment) as in the panda-example. In this case, however, given normal assumptions about how the parsing works, there will not be an active edge available to confirm the hypothesis as there was in the panda-example.

$$\begin{array}{llllll} \text{Abe: Can you} & [_{\text{VP}} & [_{\text{V}}\ \text{tip}] & \text{it} & \text{over}\ ] \\ & | & \downarrow & | & | \\ \text{Mother: Okay I'll} & [_{\text{VP}} & [_{\text{V}}\ \textbf{turn}] & \text{it} & \text{over}\ ] \end{array}$$

for you

A possible hypothesis is that alignment evidence is primary in predicting syntactic properties of innovations when it is available (as it is in corrective feedback). Other evidence can be used to support or refute the analysis deriving from alignment.

Following Montague (1974) and Blackburn and Bos (2005) compositional semantics can be predicted from syntactic information such as category. For example, for common nouns we may use the formula

$$\text{commonNounSemantics}(N) = \lambda x N'(x)$$

or, using TTR,

$$\text{commonNounSemantics}(N) = \\ \lambda r{:}\big[\text{x} : \textit{Ind}\big]\ (\big[\text{e} : N'(r.\text{x})\big])$$

Thus, we see how compositional semantics can be derived from corrective feedback in dialogue. However, compositional semantics of this kind does not reveal very much, if anything, about the details of word semantics unless we add ontological information. Before we proceed to ontological semantics we shall give a brief background on some aspects of TTR.

4 TTR

The received view in formal semantics (Kaplan, 1979) assumes that there are abstract and context-independent "literal" meanings (utterance-type meaning; Kaplan's "character") which can be regarded formally as functions from context to content; on each occasion of use, the context determines a specific content (utterance-token meaning). Abstract meanings are assumed to be static and are not affected by language use in specific contexts. Traditional formal semantics is thus ill-equipped to deal with semantic coordination, because of its static view of meaning.

We shall make use of type theory with records (TTR) as characterized in Cooper (2005; 2008) and elsewhere. The advantage of TTR is that it integrates logical techniques such as binding and the lambda-calculus into feature-structure like objects called record types. Thus we get more structure than in a traditional formal semantics and more logic than is available in traditional unification-based systems. The feature structure like properties are important for developing similarity metrics on meanings and for the straightforward definition of meanings modifications involving refinement and generalization. The logical aspects are important for relating our semantics to the model and proof theoretic tradition associated with compositional semantics. Below is an example of a record type:

$$\left[\begin{array}{lll} \text{ref} & : & \text{Ind} \\ \text{size} & : & \text{size(ref, MuchBiggerThanMe)} \\ \text{shape} & : & \text{shape(ref, BearShape)} \end{array}\right]$$

A record of this type has to have fields with the same labels as those in the type. (It may also include additional fields not required by the type.) In place of the types which occur to the right of ':' in the record type, the record must contain an object of that type. Here is an example of a record of the above type:

$$\left[\begin{array}{lll} \text{ref} & = & \text{obj123} \\ \text{size} & = & \text{sizesensorreading85} \\ \text{shape} & = & \text{shapesensorreading62} \\ \text{colour} & = & \text{coloursensorreading78} \end{array}\right]$$

Thus, for example, what occurs to the right of the '=' in the ref field of the record is an object of type *Ind*, that is, an individual. Types which are constructed with predicates like *size* and *shape* are sometimes referred to as "types of proof". The idea is that something of this type would be a proof that a given individual (the first argument) has a certain size or shape (the second argument). One can have different ideas of what kind of objects count as proofs. Here we are assuming that the proof-objects are readings from sensors. This is a second way (in addition to the progression of local resources towards general resources) that our theory interfaces with an analogue non-categorical world. We imagine that the mapping from sensor readings to types involves sampling of analogue data in a way that is not unsimilar to the digitization process involved, for example, in speech recognition. Again we have nothing detailed to say about this at the moment, although we regard it as an important part of our theory that it is able to make a connection between the realm of feature vectors and the realm of model-theoretic semantics.

Types constructed with predicates may also be *dependent.* This is represented by the fact that arguments to the predicate may be represented by labels used on the left of the ':' elsewhere in the record type. This means, for example, that in considering whether a record is of the record type, you will need to find a proof that the object which is in the ref-field of the record has the size represented by *MuchBiggerThanMe.* That is, this type depends on the value for the ref-field.

Some of our types will contain *manifest fields* (Coquand et al., 2004) like the ref-field in the following type:

$$\begin{bmatrix} \text{ref=obj123} & : & \text{Ind} \\ \text{size} & : & \text{size(ref, MuchBiggerThanMe)} \\ \text{shape} & : & \text{shape(ref, BearShape)} \end{bmatrix}$$

$\left[\text{ref=obj123:Ind}\right]$ is a convenient notation for $\left[\text{ref} : \text{Ind}_{\text{obj123}}\right]$ where $\text{Ind}_{\text{obj123}}$ is a *singleton type.* If $a : T$, then T_a is a singleton type and $b : T_a$ (i.e. b is of type T_a) iff $b = a$. Manifest fields allow us to progressively specify what values are required for the fields in a type.

An important notion in this kind of type theory is that of *subtype.* For example,

$$\begin{bmatrix} \text{ref} & : & \text{Ind} \\ \text{size} & : & \text{size(ref, MuchBiggerThanMe)} \end{bmatrix}$$

is a subtype of

$$\begin{bmatrix} \text{ref} & : & \text{Ind} \end{bmatrix}$$

as is also

$$\begin{bmatrix} \text{ref=obj123} & : & \text{Ind} \end{bmatrix}$$

5 Learning ontological semantics from corrective feedback and explicit definition

As a (modest) "proof of concept" of our approach, we will in this section provide a TTR analysis of updates to compositional and ontological semantics for the "mittens" example above. As pointed out by one of the reviewers, our approach to coordination of ontological semantics bears resemblances to work on ontology mapping and ontology negotiation on the semantic web (van Diggelen et al., 2007).

Using TTR, we can formalise ontological classes as record types:

$$\text{Thing} = \begin{bmatrix} \text{x} : \textit{Ind} \end{bmatrix}$$

$$\{\text{Class } P\} = \begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_P : P(\text{x}) \end{bmatrix}$$

We will use a function SubClass which creates a class based on a predicate P:

{SubClass C_1 C_2} = $C_1 \underset{.}{\wedge} C_2$ ("Make a subclass of C_2 based on C_1")
The $\underset{.}{\wedge}$ operator is characterized as follows. Suppose that we have two record types C_1 and C_2:

$$C_1 = \begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\text{clothing}} : \text{clothing(x)} \end{bmatrix}$$

$$C_2 = \begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\text{physobj}} : \text{physobj(x)} \end{bmatrix}$$

$C_1 \wedge C_2$ is a type. In general if T_1 and T_2 are types then $T_1 \wedge T_2$ is a type and $a : T_1 \wedge T_2$ iff $a : T_1$ and $a : T_2$. A meet type $T_1 \wedge T_2$ of two record types can be simplified to a new record type by a process similar to unification in feature-based systems. We will represent the simplified type by putting a dot under the symbol $\wedge$. Thus if T_1 and T_2 are record types then there will be a type $T_1 \underset{.}{\wedge} T_2$ equivalent to $T_1 \wedge T_2$ (in the sense that a will be of the first type if and only if it is of the second type).

$$C_1 \underset{.}{\wedge} C_2 = \begin{bmatrix} \text{x} : \textit{Ind} \\ \text{c}_{\text{physobj}} : \text{physobj(x)} \\ \text{c}_{\text{clothing}} : \text{clothing(x)} \end{bmatrix}$$

The operation $\underset{.}{\wedge}$ corresponds to unification in feature-based systems and its definition (which we omit here) is similar to the graph unification algorithm.

Given this formal apparatus, we can show how ontological semantics properties can be predicted in the glove example. Naomi's pre-gloves ontology contains (we assume) the following:

PhysObjClass = {Class physobj}
ClothingClass = {SubClass {Class clothing} PhysObjClass}
MittenClass = {SubClass {Class mitten} ClothingClass}

This ontology is shown in Figure 1, where the arrow represents the subclass relation. Provided that Naomi learns from the interaction, Naomi's post-gloves ontology may include the following (see also Figure 2):

PhysObjClass = {Class physobj}
ClothingClass = {SubClass {Class clothing} PhysObjClass}
MittenClass = {SubClass {Class mitten} ClothingClass}

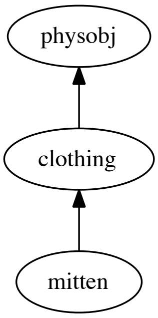


Figure 1: Naomi's "pre-gloves" ontology

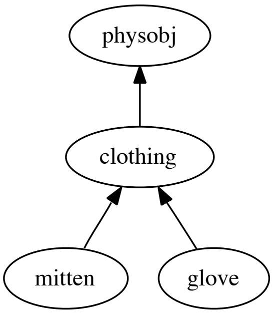


Figure 2: Naomi's "post-gloves" ontology

GloveClass = {SubClass {Class glove} ClothingClass} (from alignment of *mittens* and *gloves*)

This means that the glove class is the following type

$$\begin{bmatrix} \text{x} & : & \mathit{Ind} \\ \text{c}_{\text{physobj}} & : & \text{physobj(x)} \\ \text{c}_{\text{clothing}} & : & \text{clothing(x)} \\ \text{c}_{\text{glove}} & : & \text{glove(x)} \end{bmatrix}$$

which can be used as a refinement of the type corresponding to the compositional semantics:

$$\text{GloveCompSem} = \begin{bmatrix} \text{x} & : & \mathit{Ind} \\ \text{c}_{\text{glove}} & : & \text{glove(x)} \end{bmatrix}$$

Thus we can obtain the new function below as a refined compositional semantics:

$$\lambda r{:}\big[\text{x} : \mathit{Ind}\big]\ (\begin{bmatrix} \text{c}_{\text{physobj}} : \text{physobj}(r\text{.x}) \\ \text{c}_{\text{clothing}} : \text{clothing}(r\text{.x}) \\ \text{c}_{\text{glove}} : \text{glove}(r\text{.x}) \end{bmatrix})$$

In the "glove" example, the father's second utterance contains a partial but explicit definition of the ontology of gloves and mittens:

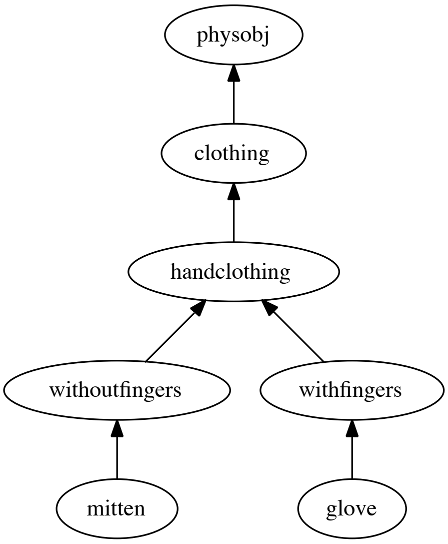


Figure 3: Naomi's ontology after explicit definition

- Father: when they have fingers in them they are called gloves and when the fingers are all put together they are called mittens.

When integrating this utterance, Naomi may modify her take on the ontological semantics (see also Figure 3):

PhysObjClass = {Class physobj}
ClothingClass = {SubClass {Class clothing} PhysObjClass}
HandClothingClass = {SubClass {Class handclothing} ClothingClass}
WithFingersClass = {SubClass {Class withfingers} HandClothingClass}
WithoutFingersClass = {SubClass {Class withoutfingers} HandClothingClass}
MittenClass = WithoutFingersClass
GloveClass = WithFingersClass

In TTR, after this update the meanings for "glove" and "mitten" will be respectively:

$$\begin{bmatrix} \text{x} : \mathit{Ind} \\ \text{c}_{\text{physobj}} : \text{physobj(x)} \\ \text{c}_{\text{clothing}} : \text{clothing(x)} \\ \text{c}_{\text{handclothing}} : \text{handclothing(x)} \\ \text{c}_{\text{withoutfingers}} : \text{withoutfingers(x)} \\ \text{c}_{\text{glove}} : \text{glove(cntxt.x)} \end{bmatrix}$$

and

$$\left[\begin{array}{l} \mathrm{x} : \mathit{Ind} \\ c_{\text{physobj}} : \text{physobj(x)} \\ c_{\text{clothing}} : \text{clothing(x)} \\ c_{\text{handclothing}} : \text{handclothing(x)} \\ c_{\text{withfingers}} : \text{withfingers(x)} \\ c_{\text{mitten}} : \text{mitten(cntxt.x)} \end{array}\right]$$

6 Conclusion and future work

By providing a basic compositional semantic resource and providing the ability to refine this with local ontologies, which may be associated with given domains or even specific dialogues, we allow for an extremely flexible view of word meaning that provides mechanisms for associating a central core of meaning with situation specific meanings that can be generated on the fly.

Future work includes exploring the relation to work on ontology negotiation on the semantic web, as well as extending our account to cover further aspects of meaning, including perceptually grounded meaning and connotations. We also wish to relate detailed accounts of semantic updates to other kinds of dialogue strategies, such as ostensive definitions and meaning accommodation (Larsson, 2008).

Acknowledgments

This research was supported by The Swedish Bank Tercentenary Foundation Project P2007/0717, Semantic Coordination in Dialogue.

References

Patrick Blackburn and Johan Bos. 2005. *Representation and Inference for Natural Language: A First Course in Computational Semantics.* CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

S. E. Brennan and H. H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:482–493.

H. H. Clark and S. E. Brennan. 1990. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*, pages 127 – 149. APA.

H. H. Clark and D. Wilkes-Gibbs. 1986. Refering as a collaborative process. *Cognition*, 22:1–39.

Eve V. Clark and Andrew D. W. Wong. 2002. Pragmatic directions about language use: Offers of words and relations. *Language in Society*, 31:181–212.

E. V. Clark. 2007. Young children's uptake of new words in conversation. *Language in Society*, 36:157–82.

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

Robin Cooper. 2008. Type theory with records and unification-based grammar. In Fritz Hamm and Stephan Kepser, editors, *Logics for Linguistic Structures*, pages 9 – 34. Mouton de Gruyter.

Thierry Coquand, Randy Pollack, and Makoto Takeyama. 2004. A logical framework with dependently typed records. *Fundamenta Informaticae*, XX:1–22.

Simon C. Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, 27:181–218.

P.G.T. Healey. 1997. Expertise or expertese?: The emergence of task-oriented sub-languages. In M.G. Shafto and P. Langley, editors, *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 301–306.

D. Kaplan. 1979. Dthat. In P. Cole, editor, *Syntax and Semantics v. 9, Pragmatics*, pages 221–243. Academic Press, New York.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In Afra Alishahi, Thierry Poibeau, and Aline Villavicencio, editors, *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, EACL*, pages 1–9.

Staffan Larsson and David Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *NLE Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*, pages 323–340.

Staffan Larsson. 2007. Coordinating on ad-hoc semantic systems in dialogue. In R. Artstein and L. Vieu, editors, *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue*, pages 109–116.

Staffan Larsson. 2008. Formalizing the dynamics of semantic systems in dialogue. In Robin Cooper and Ruth Kempson, editors, *Language in Flux - Dialogue Coordination, Language Variation, Change and Evolution*, pages 121–142. College Publications, London.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague.* Yale University Press, New Haven. ed. and with an introduction by Richmond H. Thomason.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–226, April.

Matthew Saxton. 1997. The contrast theory of negative input. *Journal of Child Language*, 24:139–161.

Matthew Saxton. 2000. Negative evidence and negative feedback: immediate effects on the grammaticality of child speech. *First Language*, 20(3):221–252.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Department of Computer Science, University of Rochester. Also available as TR 545, Department of Computer Science, University of Rochester.

Jurriaan van Diggelen, Robbert-Jan Beun, Frank Dignum, Rogier M. van Eijk, and John-Jules Meyer. 2007. Ontology negotiation in heterogeneous multi-agent systems: The anemone system. *Applied Ontology*, 2:267–303.

How mechanistic can accounts of interaction be?

Ruth Kempson[†], **Eleni Gregoromichelaki**[†], **Matt Purver**[‡],
Greg Mills[†], **Andrew Gargett**[‡], **Chris Howes** [‡]
[†]King's College London, Strand, London WC2R 2LS, UK
{ruth.kempson,eleni.gregor, andrew.gargett}@kcl.ac.uk
[‡]Queen Mary University of London, Mile End Road, London E1 4NS, UK
{mpurver,gj,chrizba}@dcs.qmul.ac.uk

Abstract

Ever since dialogue modelling first developed relative to broadly Gricean assumptions about utterance interpretation (Clark, 1996), it has been questioned whether the full complexity of higher-order intention computation is made use of in everyday conversation. In this paper, building on the DS account of *split utterances*, we further probe the necessity of full-intention recognition/formation: we do so by exploring the extent to which the interactive coordination of dialogue exchange can be seen as emergent from mechanisms of language processing, without either needing representation by interlocutors of each other's mental states, or fully developed intentions as regards messages to be conveyed (even in e.g. clarifications and completions when the content of the utterance is in doubt).

1 Introduction

The pioneering work of H. Clark (Clark, 1996) initiated a broadly Gricean program for dialogue modelling, in which coordination in dialogue is said to be achieved by establishing recognition of speaker-intentions relative to what each party takes to be their mutually held beliefs (*common ground*). However, computational models in this vein have very largely been developed without explicit high-order meta-representations of other parties' beliefs or intentions, except where dealing with highly complex dialogue domains (e.g. non-cooperative negotation (Traum et al., 2008)) or phenomena (e.g. collaborative completions (Poesio and Rieser, to appear)). With concepts such as *dialogue gameboard*, *QUD*, (Ginzburg, 1995; Larsson, 2002) and *settledness* (Asher and Gillies, 2004) largely replacing intention recognition, it is arguable that the Gricean assumptions underpinning communication should be re-considered. A parallel weakening has been taking place within another major pragmatic paradigm, that of (Sperber and Wilson, 1986). The relevance-theoretic view is that the content of an utterance is established by a hearer relative to what the speaker could have intended (relative also to a concept of mutual manifestness of background assumptions). However, (Breheny, 2006) argued that children in the initial stages of language acquisition communicate relative to a weaker 'naive-optimism' view in which some context-established interpretation is simply presumed to match the speaker's intention, only coming to communicate in the full sense substantially later (see (Tomasello, 2008) for a Gricean variant of this view).

With this weakening across all pragmatic models of the status of recognition of other interlocutor's intentions,[1] for at least some cases of communication, in this paper we set out the groundwork for an interactive model of communication using *Dynamic Syntax* (DS: Cann et al. (2005)), and examine its application to the tightly interactive dialogue phenomena that arise in cases of continuative/clarificatory/reformulatory splits among speakers. In this model, each party to the dialogue interprets the signals they receive, or plans the signals they send, egocentrically relative to their own context, without explicit (meta-)representation of the other party's knowledge/beliefs/intentions. Nevertheless, the effect of coordinated communication is achieved by relying on ongoing feedback between parties and the goal-directed action-based architecture of the grammar.

Our claim is that communication involves taking risks: in all cases where a single agent's system fails to fully determine choices to be made (either in parsing or production), the eventual choice may happen to be right, and might or might not get acknowledgement; it may be wrong and potentially get corrected, thereafter establishing explicit coordination with respect to some subpart of the communication; or, in recognition of the non-determinism, the agent may set out a sub-routine of clarification thereby delegating the choice of construal to the interlocutor before proceeding. Otherwise, a wrong choice which is uncorrected

[1] see also (Kecskes and Mey, 2008)

might threaten the viability of the exchange. Success in communication thus involves clarification/correction/extension/reformulation etc ("repair strategies") as essential subparts of the exchange. When modelled non-incrementally, such strategies might lead to the impression of non-monotonic repair and the need to revise established context. But pursued incrementally within a goal-directed architecture, these do not constitute communication breakdown and repair, but the normal mechanism of hypothesised update, context selection, and confirmation. By building on the assumption that successful communication may crucially involve subtasks of repair (see also (Ginzburg, forthcmg)), the mechanisms for informational update that underpin interaction can be defined without any reliance on (meta-) representing contents of the interlocutors' mental states.

2 Split Utterances

Switching of roles between speaking and hearing, across and within sentential structures, is characteristic of dialogue. People show a surprising facility to switch between speaker and hearer roles even mid-utterance:

(1) Daughter: Oh here dad, a good way to get those corners out
Dad: is to stick yer finger inside.
Daughter: well, that's one way. [from Lerner (1991)]

(2) A: They X-rayed me, and took a urine sample, took a blood sample. Er, the doctor
B: Chorlton?
A: Chorlton, mhm, he examined me, erm, he, he said now they were on about a slight [shadow] on my heart. [BNC: KPY 1005-1008]

(3) A: Are you left or
B: Right-handed.

The challenge of modelling such phenomena has recently been taken up by (Poesio and Rieser, to appear) (*P&R* henceforth) for German, defining a admirably fine-grained neo-Gricean model of dialogue interactivity that builds on an LTAG grammar base. Their primary aim is to model *completions*, as in (1) and (3), with take-over by the hearer because the remainder of the utterance is taken to be understood or inferrable from mutual knowledge. Their account hinges on two main areas: the assumption of recognition of interlocutors' intentions according to shared joint plans, and the use of incremental grammatical processing based on LTAG. However, their account relies on the assumption of a string-based level of syntactic analysis, for it is this which provides the top-down, predictive element allowing the incremental integration of such continuations. The question we address here is whether the more parsimonious DS model, dispensing with an autonomous string-based syntax, can provide the required *predictivity* (for this psycholinguistic notion, see Sturt and Crocker (1996)); and indeed, besides its greater economy in representational levels, such a model seems better suited to capturing such phenomena since there are cases which show that such splits do NOT involve interlocutors intending to say the same string of words/sentence:

(4) with smoke coming from the kitchen:
A: Have you burnt the
B buns. Very thoroughly.
A: But did you
B: burn myself? No. Luckily.

The explanation for B's continuation in the fourth turn of (4) cannot be string-based as then *myself* would not be locally bound (its antecedent is *you*). Moreover, in LTAG, words are defined in terms of syntactic/semantic pairings, relative to a given head, with adjuncts as a means of splitting these. However, as (1)-(4) indicate, utterance take-over can take place at any point in a sequence of words with or without a head having occurred prior to the split. Many split utterances are not joint sentential constructions; and, they couldn't be because, as (2)-(4) show, even the function of the utterance can alter in the switch of roles, with fragments playing multiple roles at the same time (in (3): question/completion/acknowledgment/answer). If the grammar necessarily induces fine-grained speech act representations such multifunctionality cannot be captured except as a case of ambiguity or by positing hidden constituent reconstruction.

The setting for the P&R analysis is one in which participants are assigned a collaborative task with a specific joint goal, so joint intentionality is fixed in advance and hence anticipatory computation of interlocutor's intentions can be defined. However, (Mills and Gregoromichelaki, in prep) argue that, even in such task-specific situations, joint intentionality is not guaranteed but rather has to evolve as a result of routinisation. In accordance with this, as (1) shows, in ordinary conversation, there is no guarantee that there is a plan genuinely shared, or that the way the shared utterance evolves is what either party had in mind to say at the outset, indeed obviously not, as otherwise such exchanges would appear otiose. Instead utterances are shaped incrementally and "opportunistically" according to feedback by the interlocutor (Clark, 1996). And, as in (2), clarification can occur well before the completion of the utterance, which then absorbs both contributions. Grammatical integration of such joint contributions must therefore be flexible enough to allow such switches, with fragment resolutions occurring incrementally before computation of intentions at the pragmatic level is even possible.

The P&R account marks a significant advance in the analysis of such phenomena as it employs a dynamic view of the grammar in their analysis. But, as we saw above, the phenomenon is more general than just *completions/extensions*, the primary target of the P&R account. Nevertheless, given the observations above, dialogue exchanges involving incremental split utterances of any type are even harder to model adopting any other static grammatical framework. First of all, in such frameworks it is usually the sentence/proposition that is the unit of syntactic/semantic analysis, and, in the absence of an incremental/parsing perspective, elliptical phenomena/fragments are defined (following Dalrymple et al. (1991)) as associated with an abstraction operation over contextually provided propositional content to yield appropriate functors to apply to the fragment. But this problematically increases parsing uncertainty, since multiple options of appropriate "antecedents" for elliptical fragments become available (one for each available abstract). In consequence, to resolve such exploding ambiguities, the parsing mechanism has to appeal to general pragmatic mechanisms having to do with recognizing the speaker's intention in order to select a single appropriate interpretation. The conundrum that opens up is that intention recognition, on which all such successful contextual resolution will have to be based, is inapplicable in such sub-sentential split utterances, in all but the most task-specific domains. In principle, attribution to any party of recognition of the speaker's intention to convey some specific propositional content is unavailable until the appropriate propositional formula is established, so recognition of fully propositional intentions cannot be the basis on which incrementally established joint utterances are based. Moreover, from a generation point of view, relative to orthodox grammar-producer assumptions, the fact that speakers are interrupted, with (possibly unintended) continuations of their utterances being provided instead, means that the original speaker's plan to convey some full proposition will have to be abandoned mid-production, with some form of radical revision initiated in adopting the role of the parser. However, the seamlessness of such switches indicates no radical revision, and it is to be expected given the psycholinguistic evidence that speakers do not start articulating with fully formed propositional contents to convey already in mind (Levelt, 1989; Guhe, 2007).

Below we set out a model of parsing and production mechanisms that make it possible to show how, with speaker and hearer in principle using the same mechanisms for construal, equally incrementally applied, issues about interpretation choice and production decisions may be resolvable without reflections on the other party's mental state but solely on the basis of feedback. As we shall see, what connects our diverse examples, and indeed underpins the smooth shift in the joint endeavour of conversation, lies in incremental, context-dependent processing and tight coordination between parsing and generation, essential ingredients of the DS dialogue model (Cann et al., 2005). Instead of data such as (1)-(4) being problematic for such an account, in fact, their extensive use illustrates the advantages of a DS account in its provision of restricted contextually salient structural frames within which fragment construal/generation take place. This results in effective narrowing down of the threatening multiplicity of interpretations by incrementally weeding out possibilities en route to some commonly shared understanding. Features like incrementality, predictivity/goal-directedness and context-dependent processing are, that is, built into the grammar architecture itself: each successive processing step relies on a grammatical apparatus which integrates lexical input with essential reference to the context in order to proceed. Such a view notably does not invoke high-level decisions about speaker/hearer intentions as part of the mechanism itself. That this is the right view to take is enhanced by the fact that, as all of (1)-(4) show, neither party in such role-exchanges can definitively know in advance what will emerge as the eventual joint proposition.

An additional puzzle for any common-ground/intention-based views is that both speakers and hearers may elect not to make use of what is well established shared knowledge. On the one hand, in selecting an interpretation, a hearer may fail to check against consistency with what they believe the speaker could have intended (as in (5) where B construes the fragment in flagrant contradiction to what she knows A knows):

(5) A: Why don't you have bean chili?
B: Beef? You KNOW I'm a vegetarian
[natural data]

On the other hand, speaker's choice of anaphoric expression, supposedly restricted to well-established shared knowledge, is commonly made in apparent neglect of their hearer:

(6) A having read out newspaper headline about Brown and Obama, upon reading next headline provides as follow-on:
A: They've received 10,000 emails.
B: Brown and Obama?
A: No, the Camerons. [natural data]

Given this type of example, checking in parsing or producing utterances that information is jointly held by the dialogue participants - the perceived *common ground* - can't be a necessary condition on such activities. Hence it is not intrinsic to utterance interpretation in virtue of which conversational dialogue takes place. So we turn to Dynamic Syntax (DS) to explore possible forms of correlation between parsing and generation as they take place in dialogue without reliance on any such construct.

3 Incrementality in Dynamic Syntax

DS is a procedure-oriented framework, involving incremental processing, i.e. strictly sequential, word-by-word interpretation of linguistic strings. The notion of incrementality in DS is closely related to another of its features, the *goal-directedness* of BOTH parsing and generation. At each stage of processing, *structural predictions* are triggered that could fulfill the goals compatible with the input, in an underspecified manner. For example, when a proper name like *Bob* is encountered sentence-initially in English, a semantic predicate node is predicted to follow ($?Ty(e \rightarrow t)$), amongst other possibilities.

By way of rehearsing DS devices, let us look at some formal details with an example, *Bob saw Mary*. The 'complete' semantic representation tree resulting after full processing of this sentence is shown in Fig 1 below. A DS tree is binary and formally encoded with the tree logic *LOFT* (Blackburn and Meyer-Viol, 1994). It carries annotations at every node which represent semantic formulae with their type information (e.g. '$Ty(x)$') based on a combination of the epsilon and lambda calculi:

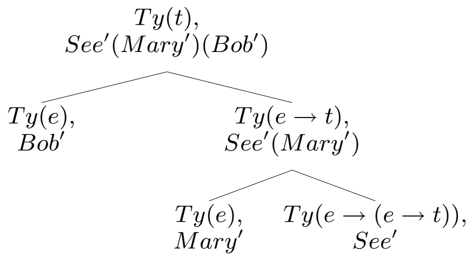


Figure 1: A DS complete tree

Such complete trees are constructed, starting from a radically underspecified goal, the *axiom*, the leftmost minimal tree in the illustration provided by Fig 2. Going through *monotonic updates* of partial or *structurally underspecified* trees, complete trees are eventually constructed. Crucial for expressing the goal-directedness are *requirements*, i.e. unrealized but expected node/tree specifications, indicated by '?' in front of annotations. The axiom says that a proposition (of type *t*, $Ty(t)$) is expected to be constructed. Furthermore, the *pointer* notated with '$\Diamond$' indicates the 'current' node in processing, namely the one to be processed next, and governs word order.

Updates are carried out by means of applying *actions* of two types. *Computational actions* govern general tree-constructional processes, such as moving the pointer, introducing and updating nodes, compiling interpretation for all non-terminal nodes. In Fig 2, the update of 1 to 2 is executed via computational actions expanding the axiom to the subject and predicate nodes, requiring the former to be processed next (given the position of the pointer). Construction of only weakly specified tree relations (*unfixed nodes*) can also be induced, characterized only as dominance by some current node, with subsequent update required. Individual lexical items also provide procedures for building structure in the form of *lexical actions*, inducing both nodes and annotations. In the update from 2 to 3, the set of lexical actions for the word *see* is applied, yielding the predicate subtree and its annotations. Unlike conventional bottom-up parsing, the DS model takes the parser/generator to entertain some predicted goal(s) (*requirements*) to be reached eventually at any stage of processing. Thus *partial trees*

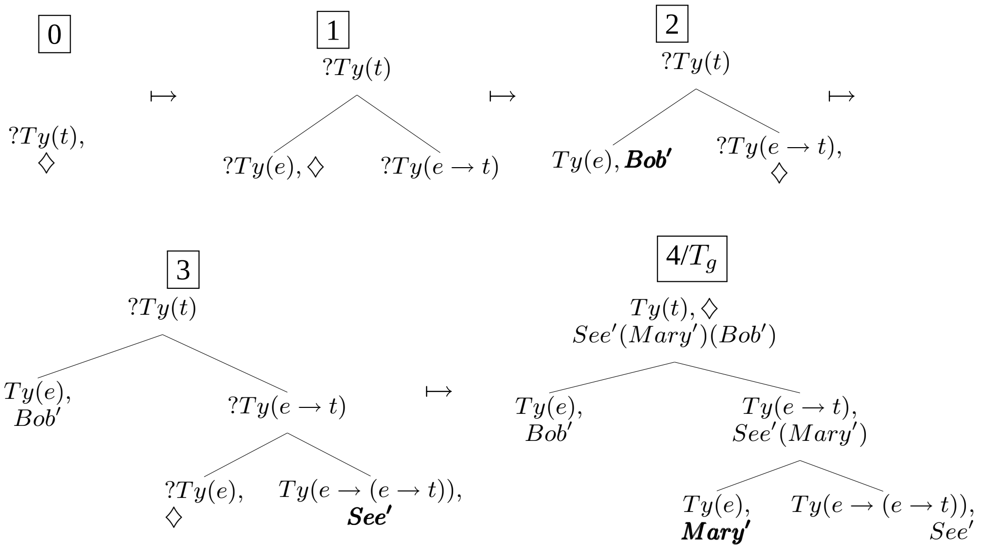


Figure 2: Monotonic tree growth in DS

grow incrementally, driven by procedures associated with particular words as they are encountered.

Individual DS trees consist of predicates and their arguments. Complex structures are obtained via a general tree-adjunction operation licensing the construction of LINK*ed* trees, pairs of trees where sharing of information occurs. The assumption in the construction of such LINKed structures is that at any arbitrary stage of development, some type-complete subtree may constitute the context for the subsequent parsing of the following string as an adjunct structure candidate for incorporation into the primary tree, hence the obligatory sharing of information in the resulting semantic representation.

Appositional structure, as in *A consultant, a friend of Jo's, left*, can then be established by defining a LINK transition as in Fig 3 from a node of type e in which a preliminary epsilon term[2] has been constructed (with all terminal nodes decorated but nonterminals not fully compiled) onto a LINKed tree introduced with a requirement to develop a term using that very same variable. A twinned evaluation rule then combines the restrictors of two such paired terms to yield a composite term (unlike the P&R account, this does not involve ambiguity of the head NP according to whether a second or subsequent NP follows). The fact that the first term has not been completed is no more than the term-analogue of the delaying tactic made available by expletive pronouns, extraposition etc, whereby a parse can proceed from some type specification of a node but without completing (*evaluating*) its formula. This strategy allows term modification when the pointer returns from its sister node immediately prior to compiling the decorations of its mother (as in *A man has won, someone you know*). Should this sequence of transitions be adopted by the hearer, in the absence of any such end-placed modification, it would constitute motivation for asking for clarification to enable a complete parse.

Such LINKed trees and their development set the scene for a general characterisation of context. *Context* in DS is defined as the storage of *parse states*, i.e., the storing of partial tree, word sequence parsed to date, plus the actions used in building up the partial tree. All fragments illustrated above are processed by means of either extending the current tree, or by constructing LINKed structures with transfer of information among them so that one tree provides the context for another. Such fragments are licensed as wellformed by the grammar only relative to such contexts (Gargett et al., 2008; Kempson et al., 2009).

4 Parsing/Generation Coordination

This architecture allows a dialogue model in which generation and parsing function in parallel, following exactly the same procedure in the same order. Fig 2 also displays the generation steps 0 to 4 of *Bob saw Mary*, for generation of this utterance follows precisely the same actions and trees from left to right as in parsing, although the complete tree is available as a *goal tree* from the start (hence the labelling of the complete tree as T_g):

[2] *Epsilon terms*, like $\epsilon, x, Consultant'(x)$, stand for witnesses of existentially quantified formulae in the epsilon calculus and represent the semantic content of indefinites. Defined relative to the equivalence $\psi(\epsilon, x, \psi(x)) = \exists x \psi(x)$, their defining property is their reflection of their containing environment, and accordingly they are particularly well-suited to expressing the growth of terms secured by such appositional devices.

Having parsed *a friend of Jo's* in *A consultant, a friend of Jo's, left*:

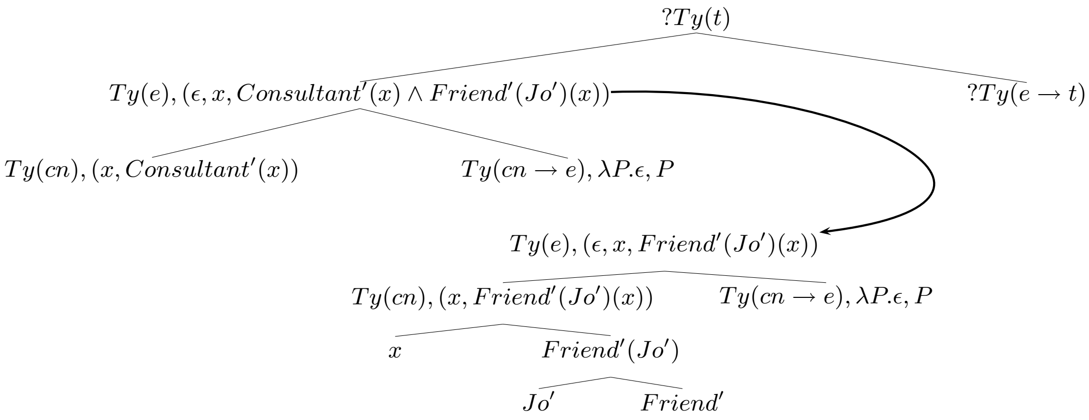

$?Ty(t)$
$Ty(e), (\epsilon, x, Consultant'(x) \wedge Friend'(Jo')(x))$
$?Ty(e \rightarrow t)$
$Ty(cn), (x, Consultant'(x))$
$Ty(cn \rightarrow e), \lambda P.\epsilon, P$
$Ty(e), (\epsilon, x, Friend'(Jo')(x))$
$Ty(cn), (x, Friend'(Jo')(x))$
$Ty(cn \rightarrow e), \lambda P.\epsilon, P$
x
$Friend'(Jo')$
Jo'
$Friend'$

Figure 3: Apposition in DS

in this case the eventual message is known by the speaker, though of course not by the hearer. What generation involves in addition to parse steps is reference to T_g to check whether each intended generation step (1, 2, 3, 4) is consistent with it. That is, a *subsumption* check is carried out as to whether the current parse tree is monotonically extendible to T_g. The trees 1-3 are licensed because each of these subsumes T_g in this sense. Each time then the generator applies a lexical action, it is licensed to produce the word that carries that action only under successful subsumption check: at step 3, for example, the generator processes the lexical action which results in the annotation 'See'', and upon success and subsumption of T_g license to generate the word *see* ensues.

For processing split utterances, two more consequences are pertinent. First, there is nothing to prevent speakers initially having only a partial structure to convey, i.e. T_g may be a *partial* tree: this is unproblematic, as all that is required by the formalism is monotonicity of tree growth, and the subsumption check is equally well defined over partial trees. Second, the goal tree T_g may change during generation of an utterance, as long as this change involves monotonic extension; and continuations/reformulations/extensions across speakers is straightforwardly modelled in DS by appending a LINKed structure annotated with added material to be conveyed (preserving monotonicity) as in single speaker utterances:

(7) A friend is arriving, with my brother, maybe with a new partner.

Such a model under which the speaker and hearer essentially follow the same sets of actions, each incrementally updating their semantic representations, allows the hearer to mirror the same series of partial trees as the producer, albeit not knowing in advance the content of the unspecified nodes. Furthermore, not only can the same sets of actions be used for both processes, but also a large part of the parsing and generation algorithms is shared. And both parties may engage with partial tree representations. Even the concept of *goal tree*, T_g, may be shared between speaker and hearer, in so far as the hearer may have richer expectations relative to which the speaker's input is processed, as in the processing of a clarification question. Conversely, the speaker may have only a partial tree as T_g, relative to which they are seeking clarification.

In general, as no intervening level of syntactic structure over the string is ever computed, the parsing/generation tasks are more economic in terms of representations. Additionally, the top-down architecture in combination with partiality allows the framework to be (strategically) more radically incremental in terms of interleaving planning and production than is possible within other frameworks. And there is evidence that such incrementality increases efficiency (Fernanda and Swets (2002):77).

4.1 Split utterances in Dynamic Syntax

Split utterances follow as an immediate consequence of these assumptions. For dialogues (1)-(4), A reaches a partial tree of what she has uttered through successive updates, while B as the hearer, follows the same updates to reach the same representation of what he has heard: they both apply the same tree-construction mechanism which is none other than their effectively shared grammar[3]. This provides B with the ability at any stage to become the speaker, interrupting to continue A's utterance, repair, ask for clarification, reformulate,

[3] A completely identical grammar is, of course, an idealisation but one that is harmless for current purposes.

or provide a correction, as and when necessary. According to our model of dialogue, repeating or extending a constituent of A's utterance by B is licensed only if B, the hearer now turned speaker, entertains a message to be conveyed (a new T_g) that matches or extends in a monotonic fashion the parse tree of what he has heard. This message (tree) may of course be partial, as in (2), where B is adding a clarificational LINKed structure to a still-partially parsed antecedent. Importantly, in DS, both A and B can now re-use the already constructed (partial) parse tree in their immediate context as a point from which to begin parsing and generation, rather than having to rebuild an entirely novel tree or subtree. In this sense, the most recent parse tree constitutes the most immediately available local "antecedent" for fragment resolution, for both speaker and hearer, hence no separate computation or definition of *salience* or speaker intention by the hearer is necessary for fragment construal.

As we saw, the hearer B may respond to what he built up in interpretation, anticipating the verbal completion as in (1)-(3). This is facilitated by the general predictivity/goal-directedness of the DS architecture since the parser is always predicting top-down goals (*requirements*) to be achieved in the next steps. Such goals are what drives the search of the lexicon (lexical access) in generation so a hearer who shifts to successful lexicon search before processing the anticipated lexical input provided by the speaker can become the generator and take over. In (3), B is, indeed, using such anticipation as, simultaneously, at least a completion of A's utterance, an acknowledgment of his understanding of the question and of his taking it up, and as a direct form of reply. Any framework that relies on complete determination of the speaker's intention in order to resolve such fragments does not allow for such multiple functionality. Instead, such fragments would have to be characterized as multiply ambiguous requiring the parser to select interpretations among a set of pre-defined options (but cf Ginzburg (forthcmg):Ch 3 for arguments in favour of this approach). Even if predetermination of such options were feasible, such a stance once more increases parsing uncertainty at the choice points so that inferential pragmatic mechanisms (appealing to deciphering speakers' intentions with reference to common ground) have to be invoked to select the appropriate update rules that should or should not apply at this juncture.

5 Summary Evaluation

With grammar mechanisms defined as inducing tree growth and used incrementally in both parsing and generation, the availability of these derivations from within the grammar shows how the core dialogue activities can take place without any other-party representation at all.[4] This then results in a view of communication that is not grounded in recognizing speaker's intentions, hence can be displayed by both young children and adults equally. The two crucial properties are the intrinsic predictivity/goal-directedness in the formulation of the DS, and the fact that both parsing and production can have arbitrary partial goals, so that, in effect, both interlocutors are able to be building structures in tandem. Because of the assumed partiality of goal trees, speakers do not have to be modelled as having fully formed messages to convey at the beginning of the generation task but can instead be viewed as relying on feedback to shape their utterance. As goal trees are expanded incrementally, completions by the other party can be monotonically accommodated even though they might not represent what the speaker would have uttered if not interrupted: as long as what emerges as the eventual joint content is some compatible extension of the original speaker's goal tree, it may be accepted as sufficient for the purposes to hand. Hence "repair" phenomena naturally emerge as "coordination devices" (Clark, 1996), devices exploiting mutually salient contexts for achieving coordination enhancement. And such jointly constructed content through cycles of "miscommunication" and "repair" is more securely coordinated (see e.g. Healey (2008)) and thus can form the basis of what each party considers shared cognitive context.

It might appear that the analysis faces the familiar exponential explosion of interpretations requiring the computation by the hearer of speaker intentions on the basis of common ground, albeit at a sub-propositional level. However, on the incremental processing view developed here, on the one hand, such speaker intentions are not available at the relevant juncture and, on the other hand, speaker intentions might not have even been formed given the partiality of the goal trees. But with feedback able to be provided/accommodated at any (sub-propositional) stage, the potential exponential explosion of interpretations can be kept firmly in check: structurally, such fragmental

[4] Note that we are not claiming that they necessarily do.

feedback can be integrated in the current partial tree representation directly (given the position of the pointer) so there is no structural ambiguity multiplication. What is notable is that for any one such intermediate check point, matching use of tree-construction processes by the parser and generator means that consistency checking can remain internal to each interlocutor's system. The fact of their mirroring each other results in their being at the same point of tree-growth and this provides a shared basis for understanding without explicit modelling of each other's information state. Even repairs may be processed relative to each interlocutor's own set of trees (background knowledge) and with no thought of what the other might have in mind. This is compatible with a mechanistic view of dialogue processing (Pickering and Garrod, 2004), though without invoking *priming*.

Of course, DS being a grammar formalism, an account of all facets of dialogue including its non-monotonic aspects is not within its remit. Nevertheless, the account provided does not preclude the representation of "intentions" as explicitly expressed and manipulated (in the form of adjoined LINKed structures), derived through the mechanisms mentioned in P&R or alternative routinisation accounts (Mills and Gregoromichelaki, in prep). Yet the dual applicability of the mechanisms, defined identically for both parsing and (tactical) generation, enables us to see how apparently shared contents can be incrementally and egocentrically derived, all context-based selections being based on the individual's own context as far as fragment resolution is concerned. Where uncertainty arises, the context-dependent repair mechanisms can take over. This, in its turn, makes possible an account of how hearers may construct interpretations that are transparently inconsistent with what both interlocutors know ((5)-(6)). Hence we suggest, contra (Tomasello, 2008), that we need to be exploring accounts of human communication as an activity involving emergent agent coordination without any required high-level mind-reading.

References

Nicholas Asher and Anthony Gillies. 2004. Common Ground, Corrections and Coordination. *Argumentation*, 17:481–512.

Patrick Blackburn and Wilfried Meyer-Viol. 1994. Linguistics, logic and finite trees. *Bulletin of the IGPL*, 2:3–31.

Richard Breheny. 2006. Communication and folk psychology. *Mind & Language*, 21(1):74–107.

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

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Mary Dalrymple, Stuart M. Shieber, and Fernando C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4):399–452.

Ferreira Fernanda and Benjamin Swets. 2002. How incremental is language production? evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, 46:5784.

Andrew Gargett, Eleni Gregoromichelaki, Chris Howes, and Yo Sato. 2008. Dialogue-grammar correspondence in dynamic syntax. In *Proceedings of the 12th SEMDIAL (LONDIAL)*.

Jonathan Ginzburg. 1995. Resolving questions, I. *Language and Philosophy*, 18(5):459–527.

Jonathan Ginzburg. forthcmg. *The Interactive Stance: Meaning for Conversation*. CSLI.

Markus Guhe. 2007. *Incremental Conceptualization for Language Production*. NJ: Lawrence Erlbaum Associates.

Patrick Healey. 2008. Interactive misalignment: The role of repair in the development of group sub-languages. In R. Cooper and R. Kempson, editors, *Language in Flux*. College Publications.

Istvan Kecskes and Jacob Mey, editors. 2008. *Intention, Common Ground and the Egocentric Speaker-Hearer*. Mouton de Gruyter.

Ruth Kempson, Eleni Gregoromichelaki, and Yo Sato. 2009. Incrementality, speaker/hearer switching and the disambiguation challenge. In *Proceedings of European Association of Computational Linguistics proceedings.*

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University. Also published as Gothenburg Monographs in Linguistics 21.

Willem JM Levelt. 1989. *Speaking*. MIT Press.

Greg Mills and Eleni Gregoromichelaki. in prep. Coordinating on joint projects. Based on talk given at the Coordination of Agents Workshop, Nov 2008, KCL.

Martin Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*.

Massimo Poesio and Hannes Rieser. to appear. Completions, coordination, and alignment in dialogue. Ms.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Blackwell.

Patrick Sturt and Matthew Crocker. 1996. Monotonic syntactic processing: a cross-linguistic study of attachment and reanalysis. *Language and Cognitive Processes*, 11:448–494.

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

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *8th International Conference on Intelligent Virtual Agents.*,

A monotonic model of denials in dialogue

Elena Karagjosova
Institut für Linguistik/Germanistik
Stuttgart University
elena.karagjosova@ling.uni-stuttgart.de

Abstract

The paper outlines a monotonic model of denial in dialogue that keeps a representation of the offensive material at the same time as it accounts for the tentative status of utterances with respect to the common ground (CG). It is cast in the Information state based approach to dialogue developed in the projects TRINDI and SIRIDUS (Cooper and Larsson, 1999; Larsson, 2002), and incorporates a notion of discourse commitments (DCs) that enables us to distinguish between information that is part of the CG and such that is merely proposed for consideration. The presented IS based model is meant as a first theoretical approximation towards an adequate DRT-based account of denial and correction.

1 Introduction

This paper deals with denials and their adequate modelling in terms of their effects in dialogue. It treats denial as a case of correction. Consider the dialogue in (1) where B denies the truth of the entire proposition expressed by A's utterance.

(1) A: The earth is flat.
B: No, it isn't.

The analysis suggested here should be also applicable to other kinds of corrections like (2), where the objection concerns only a portion of the utterance of the previous discourse participant.

(2) A: Anna ate spaghetti.
B: No, she ate salad.

Existing models of denial and correction are non-monotonic, i.e. they account of denial and correction as effectuating a removal of the corrected material from the context, see (Maier and van der Sandt, 2003) and (van Leusen, 2004). Both accounts employ a notion of context that corresponds to Stalnaker's *common ground* (CG), i.e. the commitments the discourse participants (DPs) have agreed upon (Stalnaker, 1979). A denial in these models has the effect that the CG is revised by removing the offensive material from the representation of the discourse.

However, the existing accounts of denial and correction in terms of CG-revision do not do justice to the nature of these phenomena. The core of the problem seems to be that the dialogue models employed are not fine-grained enough to treat denial properly. Intuitively, the content of utterances that are rejected does not become part of the CG in the first place. Therefore, it cannot be removed from it by means of denial. Each utterance made is only a proposal on how to update the CG. Only if it is accepted by the other DP, is it added to the CG. A sequence of an assertion and a denial represents a dialogue segment where the DPs explicitly negotiate on how the CG should be updated. In other words, existing models of denial and correction do not have a way to account for the preliminary status of utterances with respect to the CG and more specifically, of the explicit CG-negotiation that denials represent.

Another objection to the non-monotonic models of denial, and especially to the one in (Maier and van der Sandt, 2003) is that it is counterintuitive that the corrected material should completely disappear from the dialogue representation. For comparison, (van Leusen, 2004) deals with corrections in a more fine-grained model that distinguishes the discourse record (a record of all utterances contributed during the discourse, discourse history) from the discourse meaning. However, the same criticism holds for this model as well when it comes to accounting of corrections as context revisions.

The present paper outlines an alternative model of denial in dialogue. The model is monotonic since it keeps a representation of the offensive material at the same time as it accounts for the tentative status of utterances with respect to the CG. It is cast in the Information state based approach to dialogue developed in the projects TRINDI and SIRIDUS (Cooper and Larsson, 1999; Larsson, 2002), and incorporates a notion of discourse commitments (DCs) that enables us to distinguish between information that is part of the CG and such that is merely proposed for consideration. The IS based approach to dialogue provides a framework that is flexible enough to implement the more fine-grained dialogue model needed. The presented IS based model is meant as a first theoretical approximation towards an adequate DRT-based account of denial and correction. As the discussion of the existing DRT approaches will show, in order to model correction adequately in DRT, more fundamental, non-trivial modifications to the theory have to be made. This is a large scale enterprise that will be addressed in future work.

The paper is structured as follows. Section 2 provides an overview over the two existing elaborate models of denial and correction. In section 3, I briefly clarify my understanding of the relation between denials and the notion of context in terms of CG. I implement this relation within the framework of the Information state based approach to dialogue in section 4, and section 5 presents a refined account that also implements the notion of dialogue history. Finally, a summary and outlook are presented in section 6.

2 Non-monotonic models of denial

2.1 (Maier and van der Sandt, 2003)

(Maier and van der Sandt, 2003) account of denial in terms of its discourse effects, which are claimed to be “a non-monotonic correction operation on contextual information”. More closely, they argue that the primary function of denial is “to object to information which has been entered before and to remove it from the discourse record.” They model denials in an extension of DRT. In DRT, discourse is modelled in terms of abstract structures, DRSs, which represent the meaning of the discourse as it evolves. Each new sentence is interpreted on the background of the representation of the preceding discourse, the background-DRS. Thus in DRT, the notion of context is modelled by the DRS. In order to be able to model dialogue, (Maier and van der Sandt, 2003) propose an extension to standard DRT that allows keeping track of who said what and when in a dialogue.[1] In this extension, it is assumed that a DRS represents the CG of the DPs, i.e. the propositions that the DPs have agreed upon as being true. In this framework, it is not possible to model denial monotonically. As (Maier and van der Sandt, 2003, p.12) argue, “with respect to an incoming context that contains the offensive material the sentence cannot even be processed in view of the fact that this would result in a plain contradiction.” Therefore they model the effect of denial by means of the so-called “reversed anaphora”: the denial is not further processed but leaves a simple negated condition in the DRS. Then a preliminary sentence representation is constructed and merged with the background DRS. After that, a presupposition resolution mechanism with reversed anaphora collects the offensive material from the preceding utterance and moves it to the position of the negated condition. As a result of this process, the contribution of the corrected utterance is removed from the main DRS and the material it originally introduced ends up under the scope of the negation introduced by the denial. In other words, the offensive material is removed first from the dialogue representation (the CG), and then the content of the denial is added to it.

For instance, as a result of this process, the final representation of the dialogue in (3) is a DRS containing the representation of σ_2, which is the negated sentence σ_1, and the representation of σ_3.

(3) σ_1 A: The King of France walks in the park.
σ_2 B: No, he doesn’t,
σ_3 France doesn’t have a king.

However, the final representation of the dialogue in (3) does not seem satisfactory as a dialogue representation of an assertion-denial sequence since it only contains a representation of the negative assertion in σ_2 and contains no trace of the denied utterance σ_1. I.e. the representation of the dialogue in (3) hardly differs from the representation of the negative statement *The King of France does not walk in the park.* It is unsatisfactory to let the contribution of the offensive

[1] It is further refined to Layered DRT in order to be able to account for cases where only parts of the utterance are rejected while others are acknowledged.

material disappear from the dialogue representation. One reason for wanting to keep this material in a dialogue representation may be that a speaker may want to refer to it at some later stage of the dialogue. As already mentioned, in this approach non-monotonicity is necessary since the DRS represents the CG and must be kept consistent. On the other hand, the inadequacy of the proposed solution suggests that a more radical modification of DRT is needed to capture adequately the nature of dialogue and of phenomena like corrections.[2]

Another objection to the account presented in (Maier and van der Sandt, 2003) is that if a DRS reflects the CG, it is inadequate to treat the offensive material as being part of the CG, since the other DP hasn't acknowledged it yet. The same holds for the content of the denial itself, since the other DP may disagree and stick to his opinion. In the present model, the representation of the dialogue in (3) contains a representation of the denial σ_2, which means that the denial is part of the CG. In general, if a DRS reflects the CG, then it is impossible to add new sentence representations to it, since each utterance in a dialogue is only a proposal on how the CG should be updated. The CG changes only after the proposal is accepted, explicitly or implicitly. In other words, before each update of the CG, there is a grounding step (see (Traum, 1994) on grounding). In the case of denial and correction, we can speak of a dialogue segment that has the purpose of explicitly negotiating how the CG should be updated. Consequently, dialogue models should provide separate representations for the level at which the content of the CG is negotiated, and for the one that represents the result of this negotiation. As it stands, Maier and Sandt's DRT model reflects the former but is intended to represent the latter. When a speaker denies an utterance of the other DP, neither the content of the preceding utterance nor that of the denial are part of the CG. Consequently, denials cannot be identified with revisions in the CG. Denials are part of a negotiation phase in dialogue, and this is not reflected in the model.

2.2 (van Leusen, 2004)

The second nonmonotonic account of denial and correction in dialogue is proposed in (van Leusen, 2004). It is based on a more sophisticated dialogue model than the one in (Maier and van der Sandt, 2003), called Logical Description Grammar (van Leusen and Muskens, 2003), that distinguishes between the utterances made, or the dialogue history, and the discourse meaning. The former is modelled by means of "discourse descriptions", which describe the syntactic, semantic and pragmatic characteristics of the discourse and represent "the language user's knowledge of the discourse processed so far" (van Leusen, 2004). A second level constitutes the discourse meaning or context, which is a model that fits or verifies a discourse description. The incrementation of the discourse description is monotonic, including cases of corrections and denials. Corrections have an update effect only on discourse meaning.

In this model, discourse meanings and sentence meanings are DRSs. Correspondingly, contexts are DRSs, as well as the semantic contents of discourse contributions. The discourse meaning is built up in this model from the contents of the utterances of the DPs. Since DPs may disagree on certain points, it is argued, it is not necessary that all of the contextual information is believed or supported by each of the participants. Therefore it is assumed that the context at any point of the conversation represents what has been proposed for acceptance as CG by the most recent speaker.

This view is so far consistent with the position advocated in this paper. However, it is not entirely clear how the notion of context is defined in van Leusen's account. Thus, if the context only contains proposals on how to alter the CG, why does it have to be kept consistent? There may be contradicting proposals. Also, the effect of corrections on the context is modeled in terms of Gärdenfors' revision of epistemic states (Gärdenfors, 1988), which suggests that the notion of context employed in this model coincides with the notion of the CG, and the same criticism as for (Maier and van der Sandt, 2003) holds for this approach too.

[2] Another inadequacy of this model is that it does not account for the fact that it is not always the previous utterance that contains the offensive material, cf. corrections with German accented adverbial *doch*, as in (i), where the correction occurs several turns away from the turn that introduces the offensive material, see (Karagjosova, 2006) on corrections with *doch*.

(i) 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.')

3 Denials and the notion of context

As mentioned in section 1, the notion of context employed in the existing models of denial and correction is based on Stalnaker's notion of the CG. Stalnaker defines context as follows: "Think of a state of a context at any given moment as defined by the presuppositions of the participants." Presupposition is defined as "what is taken by the speaker to be CG of the participants", and the CG represents the propositions the dialogue participants have agreed upon, or mutually believe.

Based on this notion of context as CG, Stalnaker defines assertion in terms of its effects on the context, namely, the content of an assertion changes the context (i.e. the CG) by reducing the context set (i.e. all possible situations/worlds incompatible with what is said are eliminated). Stalnaker specifies further that this effect is only achieved provided there are no objections from the other DP. "This effect is avoided only if the assertion is rejected. " In a footnote (footnote 9 on p. 324), Stalnaker argues further that "to reject an assertion is not to assert or assent to the contradictory of the assertion, but only to refuse to accept the assertion. If an assertion is rejected, the context remains as it was." More exactly, rejection of an assertion blocks the effect assertions have on the context, namely adding its contents to the CG.

Thus we find support for our criticism of the existing models of denial in Stalnaker's definition of assertion and its effects on the CG. A denial is a rejection to add the contents of a previously made contribution to the CG. The corrected material, i.e. the previous commitment, is thus not yet part of the CG, since its content has not been agreed upon yet. The corrected material, as well as the correction itself, are just *proposals* on how to update CG.

In the next two sections I implement an alternative, monotonic account of denial in the framework of the Information State based approach to dialogue.

4 The Information State based approach to dialogue

In this section, I use the framework of the Information State based approach to dialogue to implement a model of denial that does not assume CG revision.

The information state (IS) is an abstract data structure that represents information available to the DPs at any given stage of the dialogue. It is based on Ginzburg's (Ginzburg, 1998) notion of the dialogue gameboard which in turn is a modification and elaboration on Stalnaker's "common ground" and Lewis' "conversational scoreboard" (Lewis, 1979).

Ginzburg's dialogue gameboard is a structure that includes propositions, questions and dialogue moves.[3] The IS is an adaption of Ginzburg's DGB. The IS is a flexible construct that allows adding more complexity depending on the requirements of the dialogue phenomena that are modelled. This is also the strategy that I will follow here. I will start with a basic IS model, namely the IS used in (Larsson, 2002) in an implementation of the dialogue system IBiS, and see how far this IS can get us. It will turn out that additional complexities have to be added.

The basic structure of the IS is represented in figure 1 on page 5. The dialogue information is divided into two basic records: private and shared information. The record of information private to the DPs contains an agenda of actions the DP (by default the system) needs to perform in the near future (/PRIVATE/AGENDA), a dialogue plan for more long-term actions (/PRIVATE/PLAN), and a set of beliefs (/PRIVATE/BEL). Another record represents the shared information, information that is public to the DPs (system and user), containing a set of mutually agreed-upon propositions (/SHARED/COM), a stack of questions under discussion (/SHARED/QUD) and information about the latest utterance (/SHARED/LU): the speaker and the speech act/move realised by the utterance (assuming for simplicity that each utterance realizes only one move). The propositions in /SHARED/COM need not be actually believed by the DPs but they have committed to them for the purpose of the conversation.

Let us examine how the dialogue information is recorded in the IS in the case of denial. Consider the exchange in (4).

(4) A: The earth is flat.
B: No, the earth is not flat.

After the first utterance, the IS looks like in figure 2. Here, the record /PRIVATE/BEL contains the information about the belief of the speaker that the earth is flat. In (Larsson, 2002), this slot is foreseen for utterances of the system, where for in-

[3] In this model, each DP has his own version of the DGB and there may be mismatches. This is also the model on which the IS in (Cooper and Larsson, 1999) is based.

$$\begin{bmatrix} \text{PRIVATE} & : & \begin{bmatrix} \text{PLAN} & : & \text{StackSet(Action)} \\ \text{AGENDA} & : & \text{Stack(Action)} \\ \text{BEL} & : & \text{Set(Prop)} \end{bmatrix} \\ \text{SHARED} & : & \begin{bmatrix} \text{COM} & : & \text{Set(Prop)} \\ \text{QUD} & : & \text{Stack(Quest)} \\ \text{LU} & : & \begin{bmatrix} \text{SPEAKER} & : & \text{Participant} \\ \text{MOVE} & : & \text{Move} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 1: IBiS information state type

$$\begin{bmatrix} \text{PR} & : & \begin{bmatrix} \text{AG} & : & \\ \text{BEL} & : & \{\text{flat(e)}\} \end{bmatrix} \\ \text{SH} & : & \begin{bmatrix} \text{COM} & : & \{\,\} \\ \text{QUD} & : & <?\text{flat(e)}> \\ \text{LU} & : & \begin{bmatrix} \text{SP} & : & \text{A} \\ \text{MV} & : & \text{assert(flat(e))} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 2: *A: The earth is flat*

$$\begin{bmatrix} \text{PR} & : & \begin{bmatrix} \text{AG} & : & \\ \text{BEL} & : & \{\neg\text{flat(e)}\} \end{bmatrix} \\ \text{SH} & : & \begin{bmatrix} \text{COM} & : & \{\} \\ \text{QUD} & : & <?\text{flat(e)}> \\ \text{LU} & : & \begin{bmatrix} \text{SP} & : & \text{B} \\ \text{MV} & : & \text{deny(flat(e))} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 3: *B: The earth is not flat*

stance after a database search the system writes the result of the search into this field in order to communicate this result in a next utterance. The field /SHARED/COM is assumed to be empty at the start of the dialogue. It will not change at this stage since the content of the utterance does not become CG before it has been accepted by the other DP. In general, the update of the IS is governed by update rules defined for handling various dialogue moves, as well as for handling plans and actions. Concerning the CG, different grounding strategies may be assumed, such as optimistic (content of utterance automatically added to CG), caucious (content added but can be retracted if DP objects to it) and pessimistic grounding (content of utterance added to CG only after positive evidence for grounding) (Larsson, 2002). I will assume a pessimistic grounding strategy because it seems to reflect more adequately the nature of corrections. Following this strategy, the CG will not get updated after this first utterance until positive feedback is received. The utterance of the assertion has the effect that the corresponding yes/no-question is pushed on top of the QUD-stack, where it stays until it is resolved, i.e. gets accepted.[4]

I leave the field /PRIVATE/AGENDA empty since at this point it is not relevant for the current investigation. I also ignore completely the field /PRIVATE/PLAN for the same reason.

After the second utterance, the IS gets updated again and looks like in figure 3. The field /PRIVATE/BEL contains now the belief of the current speaker B that he communicates via utterance (4-B). The CG is still empty. Ginzburg's utterance processing protocol foresees that when an assertion is rejected, its content is not added to the CG. The only effect it has is that the corresponding yes/no-question is pushed on /SHARED/QUD.

The record /SHARED/LU is updated with information about the current move. This record only keeps information about the latest move. The next move overrides it with its own information.

If A agrees with B, then the IS will look like the IS in figure 4:

$$\begin{bmatrix} \text{PR} & : & \begin{bmatrix} \text{AG} & : & \\ \text{BEL} & : & \{\neg\text{flat(e)}\} \end{bmatrix} \\ \text{SH} & : & \begin{bmatrix} \text{COM} & : & \{\,\neg\text{flat(e)}\,\} \\ \text{QUD} & : & <> \\ \text{LU} & : & \begin{bmatrix} \text{SP} & : & \text{A} \\ \text{MV} & : & \text{accept}(\neg\text{flat(e)}) \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 4: *A: You are right, the earth is not flat*

Here, the speaker holds the belief that the earth is not flat. The CG will be updated with this proposition, and the question will be removed from the QUD-stack since it is resolved.

Implementing denial in the basic IS under the pessimistic grounding strategy captures in a way the preliminary status of utterances with respect to the CG. However, it does not provide means for keeping track of the actual commitments of the DPs in the course of the dialogue, but only reflects those that the DPs have agreed upon. This and several other points require adjustments to the simple IS and the rules for its update in order to model denial more adequately in this framework. This issue will be the subject of the next section.

[4] Otherwise it remains unresolved.

5 A modified IS for dealing with denials

In order to be able to deal more adequately with denials and corrections in this framework, I suggest some adjustments to the structure of the IS and its contents, as well as to the rules of its update. The modified IS is presented in figure 5 on page 7.

First of all, we need to distinguish between different layers of information in dialogue, CG and discourse history. (Gunlogson, 2003) proposes a dialogue model that allows to keep track of the discourse commitments (DCs) of the DPs. DCs are beliefs publicly attributed to each participant in the conversation. I.e. if A says p then it becomes CG that A believes p. A public belief of a DP does not have to be mutual. I.e. if it is CG that A believes p, from this does not follow that it is CG that p. Thus in a way, DCs capture the notion of dialogue history. In this model, dialogue history is part of the CG: it is in the CG that A has committed himself to p. Gunlogson adopts Stalnaker's definition of the CG as the set of propositions representing what the participants in a conversation take to be mutually believed, or at least mutually assumed for the purpose of the discourse.

However, implementing directly the DCs as part of the CG will not lead us far in the case of denial and correction, since in the case the DP accepts the denial, the CG must be revised or else it will become inconsistent. However, this would mean that we remove a commitment made by a speaker from the dialogue record, which is not satisfactory: intuitively, even if the DP makes a contradictory commitment, the fact that he has made the earlier commitment remains. A more adequate solution will be therefore to separate the DC from the CG. I therefore implement Gunlogson's discourse commitments as a separate field of the IS. The DCs represent the propositions that the DPs have committed to in the course of the conversation. Each utterance (at least each assertion) leads to updating the DC (the field /SHARED/DC in figure 5 on page 7) with the information that the speaker believes the proposition expressed by the utterance. E.g., after A's utterance in (4), /SHARED/DC is updated with the information $B_A flat(earth)$. By means of implementing the DCs, we can keep track of the dialogue history, a record of all utterances contributed by the DPs in the course of the entire dialogue, independently of the CG status of their contents. In the field /SHARED/DC, mutually contradicting utterances of the DPs can coexist. The data type is assumed to be an ordered set (although it may be inconsistent) of beliefs.

Second, since we model interaction between human DPs, we need a way to keep track of both DP's beliefs and commitments. I.e. we need to be able to represent the beliefs of the DPs separately. This can be done by using a belief operator B indexed with the speaker of the utterance and holder of the respective belief. E.g., $B_A flat(earth)$. In other words, the field /PRIVATE/BEL is a set of beliefs. The update will not overwrite the information in this field, but augment it with the beliefs of the next DP.[5]

Thus the field /SHARED/DC will partly contain the same information as the field /PRIVATE/BEL. The difference will be that while we cannot retract commitments, we can revise belief states, i.e. delete certain beliefs from /PRIVATE/BEL. Thus the information in /PRIVATE/BEL is not redundant but can be used to model the dynamics of the belief states of the DPs during the exchange. As already said, DPs need not actually hold these beliefs, but it suffices that they act as if they were.

The CG is represented by a separate field. In order to avoid confusion with the Discourse Commitments, I call the field that records the CG /SHARED/BEL (instead of "shared commitments"), since it concerns the propositions that the DPs mutually believe.[6] The shared believes correspond to the notion of CG, i.e. commitments the DPs have agreed upon. The data type is a set of propositions.[7] Note that the CG does not include information that is merely public, or manifest, to the DPs, such as the information captured by the other subfields in the SHARED-field, but rather concerns only the content of the utterances.

By separating the DC from the CG we can capture the CG-negotiating effect of denial[8] and

[5] Another possibility is to have different copies of the IS for each DP, as in (Cooper and Larsson, 1999). However, this solution will unnecessarily complicate matters and will not be further pursued for the time being.

[6] Note that in (Cooper and Larsson, 1999) /SHARED/COM is /COMMON/BEL, which reflects more adequately the intended purpose of this field as a set of agreed upon, or commonly believed, propositions.

[7] Actually, the CG may contain not only propositions, but also beliefs attributed to other DPs, or introspective belief, which means that it is a set of propositions and beliefs. I will ignore this potential complication for the time being.

[8] There exist IS-based dialogue models that provide means for modelling the process of negotiating content, such as the PENDING field in (Ginzburg and Cooper, 2004) and the list of ungrounded discourse units in (Poesio and Traum, 1998)

$$\begin{bmatrix} \text{PRIVATE} & : & \begin{bmatrix} \text{BEL} & : & \text{Set(Bel)} \end{bmatrix} \\ \text{SHARED} & : & \begin{bmatrix} \text{DC} & : & \text{Set(Bel)} \\ \text{BEL} & : & \text{Set(Prop)} \\ \text{QUD} & : & \text{Stack(Quest)} \\ \text{MOVES} & : & \text{Set(Moves)} \end{bmatrix} \end{bmatrix}$$

Figure 5: IS modified

need not assume that denial updates nonmonotonically the CG.[9] A revision becomes only necessary within the private beliefs of the DPs (the field /PRIVATE/BEL) in case the corrected DP accepts the correction. Note that assuming that the private beliefs of the DP can be revised does not make our model of denial less monotonic. The monotonicity concerns only the DC field, which accumulates all utterances made during the conversation independently of whether their contents are accepted or rejected by the other DP.

Further, as already said, each assertive utterance results in raising the respective yes/no-question in QUD, i.e. pushing it on top of the QUD-stack.

Another change concerns the LU-record. In standard IS there is no relation between the utterance and its utterer after the IS gets updated - after the next utterance, the speaker is a different person, and we do not have a way to relate the contents of the utterance with its originator beyond the respective turn. LU only shows who the last speaker was. Having represented the DCs, we do not need information about the latest speaker. We keep however the information about the move realized by the utterance, where in order to keep track of who realized which move, we index the moves with the respective DP, e.g. $assert_A(flat(earth))$. Also, in order to have a more complete record of the course of the dialogue, we do not let the dialogue move-field be overwritten after each update, but make sure that it gets augmented with the next moves.[10] It would also be useful that the information in the move-field is ordered, i.e. we assume that it has the data type ordered set.[11]

I also ignore for the time being the fields /PRIVATE/AGENDA and /PRIVATE/PLAN, since they are not immediately relevant for my purpose.

All updates must be handled by respective update rules, whose definition however I have to ignore for the time being.

Let us go through an example to see how this model works. Figure 6 reflects the IS after the first utterance in (4). It contains the private belief of the speaker (under the assumption of cooperativity) that he communicates with his utterance. In the shared record, the CG is empty, for the sake of simplicity, i.e. this is the first utterance in a dialogue. The QUD is whether the earth is flat, and the communicated belief is recorded as a DC of A in /SHARED/DC.

$$\begin{bmatrix} \text{PRIV} & : & \begin{bmatrix} \text{BEL} & : & \{\text{B}_A\ \text{flat(earth)}\} \end{bmatrix} \\ \text{SH} & : & \begin{bmatrix} \text{DC} & : & \{\text{B}_A\ \text{flat(earth)}\} \\ \text{BEL} & : & \{\,\} \\ \text{QUD} & : & <?\text{flat(earth)}> \\ \text{MV} & : & \{\ \text{assert}_A\text{(flat(earth))}\ \} \end{bmatrix} \end{bmatrix}$$

Figure 6: *A: The earth is flat.*

The IS after the second utterance is represented in Figure 7. It reflects in addition the private belief of the speaker B, which is just opposite to what A asserts. In the shared record, the CG is still empty, since the DPs have not yet agreed on a proposition. The field /SHARED/DC is updated by the DC of B. Topmost on QUD is still the question whether the earth is flat.

$$\begin{bmatrix} \text{PRIV} & : & \begin{bmatrix} \text{BEL} & : & \{\text{B}_A\ \text{flat(earth)}, \\ & : & \text{B}_B\neg\text{flat(earth)}\ \} \end{bmatrix} \\ \text{SH} & : & \begin{bmatrix} \text{DC} & : & \{\text{B}_A\ \text{flat(earth)}, \\ & : & \text{B}_B\ \neg\text{flat(earth)}\} \\ \text{BEL} & : & \{\,\} \\ \text{QUD} & : & <?\text{flat(earth)}> \\ \text{MV} & : & \{\text{assert}_A\text{(flat(earth))}, \\ & : & \text{deny}_B\text{(flat(earth))}\ \} \end{bmatrix} \end{bmatrix}$$

Figure 7: *B: The earth is not flat.*

Suppose that after some convincing argumentation of B, A finally accepts B's counterproposal on how to update the CG, namely with the proposition that the earth is not flat. This situation is reflected in figure 8. Then, this proposition will

[9] Of course, there may be situations in a dialogue where the CG has to be revised, e.g. when both DPs adopt a belief that contradicts their earlier common knowledge.

[10] A similar strategy is also allplied in other dialogue models, see e.g. (Ginzburg and Fernandez, 2005).

[11] It may actually be more reasonable to assume a sequence or a stack, since there could be multiple entries. The same holds for the DC field.

be added to the CG, here the field /SHARED/BEL. The QUD will be empty - the question whether the earth is flat is resolved and can be popped out from the QUD-stack. The beliefs of A will be revised: the old abandoned belief of A that the earth is flat will be deleted. This can be reflected in the field /PRIVATE/BEL by either simply removing the respective belief from the set, or by marking it somehow as not held anymore (e.g. by crossing the respective belief out), if we want to be able to capture the dynamics of the DPs' beliefs. In the example, I choose the first option for simplicity.

$$\left[\begin{array}{lll}
\text{PRIV} & : & \left[\begin{array}{lll} \text{BEL} & : & \{ \mathrm{B}_B \neg \text{flat(earth)}, \\ & : & \mathrm{B}_A \neg\text{flat(earth)} \} \end{array}\right] \\
\text{SH} & : & \left[\begin{array}{lll}
\text{DC} & : & \{ \mathrm{B}_A \text{ flat(earth)}, \\
 & : & \mathrm{B}_B \neg\text{flat(earth)}, \\
 & : & \mathrm{B}_A \neg\text{flat(earth)} \} \\
\text{BEL} & : & \{ \neg\text{flat(earth)} \} \\
\text{QUD} & : & <> \\
\text{MV} & : & \{ \text{assert}_A\text{(flat(earth))}, \\
 & : & \text{deny}_B\text{(flat(earth))}, \\
 & : & \text{accept}_A(\neg\text{(flat(earth)))} \}
\end{array}\right]
\end{array}\right]$$

Figure 8: *A: You are right, the earth is not flat*

6 Summary and outlook

In this paper I present a model of denial in dialogue that assumes that denial does not revise the CG but represents a phase in a dialogue with the purpose to negotiate the contents of the CG. I implement this idea in the IS based approach to dialogue and argue that it is important to be able to keep track of the dialogue history in order to deal adequately with denials. An obvious drawback of the proposed implementation is that the IS and especially the fields private beliefs, DC, and moves can become extremely long for realistic applications. But since the purpose of this investigation is a theoretical one, this fact is irrelevant for the time being. The ultimate goal of the present investigation that will be pursued in future work, is the development of a DRT-based model of denial and correction in dialogue that distinguishes between the CG and the dialogue history, and takes into account the private beliefs of the DPs. A DRT-based model should also be able to provide a proper semantics for these notions, an issue that was neglected in the present paper.

Acknowledgement I would like to thank three anonymous reviewers for detailed comments and suggestions.

References

Robin Cooper and Staffan Larsson. 1999. Dialogue moves and information states. In *Proceedings of the Third IWCS*, Tilburg.

Peter Gärdenfors. 1988. *Knowledge in flux. Modelling the dynamics of epistemic states.* MIT Press, Cambridge, Massachusetts.

Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and philosophy*, (27):297–365.

Jonathan Ginzburg and Raquel Fernandez. 2005. Action at a distance: the difference between dialogue and multilogue. In *Proceedings of DiaLor.*

Jonathan Ginzburg. 1998. Clarifying utterances. In *Proceedings of TwentDial.*

Christine Gunlogson. 2003. *True to form. Rising and falling declaratives as questions in English.* Routledge, New York.

Elena Karagjosova. 2006. Correction and acceptance by contrastive focus. In *Proceedings of BranDial.*

Staffan Larsson. 2002. *Issue-based dialogue management.* Ph.D. thesis, Gteborg University.

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

Emar Maier and Rob van der Sandt. 2003. Denial and correction in layered DRT. In *Proceedings of DiaBruck.*

Massimo Poesio and David Traum. 1998. Towards an axiomatization of dialogue acts. In *Proceedings of TwenDial.*

Robert Stalnaker. 1979. Assertion. In P. Cole, editor, *Syntax and Semantics*, volume 9. Academic Press, London.

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

Noor van Leusen and Reinhard Muskens. 2003. Construction by description in discourse representation. In J. Peregrin, editor, *Meaning, the Dynamic turn.* Elsevier.

Noor van Leusen. 2004. Incompatibility in context. A diagnosis for correction. *Journal of Semantics*, 21(4).

Dialogue Management as Interactive Tree Building

Peter Ljunglöf
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
peb@ling.gu.se

Abstract

We introduce a new dialogue model and a formalism for limited-domain dialogue systems, which works by interactively building dialogue trees. The model borrows its fundamental ideas from type theoretical grammars and Dynamic Syntax. The resulting dialogue theory is a simple and light-weight formalism, which is still capable of advanced dialogue behaviour.

1 Background

1.1 Dialogue models beyond finite-state

A finite-state dialogue system employs dialogue states, connected by transitions, which represent where the dialogue participants are in the progress of the dialogue. This is a very low-level formalism, which only is feasible for very limited dialogue domains. The dialogues become system-driven – there is not much room for the user to take initiatives. A number of formalisms have been introduced to improve on this, that are based on richer, more powerful models of dialogue structure. Here are a few examples:

Form-based dialogue systems A form-based dialogue system divides different tasks into forms, similar to web forms, containing slots to be filled. VoiceXML (Oshry, 2007) is a W3C standard for writing form-based systems. This is a more powerful formalism than finite-state, but it too becomes difficult to manage when the complexity of the domain increases. One main reason for this is that form-based systems cannot handle underspecified or ambiguous information in a good way. Although the user is allowed to take some initiative within a form, it is the system that drives the dialogue on a higher level.

Dialogue grammars The idea of modelling dialogue in terms of a grammar is based on the idea of adjacency pairs, which describe facts such as that questions are generally followed by answers, proposals by acceptances, etc. Grammar-based dialogue systems were quite popular in the 1990's (Jönsson, 1997; Gustafson et al., 1998), but tend to be better at representing the surface linguistic expression involved in dialogue rather than the semantic content and its relation to context which is very often of central importance in determining the range of options available to a dialogue participant at a given point in a dialogue.

Plan- and logic-based approaches Plan-based dialogue systems construct or infer plans for fulfilling the goals of the dialogue participants. This is accomplished by using AI techniques such as planning and plan recognition. The related logic-based approach represents dialogue and dialogue context in some logical formalism. These systems tend to be computationally complex, since they perform general AI reasoning or theorem proving. For examples see, e.g., Allen et al. (2001), Sadek et al. (1997) and Smith et al. (1995).

The information state update approach To overcome the limitations of form-based systems, a theory of dialogue modelling was introduced, known as the information state update (ISU) approach (Larsson and Traum, 2000). It is based on a structured information state to keep track of dialogue context information. The information state is updated by update rules which are triggered by dialogue moves performed by the participants in the dialogue. The ISU approach enables a modular architecture which allows generic solutions for dialogue technology.

However, there are problems with ISU-based dialogue systems, such as the GoDiS dialogue manager (Larsson et al., 2000; Larsson, 2002). The update rules tend to get very complicated, making it difficult to foresee the side effects of changing a rule, or adding a new one.

Dynamic Syntax Dynamic Syntax is a combined syntactic and semantic grammatical theory (Kempson et al., 2001; Cann et al., 2005), which takes into account dialogue phenomena such as clarifications, reformulations, corrections, and acknowledgements.

The idea is that syntactic trees represent simple propositional sentences, and trees can be connected by links to form complex utterances. Dynamic Syntax can be seen as a kind of ISU formalism; the trees are built incrementally word-by-word, where an incomplete tree corresponds to an incomplete utterance. Words whose function is not determined yet (e.g., whether a noun in initial position should act as a subject or an object), are added as unfixed nodes below the current tree. Further on, when the interpreter has read some more words and its function has been determined, an unfixed node becomes a fixed part of the tree.

Since the minimal linguistic units in Dynamic Syntax are words, it is in practise only used for analysing single sentences or short dialogue exchanges. For full-size dialogues, the input resolution is too fine-grained.

Dialogue as proof editing Ranta and Cooper (2004) describe how a dialogue system can be implemented in a syntactical proof editor based on type theory, originally developed for editing mathematical proofs. Metavariables in the proof term represents questions that needs to be answered by the user so that the system can calculate a final answer. This is very similar to the Prolog-style proof-searching dialogue of Smith et al. (1995), but with a foundation in type-theory. However, Ranta and Cooper only support information-seeking dialogues, and the backbone is a fairly simple form-based dialogue system. Furthermore, there is no account for underspecified answers, anaphoric expressions, or ambiguous answers.

Our proposed dialogue model can be seen as a development of the approaches of Ranta and Cooper, and Smith et al., using ideas from Dynamic Syntax and ISU to make the system more flexible.

1.2 The Logic of Finite Trees

Dynamic Syntax is based on the underlying Logic of Finite Trees (Blackburn and Meyer-Viol, 1994), a logical theory which makes it possible to interactively build a tree in a logically well-founded manner. We will use two concepts from this logic; unfixed nodes and linked trees:

Unfixed nodes An unfixed node is a subtree which we know should be attached somewhere below a given node, but we do not yet know exactly where. Figure 1 contains three unfixed nodes: the A node dominates the B node, while the C node dominates both the D and E nodes. This means that in the final tree, C will contain both D and E as descendants. Note that this doesn't say anything about the order between D and E, it can even be the case that one of them will dominate the other in the final tree.

In Dynamic Syntax, unfixed nodes are used when the syntactic function of a phrase is unknown. E.g., a noun in initial position can function as a subject or an object, depending on the context. In the framework we introduce, unfixed nodes will be used for representing underspecified and/or ambiguous information.

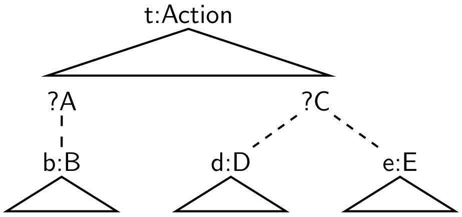


Figure 1: Unfixed nodes in a tree.

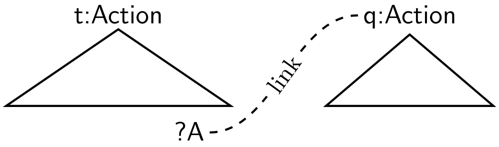


Figure 2: Two linked trees.

Linked trees Any node in a tree can have links to other trees. A link between two trees does not say that one of them dominates the other, it is merely a link between tree nodes. We assume that all links are labelled, as in figure 2.

In Dynamic Syntax, linked trees are used for, e.g., relative clauses, prepositional phrases, definites, anaphoric expressions and such things, whereas we will used them for question answering, sub-dialogues, and anaphoric references.

1.3 Type-theoretical grammar

Type theory is based on the Curry-Howard correspondence – "formulae-as-types" – where types correspond to propositions and terms correspond to proofs. To prove a proposition T we have to build a syntactic term $t : T$. An interactive proof editor builds a term interactively, where a metavariable $?T$ is used for an unknown subterm of type T. As Ranta and Cooper (2004) noted, $?T$ can be seen as a question posed by the system, "Which term of type T do you want to put here?".

Grammatical Framework (GF) is a grammar formalism based on type theory (Ranta, 2004). The main feature is the separation of abstract and concrete syntax, which makes it very suitable for writing multilingual grammars. The abstract part of a GF grammar defines a set of abstract syntactic structures, called abstract terms or trees; and the concrete part defines a relation between abstract structures and concrete structures. This separation of abstract and concrete syntax is crucial for our treatment of dialogue systems. A rich module system also facilitates grammar writing as an engineering task, by reusing common grammars.

Abstract syntax The abstract theory of GF is a version of Martin-Löf's (1984) type theory.

A grammar consists of declarations of categories and functions. In figure 3 is an example grammar, which we will use as our example domain. With the declaration route(?Dest,?Dept):Route, we mean that route(x,y):Route whenever x:Dest and y:Dept.[1] Furthermore, the grammar can contain function definitions, which we will use for calculating dialogue actions.

Concrete syntax GF has a *linearization* perspective to grammar writing, where the relation between abstract and concrete is viewed as a compositional mapping from abstract to concrete structures, called linearization terms. Linearizations are written as terms in a typed functional programming language, which is limited to ensure decidability in generation and in parsing.

It is possible to define several concrete syntaxes for one particular abstract syntax. Multilingual grammars can be used as a model for interlingua translation, but also to simplify localization of language technology applications such as dialogue systems.

Since this article is about the abstract dialogue model, and not about parsing and generation, we will not give any examples of linearization definitions. Examples of GF linearizations for dialogue systems can be found in Bringert et al. (2005) and Ljunglöf and Larsson (2008).

2 A tree-based ISU dialogue model

Our proposed dialogue model is an ISU model in that it operates on an information state which is modified by update rules. However, the information state is not a flat representation of plans and questions under discussion, as in, e.g., the GoDiS dialogue manager (Larsson, 2002). Instead the information state is represented by an incomplete tree in a similar way as is done in Dynamic Syntax, where incomplete nodes in the tree correspond to information that remains to be given.

In contrast with Dynamic Syntax, the minimal linguistic units are user and system utterances, and not single words. This makes it possible to model practical full-length dialogues, instead of being restricted to single sentences or short dialogue exchanges.

The goal of the dialogue is to build a tree, and when this tree is completed, it represents a task which the user wants the system to perform in some way. This is similar to a form in a form-based system, and a dialogue plan in an ISU system such as GoDiS, but it has a hierarchical, tree-based, structure instead of being flat. Using a tree-based information state means among other things that we can treat tasks, issues, plans and forms in exactly the same way as we treat the ontology of individuals, properties and predicates, thus simplifying the underlying logic. The use of trees, here, is related to the use of dialogue trees in, for example, work by Lemon et al. (2002), and are also found in dialogue grammar approaches. However, the kinds of trees we are using and the relationships we express between them are more complex. The main difference is that we used unfixed nodes and linked trees, which adds flexibility to the dialogue which has been a problem for grammar-based systems.

2.1 Specifying the dialogue domain

Simliar to our previous work (Bringert et al., 2005; Ljunglöf and Larsson, 2008), we use a type theoretical grammar to specify all aspects of the dialogue domain – tasks, issues, plans and forms, as well as individuals, properties and predicates. We can then make use of type checking for constraining the dialogue trees, and type checking can also be used when interpreting user utterances and when providing the user with suggestions of what to say next.

Another advantage with using a type theoretical grammar formalism, is that it is a multiple-level formalism, which can be used to specify the concrete user and system utterances which correspond to the tree structures that are used in the information state. Furthermore, Grammatical Framework is a multiple-language formalism, meaning that we can specify the dialogue domain as the language-independent part of the grammar, which is shared with all different language-dependent parts. Finally, type-checking is used to ensure that the different grammar instances are sound with respect to the dialogue domain.

To specify a dialogue domain, we have to declare all possible ways of forming trees. As already mentioned, an example travel agency domain is shown in figure 3, where with the declaration route(?Dest,?Dept):Route, we mean that route(x,y):Route whenever x:Dest and y:Dept. In this domain the user can book an event, ask for the price of an event, and ask when something happens. The possible events are oneway and round trips, hotel stays and conferences.

An example dialogue tree according to the specification is book(oneway(route(to(lon),from(gbg)), tomorrow)), which is also shown in figure 4. The concrete syntax defines translations between trees and utterances, and one possible translation of the example tree is "Book a oneway trip tomorrow from Gothenburg to London". We assume that the concrete syntax also defines translations of shorter phrases, such as "Book a oneway trip tomorrow", book(oneway(route(?,?),tomorrow)), and "A trip to London", route(to(lon),?).

[1] Note that we use a different GF grammar syntax than is common, to emphasise the similarities with tree-building and incomplete trees.

book(?Event), how-much(?Price), when(?Date) : Action
event-price(?Event) : Price
oneway(?Route,?Date), return(?Route,?Date,?Date),
hotel(?City,?Date), conf(?Conference) : Event
today, tomorrow, date(?Month,?Day),
conf-date(?Conference,?Year) : Date
route(?Dest,?Date) : Route
to(?City) : Dest
from(?City) : Dept
2008, 2009, . . . : Year
jan, feb, . . . : Month
1st, 2nd, . . . : Day
lon, gbg, . . . : City
acl, diaholmia, . . . : Conference
€450, €600, . . . : Price

Figure 3: Example domain

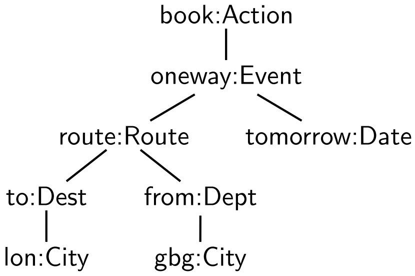


Figure 4: A completed dialogue tree

2.2 Dialogue as interactive tree building

The dialogue system tries to build a complete tree by successive refinement. In the middle of the dialogue, we represent the uninstantiated parts of the tree with metavariables. In this framework the metavariables are typed (which we write as $?T$) – when a new variable is created we can always infer its type from the types of the constants in the tree.

During the dialogue there can be several active dialogue trees, but there is always one current tree, and in that tree there is one single node which has focus. The focus node is highlighted like ⋆this⋆ in our example trees. The dialogue tree and its focus are operated with commands, such as changing focus to another node, inserting subtrees below the focus node, refining metavariables, etc.

The general idea is that the system moves the focus to a metavariable node, and asks the user to refine that node. User utterances are translated to incomplete subtrees, which the system tries to incorporate. If the user utterance is of the same type as the focused metavariable, the tree can be extended directly. Otherwise the system tries to add the utterance as an unfixed node below the focus, if possible, or tries to change focus to another metavariable which has the correct type.

2.3 System-driven dialogue

The dialogue starts with an incomplete tree, with only one metavariable stating the final type of the tree. In the example domain this final type is Action, so the initial tree is ?Action. The system then asks the question "What do you want to do?".

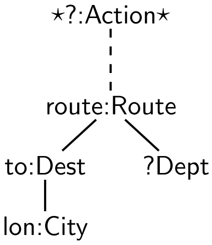


Figure 5: An incomplete dialogue tree.

Direct answer If the user gives a direct answer "Book a oneway trip tomorrow", book(oneway(route(?,?),tomorrow)), it is inserted at the focus node.

Being helpful If the user asks for help, or remains silent for a while, the system tries to refine the focus node itself. According to the specification, there are three possible actions, so the node is refined to the disjunction ?(book∨how-much∨when):Action. This is interpreted as an alternative question, "Do you want to book an event, ask for the price, or know a date?".

2.4 Handling underspecified information

The user is not required to always give direct answers to the system's questions; (s)he can, e.g., give underspecified answers. For incorporating underspecified information we use unfixed tree nodes, which is similar to how Dynamic Syntax does it: If the syntactic function of a word is unknown, its corresponding node in the tree becomes underspecified; e.g., a noun in initial position can be used as subject or object, and we cannot know which until more words are incorporated. This also corresponds to clarifications in GoDiS, within a single plan or between different plans.

If the user answers "A trip to London" (route(to(lon),?)), it is not a direct answer to the question ?Action. But since the answer type Route is dominated by Action, the system adds the answer tree as an unfixed node to the focus node. This is shown in figure 5.

Now, there are (at least) three different refinement strategies, depending on how the system searches for new metavariable nodes. We call these strategies top-down, bottom-up and "bottom-down".

Top-down refinement After this the system tries to refine the focus using the dominated tree as a constraint. Of the three possible Action refinements, only book and price can dominate a Route, so the focus node is refined to ?(book∨how-much):Action. This is shown in figure 6.

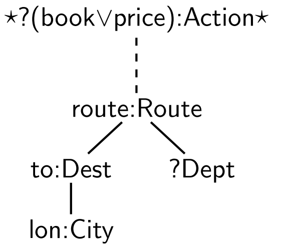


Figure 6: Top-down refinement of figure 5

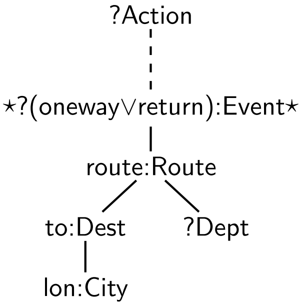


Figure 7: Bottom-up refinement of figure 5

The same thing will happen later in the dialogue, when the system wants to know which event to book (assuming that was what the user intended). When trying to refine the ?Event metavariable, only two of the four possible events can dominate a Route, so the node is refined to ?(oneway∨return):Event.

Bottom-up refinement Top-down refinement tries to connect a metavariable node with its unfixed tree by successively refining the dominating node. An alternative strategy is to instead connect the nodes by refining the dominated node. We call this strategy bottom-up refinement. The idea is that whenever the focus node has an unfixed child, the focus is moved to that child and refinement is done upwards. This means that when bottom-up refining the user answer "A trip to London", the system asks whether the user meant a oneway or a return trip, as shown in figure 7.

Furthermore, there are two different flavours of this dialogue strategy – non-aligned and aligned refinement. The most straight-forward variant of bottom-up refinement is to collect the possible immediate parents of the dominated node in the alternative question. Now, assume that the user only answered "London" to the initial ?Action question. There are three possible parents to a City – to:Dest, from:Dept and hotel:Event – which means that the system will have to ask the question ?(to:Dest∨from:Dept∨hotel:Event), which could be translated as "Do you mean to London, from London or a hotel in London?".

If it feels awkward to ask alternative questions about terms of different types, we can use *aligned* bottom-up refinement instead. In this variant, we collect the closest possible parents *all having the same type.* Since both Dest and Dept are dominated by Event, the question we get in our example is ?(oneway∨return∨hotel):Event.

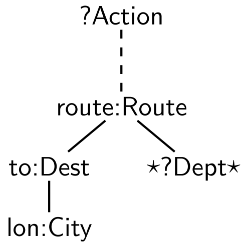


Figure 8: Bottom-down refinement of figure 5

"Bottom-down" refinement A third possible dialogue strategy, which we call "bottom-down" refinement, is to immediately dig into the tree that the user provided and try to complete that tree, before returning to the original top-level question. This means that after the system has attached the user answer "A trip to London" as an unfixed child of the ?Action node, focus is moved to the first metavariable in the given tree. The next question will therefore be "From where do you want to go?", as shown in figure 8. When the dominated tree is completed, the system can either proceed by top-down refining the dominating Action node, or by bottom-up refining the dominated Route node.

2.5 When the dialogue tree is complete

After hopefully a successful interaction, the dialogue tree is completed and represents an action that the user wants the system to execute. We model this with functional definitions, mapping the trees into action descriptions that the system can execute. In our example domain we can distinguish two kinds of actions:

Answering a question Some of the trees in the dialogue domain represent questions asked by the user. In our example both how-much(?Price) and when(?Date) represent user questions. To answer the question the system needs to consult a database, which can be encoded as function definitions in the domain:[2]

```
def conf-date(acl, 2009) = date(aug, 2)
def conf-date(...) = ...
def event-price(oneway(route(gbg,lon),
                        tomorrow) = €450
def event-price(...) = ...
```

After the dialogue tree is completed, the system evalutes it into an answer which then can be told to the user. The evaluated tree is added as a new

[2] Note that we are allowed to generate these function definitions automatically from the database in advance, or even on demand.

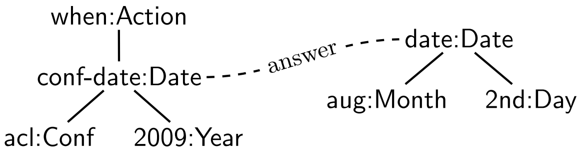


Figure 9: Answering a question

tree with a link to the original completed tree. The exact phrasing of the answer is specified in the concrete syntax.

For example, suppose the user asks "When is ACL?", which is recognised as when(conf-date(acl,?)). The system moves the focus to the metavariable, asking "Which year do you mean?", to which the user answers "2009". Now we get the final tree when(conf-date(acl,2009)), which is reduced by the function definitions to when(date(aug,2)). The system uses the concrete syntax to translate this into the answer "ACL starts 2nd August". The final dialogue state contains the question tree and the answer tree, connected by a link, as shown in figure 9.

Performing an action If the final tree is a booking of an event, the system needs a way of communicating with the outside world. Our simple solution is to attach side-effects to the type-theoretical function definitions. The problem with this is that the side-effects will reside outside the logical framework, which means that we cannot rely on type-checking or logical reasoning, for outside world interaction.[3] For example, the result of applying the term book(*event*) could be success(booking(*id*)) if the booking succeeds with booking number *id*, or failure(*reason*) otherwise, where *reason* is some explanation of why the booking failed.

2.6 Sub-dialogues

One important property of a flexible dialogue system is the possibility to engage in a sub-dialogue, and when that dialogue is finished, to return to the original dialogue. Here is an example from our example domain:

> **U:** "I'd like to book a trip from Gothenburg to Singapore"
>
> **S:** "When do you want to leave?"
>
> > **U:** "When is ACL?"
> >
> > **S:** "Which year do you mean?"
> >
> > **U:** "This year"
> >
> > **S:** "ACL starts 2nd August."
>
> **U:** "Okay, I'll leave the day before"

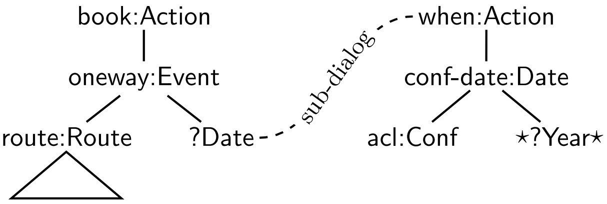


Figure 10: Engaging in a sub-dialogue

> **S:** "I have booked a trip to Singapore on 1st August."

We treat sub-dialogues in a similar way as Dynamic Syntax treats relative clauses: Since a relative clause is a statement of its own, it is interpreted as a complete tree, which is connected with a link to the referring node in the main tree. And in the same way a sub-dialogue is a tree of its own, with a link to the node where the sub-dialogue was introduced in the main dialogue tree. Figure 10 shows the dialogue state when the system asks the question "Which year do you mean?". After the user has answered this question, the system will create an answer tree just as in figure 9, after which focus is returned to the remaining ?Date metavariable in the main tree.

To be able to recognise the final user utterance correctly, the system needs to be able to handle anaphoric references ("the day before") by following sub-dialogue links.

2.7 Anaphoric expressions

In Dynamic Syntax, linked trees are also used for anaphoric references. A pronoun, or a definite noun phrase, suggests that there is a matching reference somewhere in the context. We treat anaphora in a similar way, by linking the anaphoric node to a previous dialogue tree.

> **U:** "How much is a flight from Gothenburg to London tomorrow?"
>
> **S:** "It costs €450."
>
> **U:** "Okay, book it"
>
> **S:** "I have booked a flight to London tomorrow."

After the first two utterances we have two dialogue trees – one representing the user question which is linked to the answer tree. Since these trees are completed, the next user utterance creates a new dialogue tree. The pronoun "it" is translated to a special constant it:Event, which triggers a lookup in the dialogue context for a matching subtree of the same type. The system finds a matching tree and creates an anaphoric link, as is shown in figure 11. When executing the booking, the system can use the event referred to by the link.

[3] Note that this is a problem that is shared with other existing dialogue frameworks. A more logically sound solution would be to, e.g., use a *state monad* for modeling the outside world (Wadler, 1997).

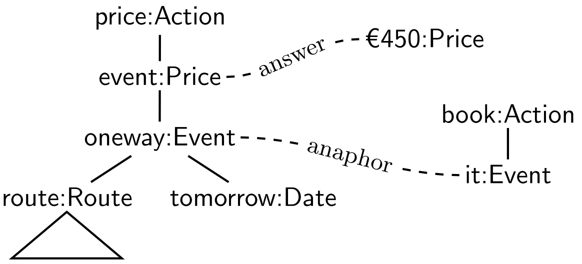


Figure 11: Handling anaphoric expressions

3 Discussion

We have introduced a dialogue model which works by interactively building dialogue trees. The model is a development of the "dialogue as proof editing" idea by Ranta and Cooper (2004), enhanced with a treatment of underspecification and references inspired from Dynamic Syntax.

Specifying user and system utterances By using a type-theoretical grammar formalism such as Grammatical Framework, we can specify all user and system utterances together with the abstract specification. The type checker can be used for catching errors in the specification, and the modular features of GF can be used for reusing grammar resources.

Questions under discussion The metavariables in the active dialogue tree correspond to the QUD (Questions Under Discussion), introduced by Ginzburg (1996). The QUD is a partial ordered set, and its topmost element corresponds to the focus node in our framework. The partial ordering of the QUD is implicit in our model, in the domain specification together with the order in which the algorithm searches the tree for metavariables.

Unified treatment of plans and items If we look at the domain specification in figure 3, we see that there is no conceptual difference between the plans (e.g., asking for the price or specifying an event) and the individual entities (e.g., the cities, dates or conferences). In fact, a declaration of a function such as route(?Dest,?Dept), both defines a dialogue plan (asking for a destination and a departure city) and the resulting individual (a specific route between two cities). This is in contrast with traditional form-based systems such as VoiceXML, and ISU systems such as GoDiS, where plans and individuals are separate concepts.

Unfixed tree nodes We use unfixed tree nodes for representing underspecified information, much in the same way as Dynamic Syntax does. The underlying Logic of Finite Trees automatically uses these nodes as constraints on the dominating nodes. We have described several different strategies for handling underspecified information (top-down, bottom-up and "bottom-down"), which then correspond to different strategies for accommodation in existing ISU dialogue models.

Links between trees Similar to Dynamic Syntax, we use links between trees for question answering, sub-dialogues and anaphoric expressions. The GoDiS dialogue manager handles sub-dialogues by having a stack of active plans, but it has no treatment of anaphoric references.

Function definitions for system replies Type-theoretical function definitions represent system replies to user questions and requests. This corresponds to external database calls in other formalisms, the difference being that we are using a well-founded logical theory, hopefully making it easier to reason logically about the properties of the system.

3.1 Future work

There are some issues that we have not addressed in this article, which are necessary for a working dialogue system.

Feedback We have not described how feedback should be treated. The reason is that since feedback cannot be defined in terms of the dialogue tree, its treatment is an orthogonal matter. Our aim is to incorporate the Interactive Communications Management (ICM) of Larsson (2002) into the system. This means that we need to add feedback information to the dialogue state, in parallel with the linked dialogue trees.

Corrections We have not described how the user can correct erroneous information in the dialogue tree. To be able to do this we need commands for deleting and changing tree nodes, as well as a functioning feedback system for clarifying the corrections.

Implementation We have rudimentary implementations in the programming languages Haskell and Python, but they need much more work to be useable as dialogue systems.

Acknowledgments

I am grateful to Staffan Larsson and four anonymous referees for constructive comments and suggestions.

References

James Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI Magazine*, 22(4):27–37.

Patrick Blackburn and Wilfried Meyer-Viol. 1994. Linguistics, logic, and finite trees. CWI Report CS-R9412, Centrum voor Wiskunde en Informatica, Amsterdam, Netherlands.

Björn Bringert, Robin Cooper, Peter Ljunglöf, and Aarne Ranta. 2005. Multimodal dialogue system grammars. In *Proc. Dialor'05, 9th Workshop on the Semantics and Pragmatics of Dialogue*, Nancy, France.

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

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. In J. Seligman, editor, *Language, Logic and Computation, volume 1*, CSLI Lecture Notes. CSLI Publications.

Joakim Gustafson, Patrik Elmberg, Rolf Carlson, and Arne Jönsson. 1998. An educational dialogue system with a user controllable dialogue manager. In *Proc. ICSLP'98, 5th International Conference on Spoken Language Processing*, Sydney, Australia.

Arne Jönsson. 1997. A model for habitable and efficient dialogue management for natural language interaction. *Natural Language Engineering*, 3(2-3):103–122.

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

Staffan Larsson and David Traum. 2000. Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, 6(3–4):323–340.

Staffan Larsson, Peter Ljunglöf, Robin Cooper, Elisabet Engdahl, and Stina Ericsson. 2000. GoDiS – an accommodating dialogue system. In *Proc. ANLP–NAACL'00 Workshop on Conversational Systems*, Seattle, Washington.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Department of Linguistics, University of Gothenburg.

Oliver Lemon, Alexander Gruenstein, and Stanley Peters. 2002. Collaborative activities and multi-tasking in dialogue systems. *TAL: Traitment Automatique des Langues*, 43(2):131–154.

Peter Ljunglöf and Staffan Larsson. 2008. A grammar formalism for specifying ISU-based dialogue systems. In *Proc. GoTAL'08, 6th International Conference on Natural Language Processing*, number 5221 in Springer-Verlag LNCS/LNAI, Gothenburg, Sweden.

Per Martin-Löf. 1984. *Intuitionistic Type Theory*. Bibliopolis, Napoli.

Matt Oshry, editor. 2007. *Voice Extensible Markup Language (VoiceXML) 2.1*. W3C Recommendation, `http://www.w3.org/TR/voicexml21/`.

Aarne Ranta and Robin Cooper. 2004. Dialogue systems as proof editors. *Journal of Logic, Language and Information*, 13(2):225–240.

Aarne Ranta. 2004. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189.

M. David Sadek, Philippe Bretier, and Franck Panaget. 1997. Artimis: Natural dialogue meets rational agency. In *Proc. IJCAI'97, 15th International Joint Conference on Artificial Intelligence*, Nagoya, Japan.

Ronnie W. Smith, Alan W. Biermann, and D. Richard Hipp. 1995. An architecture for voice dialog systems based on prolog-style theorem proving. *Computational Linguistics*, 21(3):281–320.

Philip Wadler. 1997. How to declare an imperative. *ACM Computing Surveys*, 29(3):240–263.

The Non-Individuation Constraint Revisited: When to Produce Free Choice Items in Multi-Party Dialogue*

Vladimir Popescu[1], **Jean Caelen**[2]
[1]Laboratoire Informatique d'Avignon,
University of Avignon, France
vladimir.popescu@univ-avignon.fr
[2]Laboratoire d'Informatique de Grenoble,
CNRS, France
jean.caelen@imag.fr

Abstract

In this paper we establish a set of conditions on the production of free choice items (FCI) in multi-party dialogue. Thus, we first observe that indefinite constructions are produced when speakers try to lead their addressees to access general, scalar rules, called *topoï*. These rules are used in reaching certain conclusions. However, the hearers need to be lead to access topoï when they do not manage to do this directly from definite sentences. The ability of the hearers to access topoï from definite sentences is assessed by inspecting the history of their *public commitments* in dialogue: if certain commitments are made, then it is *abductively* inferred that a certain topos was used; if so, then the hearers do not need to be "exposed" to utterances containing indefinite constructs. Secondly, an indefinite construction can be linguistically materialized as a FCI when it is not reducible to a referential situation (the *non-individuation constraint*). We thus propose a way of formalizing the non-individuation constraint in a multi-party dialogue setting, using public commitments as actual worlds, and a λ calculus-based formalism for matching the production of indefinite constructs to the accesses to topoï.

1 Introduction

Usually, FCIs (i.e., indefinite words such as 'any' and sometimes 'every' in English, or 'n'importe quel' and 'tout' in French) are studied in an *interpretation* context, i.e., for deciding when and why an utterance containing a FCI is felicitous, and another one is not (Giannakidou, 2001), (Jayez and Tovena, 2004). In this paper, generation aspects are studied, i.e., when it is appropriate to produce utterances containing FCIs (e.g., '*Every* student knows that' in English, or '*N'importe quel* étudiant sait ça' in French), and this, in a multi-party dialogue context.

For this, we link the notion of FCIs to that of argumentative *topoï*, i.e., general, scalar rules, of the form 'The more / the less P, the more / the less Q', to be read as 'if P (or $\neg P$) to a certain extent, then Q (or $\neg Q$) to a certain extent' (Anscombre, 1995). More precisely, we assume that, for generality, topoï are stored as general rules, λ-abstracted over the particular *types* (viz. human, student, book, hammer, ...) or *features* (viz. size, quantity, identity) of the entities involved in the rules (Popescu and Caelen, 2008).

Thus, assuming that indefinite constructions signal abstractions over the particular features of the entities, it results that utterances containing indefinite determiners (e.g., 'some books') can constitute (or readily imply, in a logical sense) the left side of a topos. Moreover, knowing that FCIs are a particular form of indefinite constructions, we can conclude that a FCI facilitates the access to topoï, from the perspective of the addressee of the utterance that contains it.

Thus, in a dialogue, whenever a speaker wants a hearer to access a certain topos for reaching a certain conclusion, she produces an utterance containing an indefinite construction. And, if this indefinite construction is not reducible to a referential situation (Jayez's non-individuation constraint – NIC (Jayez and Tovena, 2004)), then it is realized, for example, as 'any' in English, or as 'n'importe quel' or 'tout' in French. In order to give a precise formalization of this process, we need to tackle two issues:

*This research has been partly funded by the National Research Authority of the French Government (ANR), under the grant n^o ANR-007-008 (AVISON).

1. deciding when it is necessary to explicitly facilitate the access to a topos (i.e., when the addressee of an utterance is, a priori, not able to access the topos directly from the definite utterance), by using an indefinite construction;

2. deciding when it is possible to realize the indefinite construction as a FCI (i.e., when the NIC is met).

For the first issue, we rely on the public commitments (Kibble, 2006) of the interlocutors: if an interlocutor already committed, in the same dialogue, to a conclusion that *would* have been derived by using a topos, then one infers that this interlocutor has already had a *recent* access to the topos, hence it is very likely that she or he might access it again if necessary. Otherwise, one infers that the access to the topos has to be facilitated by λ-abstracting over certain entities in the utterances. The commitments are derived from the (Segmented) Discourse Representation Structure (SDRS) that each dialogue participant builds, as her / his view on the dialogue (Lascarides and Asher, 2009). The SDRSs for the speakers are determined in the framework of Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003).

The second issue is tackled by adapting Jayez's formalization of NIC (Jayez and Tovena, 2004) to generation, and extending it to a multi-party dialogue context. Thus, the "worlds" are the speakers' public commitments; the hybrid semantics notion of a clause being true *at* a certain world (Blackburn, 2000) is replaced with the notion of a clause being *entailed* from a public commitment (Lascarides and Asher, 2009), and the multi-party interactional context is accounted for by explicitly individualizing the commitments of each dialogue participant, and by studying the (set-theoretic) relations between these commitments.

Both these issues are given a unified formalization by using a non-typed λ calculus for representing the "indefiniteness". However, the entities on which these λ-abstractions apply are semantically typed (viz. agent, object, patient, and modifier[1]).

In this paper, after first presenting the unified λ calculus-based formalism used throughout the paper, we discuss aspects related to generating indefinite constructs in dialogue, namely the issue of accessing argumentative topoï. Then, we show how public commitments can be used as an *abductive* "hint" for deciding whether an interlocutor has already had access to a topos in the current dialogue. We also provide an extension of Jayez's NIC (Jayez and Tovena, 2004) to multi-party dialogue contexts. Finally, an extended example of a multi-party dialogue is presented for demonstrating the adequacy of the proposed framework.

2 Generating Free Choice Items in Multi-Party Dialogue

2.1 Theoretical Issues

We start from (Jayez and Tovena, 2004)'s study, that we extrapolate to multi-party dialogue utterance production. Thus, according to (Jayez and Tovena, 2004), FCIs satisfy three criteria: (i) they are not natural in affirmative episodic utterances; (ii) they are possible in generic and/or imperative and/or conditional utterances; (iii) FCIs implicate that the entities they are applied on in utterances can be freely chosen between the members of a set of entities.

For utterance production, Jayez's NIC is equivalent to the situation of producing a λ-abstracted utterance, where the β-reduction process is blocked (i.e., $\lambda p.Q(p)@\pi$ is impossible); this is equivalent to saying that a FCI is not reducible to a referential situation (Jayez and Tovena, 2004).

The NIC should be verified when an utterance ought to contain an indefinite construction (signaled, at a semantic level, by a λ-abstraction over an entity in the utterance). This indefinite construction could be specified at a semantic level in order to facilitate the access to certain topoï (Anscombre, 1995), (Popescu and Caelen, 2008). This is, in turn, necessary for the addressee of an utterance to reach certain conclusions (hinted at by the speaker), by way of these topoï. The speaker thus increases the argumentative *strength* (Popescu and Caelen, 2008) of its utterances.

Consider, for instance: '*Any* house would be OK for me!'; a part of its semantic form (that emphasizes the logical object of the utterance) is:

$\lambda X.(\textsf{[object]}(X) \wedge \textsf{equals}(X, \text{'house'}) \wedge ...)$.

Via such an expression, its addressee can reach a topos of the form: 'The more one has a house, the happier one is', i.e., in logical form:

$(\lambda X \lambda Z.(\textsf{[object]}(X) \wedge \textsf{equals}(X, \text{'house'})) \wedge$

[1] A semantic type of predicate is also needed for specifying the logical form of an utterance, but in this study abstractions (whence indefinite constructions) over predicates are not considered.

$\mathsf{[agent]}(Z) \wedge \mathsf{have}(Z,X))_{+}, (\lambda Y.(\mathsf{[agent]}(Y) \wedge \mathsf{happy}(Y)))_{+} \wedge [Z \equiv Y]$.

The predicates $\mathsf{[object]}$ and $\mathsf{[agent]}$ designate the semantic roles of the object of the action reported in an utterance, and the agent performing this action, respectively; $\mathsf{equals}/2$ is true if and only if its two arguments are bound to the same value; the last conjunct is a procedure that states the identity of the variables Z and Y; the lower index $+$ of a logical expression stands for a positive scalar value (i.e., 'the more') applied to the expression.

The usage of abstractions for facilitating the access to topoï is needed because, unlike the "ideal" situation assumed in (Popescu and Caelen, 2008), where addressees automatically perform the required λ-abstractions for accessing appropriate topoï, real dialogue agents (e.g., humans) have only partial reasoning capabilities (i.e., either they just do not perform the required λ-abstractions, or they do not perform them in due time – they perform them too late, i.e., not before the interlocutor's *subsequent* speech turn). In multi-party dialogue the situation is even thornier, because certain participants might be able to perform λ abstractions, certain might not. The use of indefinites is thus a means to tune this ability for *certain* addressees, which might yield a behavior of *selective cooperativity* in dialogue.

We will illustrate the formalization proposed for representing FCIs by considering an example: 'Any book is a waste of time', or, in logical form:

$\lambda X.(\mathsf{[object]}(X) \wedge \mathsf{equals}(X, \text{'book'}))$,

with: $\nexists \xi | \lambda X.(\mathsf{[object]}(X) \wedge \mathsf{equals}(X, \text{'book'}))@\xi$ (i.e., the β-reduction on X is blocked). This will be, by convention, written in a condensed form as:

$\lambda X.(\mathsf{[object]}(X) \wedge \mathsf{equals}(X, \text{'book'}))\neg@$.

When several variables are involved, those where λ abstractions are possible are marked by the β-reduction operator, preceded by the modal possibility operator ($\Diamond$). Thus, for '*Any* book makes us waste *some* time (reading it).', we have:

$\lambda X \lambda Y.(\mathsf{[object]}(X) \wedge \mathsf{equals}(X, \text{'book'}) \wedge \mathsf{[mod]}(Y) \wedge \mathsf{equals}(Y, \text{'time'}) \wedge \mathsf{waist}(.., X, Y))\Diamond@\neg@$.

Thus, here the β-reduction on Y can be performed.

The multi-party dialogue context imposes constraints concerning the selectivity of the speakers, according to their *dynamic profile*, i.e., their demonstrated ability to perform λ-abstractions for accessing topoï. The dynamic profiles of the speakers are *dialogue-wise*, in the sense that they are not persistent from one conversation session to another. These profiles are captured via the *public commitments* of the speakers: if a speaker commits herself to a fact, then she *must have* performed the required reasoning for this, e.g., access some topoï for deriving certain conclusions (associated – i.e., resulting from, or leading to – that fact). The reliance on public commitments in this way for determining the speakers' ability of accessing topoï is a form of abductive reasoning (i.e., $(P \Rightarrow Q) \wedge Q / > P$, where ">" means "normally", deafeasibly (Asher and Lascarides, 2003)). The commitments are expressed as user-specific SDRSs (cf. (Lascarides and Asher, 2009)).

A thorny issue concerning the abductive reasoning discussed above concerns the uniqueness of the premise (Hobbs et al., 1993): how do we know that a hearer committed to a fact by accessing a certain topos, and not in another way (e.g., by trusting the speaker, by following her order, or by modus ponens-like reasoning on facts in her/his own knowledge base)? An answer is that, in our case, we assume no a priori concerning trust (i.e., interlocutors do not a priori trust each other), social hierarchies are not assumed between dialogue partners (i.e., there are no orders simply followed) and, moreover, that abductive reasoning is not fragile, i.e., when a speaker *might* have gotten committed to a fact via a topos, we assume that this was, indeed the case. However, we should relax this constraint and provide a more fine-grained distinction between the situation where a topos is more likely to have been used, or static knowledge might have been used.

A general procedure for producing FCIs goes as follows:

1. for an utterance to generate (labeled by π, with $K(\pi)$ its logical form), check if it has the potential of facilitating the addressee to reach a certain conclusion (or, in another parlance, to commit him/herself to a certain fact), via a topos, τ; if so, then go to step 2; otherwise, feed the utterance into a surface realizer and stop;

2. check whether the addressee has the ability to access this topos τ directly from the non-indefinite form of the utterance (i.e., check whether that topos might have been already

used for reaching some facts in the current commitment store of the addressee); if so, then feed the utterance into a surface realizer and stop; otherwise, go to step 3;

3. perform a λ-abstraction over some relevant entities or the determinants of these entities in $K(\pi)$, so that the abstracted logical form, denoted by $\overline{K(\pi)}$ can constitute a premise for τ (i.e., $\tau = (\{|\neg\}\overline{K(\pi)}, \{|\neg\ K(\pi'))$, where $K(\pi')$ is the conclusion to be reached);

4. if β-reduction is possible by relying on the current contents of the commitment stores of the addressees of utterance π, then generate the λ-abstracted entities as indefinites; otherwise, generate them as FCIs (e.g., in English, 'any').

The first step of the algorithm is checked by performing all the possible combinations of λ-abstractions on the determiners (modifiers in our parlance, as discussed above) and by matching the abstracted logical forms of the utterance, to topoï premises. Then, the appropriate potentially useful λ-abstracted logical forms are kept for the third step of the algorithm, if the second step is not successful (i.e., the user can directly access the required topos from the non-abstracted logical form – i.e., non-indefinite utterance).

The second step of the algorithm is basically tackled by inspecting the content of the commitment store of the addressee after each dialogue *round*[2]: for each fact that the addressee is committed to (a fact is an SDRS, that represents the "view" of the addressee on the dialogue that has been taken place so far (Lascarides and Asher, 2009)), it is checked, based on the whole commitment store of the speaker, how this fact might have been "reached", from a logical point of view: if this fact could have been obtained by using a (optionally, β-reduced) topos as a premise[3], then it is inferred that this topos is already "fresh" in the memory of the addressee, hence, it is very likely that it is accessed again, if needed.

For this, we set, for each accessible rule or fact for performing reasoning, a priority, in inverse proportion with the recency of its access; this is practically handled by putting each newly accessed knowledge rule or fact in a stack. Then, when reasoning must be performed, first the stack is checked for each rule or fact and, if no appropriate rule or fact is found in the stack, then the commitment store is checked[4], and finally, the static knowledge base (e.g. a task or domain ontology for artificial agents (Caelen and Xuereb, 2007)). Once such a fact or rule is actually *used* in performing the reasoning, it is placed in the stack.

The results of the first two steps of the procedure are combined so that the appropriate λ abstraction of $K(\pi)$ is used as a premise for selecting, in the third step, the appropriate topos τ, that, according to the second step, the addressee might *not* have reached directly from the non-abstracted logical form.

By far the most difficult, the fourth step of the algorithm boils down to implementing Jayez's non-individuation constraint in the context of utterance production in multi-party dialogue. Deciding whether a β-reduction of a λ-abstracted utterance is blocked is a delicate task, because reasoning is needed on the joint commitments of the speaker *and* addressees. For this, we start from Jayez's formalization of NIC (Jayez and Tovena, 2004), where the hybrid logic "at" (@)operator is replaced by the notion of entailment, i.e., an expression such as $@_w\Phi$, read as 'Φ is true *at* w, where w is a (possible or real) world' is replaced by $w \models \Phi$, read as 'Φ is entailed from w', which is less restrictive than the former, because in our case we consider that the worlds are the interlocutors' public commitments, which are real from the perspective of each 'owner' of such a commitment store, and a clause is true 'at' such a commitment if it already is in that commitment. However, all we need here is that the clause can be inferred from that commitment and, optionally, static knowledge (from a knowledge base).

Thus, when a speaker L_{i_0} wants to produce an utterance to addressees L_i specified by a set $I \subseteq \{1, ..., N\} \setminus \{i_0\}$, where N is the number of speakers in the multi-party dialogue, the β-reduction of the λ-abstracted logical form $\overline{K(\pi)}$ is possible when either one of four constraints are

[2] A round in dialogue is a series of speech turns, produced by each speaker before the same speaker produces a new speech turn.

[3] The topoï are represented as λ-abstractions over entities, or over determiners of the entities – see above, but also (Popescu and Caelen, 2008).

[4] Note that this is not a technical redundancy, because in the stack of each interlocutor we put only rules or facts that she/he has *accessed*, i.e., read from the knowledge base *or* from her/his commitment store, not the facts *resulted* from these reasoning processes and placed in the commitment store.

met (they mirror Jayez's constraints (Jayez and Tovena, 2004)). First, we assume, in line with (Jayez and Tovena, 2004), that the logical form of the utterance π can be written as:

$K(\pi) = \mu_1(\{\exists|\forall\}K(P)\mu_2(K(Q)))$,
where μ_1 and μ_2 are modal operators (semantically, $\Box$ or $\Diamond$, and textually, verbs such as 'need', 'must', or, respectively, 'might', 'could')[5], and P and Q are clauses (that optionally contain negations, $\neg$). Thus, from the perspective of the speaker, L_{i_0} ($CS^{+}_{L_{i_0}}$ is the result of a single update of $CS_{L_{i_0}}$, the commitment store of L_{i_0}, and $\leftarrow$ is the assignment operation):

1.(a) $\bigcup_{\Phi}\{\Phi : CS_{L_{i_0}} \models \Phi \wedge CS_{L_{i_0}} \models \mu_1\mu_2\Phi\} \models \exists X : P(X) \wedge Q(X)$;

1.(b) $\bigcup_{\Phi}\{\Phi : CS_{L_{i_0}} \models \Phi \wedge CS_{L_{i_0}} \models \mu_1\mu_2\Phi\} \models \exists X : P(X) \wedge \neg Q(X)$;

2.(a) $CS_{L_{i_0}} \models \exists X : P(X) \wedge \forall\Gamma : \Gamma \equiv (\mu_1(\{\exists|\forall\}K(P')\mu_2(K(Q')))) \wedge CS^{+}_{L_{i_0}} \leftarrow CS_{L_{i_0}} \cup \{\Gamma\} \Rightarrow CS^{+}_{L_{i_0}} \models P(X) \wedge Q(X)$;

2.(b) $CS_{L_{i_0}} \models \exists X : P(X) \wedge \forall\Gamma : \Gamma \equiv (\mu_1(\{\exists|\forall\}K(P')\mu_2(K(Q')))) \wedge CS^{+}_{L_{i_0}} \leftarrow CS_{L_{i_0}} \cup \{\Gamma\} \Rightarrow CS^{+}_{L_{i_0}} \models P(X) \wedge \neg Q(X)$.

Again, following, in spirit, (Jayez and Tovena, 2004), for each sequent of the form $CS_L \models \Phi$, we rewrite the expressions above, by replacing CS_L with $\overline{CS_L}$, where $\overline{CS_{L_i}} \subseteq CS_{L_i}$ is the minimal commitment store such that $\overline{CS_{L_i}} \models \Phi$.

The first two constraints specify when utterances can describe referential situations associated with descriptive linguistic performance (i.e., a particular state of a world is described), whereas the latter two concern referential situations associated with exhaustiveness, i.e., utterances containing FCIs can satisfy the constraints 2 while given a universal interpretation, e.g. 'He read *any* book on the reading list' (lit. 'He read *every* book on the reading list')[6].

For extending this to multi-party dialogue, we consider that L_j, with $j \in J \subseteq \{1, ..., N\} \setminus \{i_0\}$, is an addressee of utterance π. Thus, $\overline{K(\pi)}$ is β-reducible if the facts that are not in both L_{i_0} and L_j's commitment stores at the same time, do not entail the falsity of $\exists X : P(X) \wedge \{|\neg\}Q(X)$ (the referentiality condition). In formal terms, this boils down to:

$CS_{L_{i_0}} \Delta CS_{L_j} \not\models \neg(\exists X : P(X) \wedge \{|\neg\}Q(X))$,
where Δ is the symmetric difference operator (for two sets A and B, $A \Delta B = (A \backslash B) \cup (B \backslash A)$). Otherwise, the β-reduction of the λ-abstraction $\overline{K(\pi)}$ of the semantic form $K(\pi)$ of utterance π is not possible. In a *cooperative* multi-party dialogue setting[7], if L_{i_0} addresses her current turn to a set $\{L_j : j \in J \subseteq \{1, ..., N\} \setminus \{i_0\}\}$ of interlocutors, then if there exists at least one j in J such that the referentiality condition above is fulfilled, then the indefinite marker is not realized as a FCI.

However, as pointed out in (Jayez and Tovena, 2004), the β-reduction of the λ-abstracted form of π is also blocked when, although the actual λ-abstracted π is referential, its vericonditional status is deduced from a fact (or a rule) that does not make reference to particular individuals (e.g., a *hard* topos (Popescu and Caelen, 2008), that is, a natural law of the form 'The more an x is greater than a value δ, the better x is').

We formalize this idea by stating that the β-reduction of the λ-abstracted form of π is also blocked when there is a hard topos τ such that $CS_{L_{i_0}} \models K(\pi) \wedge CS_{L_{i_0}} \setminus \{\tau\} \not\models K(\pi)$. However, according to (Jayez and Tovena, 2004), τ can also be simply a λ-abstracted clause with a non-β-reducible term (by virtue of the NIC, i.e., the constraints 1 and 2 above).

2.2 Multi-Party Dialogue Examples

The various situations that the mechanism proposed here has to deal with for generating FCIs are illustrated by the tree depicted in Figure 1, where decisions are made according to the following pragmatic constraints:

(i) the addressees (must / do not need to) access a topos for reaching a certain conclusion,

(ii) this topos (must / does not need to) be elicited by using indefinite constructions,

(iii) the NIC (is / is not) satisfied,

(iv) the indefinite utterance (depends on / does not depend of) a hard topos.

The numbers between parentheses identify the possible paths in the tree.

[5] These operators can also be void, e.g., for partially or purely assertive utterances.

[6] This example is borrowed from (Jayez and Tovena, 2004).

[7] The concept of "cooperative" dialogue is understood here in Gricean terms, i.e., the interlocutors are sincere, do not try to offend each other and respect the maxims of quality, quantity, relevance and manner (Asher and Lascarides, 2003).

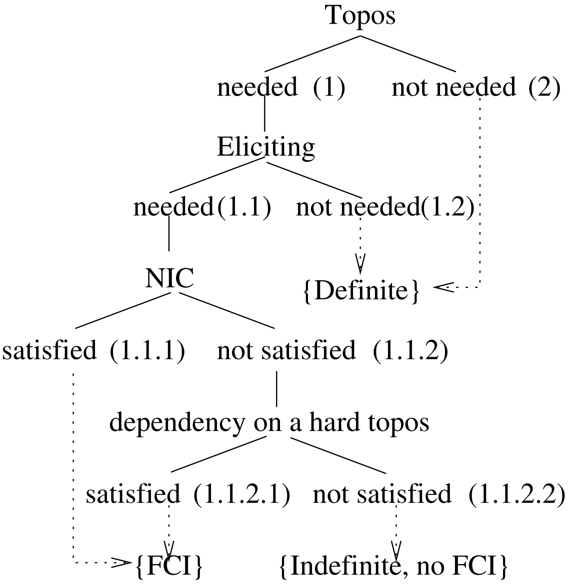


Figure 1: Decisions on the generation of FCIs.

From Figure 1 and from the manner the NIC is stated (in terms of public commitments), it results that in a dialogue, the number of FCIs produced by the interlocutors tends to lower as the dialogue progresses, unless new topoï are brought forth. This can be seen from the following example of dialogue between four speakers, concerning a book reservation topic.

L_1: Hello, I would like to read a book by A. Uthor.

L_2: Take this one, it is better than *any* other!

L_1: OK, but how about this one (another book, different from L_2's referent – n.a.), what do you think?

L_2: Yes, that one is good as well.

L_3: But, sir, how about the book "B. O. O. K." by A. Uthor?

L_1: That one as well, it is better than *any* other book.

L_4: Oh, yeah, all the customers have taken **any* book of this author!

L_3: I have read this one, it was better than *any* of A. Uthor's books!

The *any* in L_2's first turn is justified by the fact that we are in a situation that corresponds to path (1.1.1) on the tree in Figure 1. This is true, because L_2 needs to elicit the topos 'the more a book is better than other comparable books, the more interesting it is for the reader' or, in λ-abstracted form:

$$\tau = (\lambda X \lambda Y.(\textsf{[object]}(X) \wedge \textsf{equals}(X, \text{'book'}) \wedge \textsf{[patient]}(Y) \wedge \textsf{equals}(Y, \text{'book'}) \wedge \textsf{better}(X, Y)))_+, (\lambda Z \lambda T.(\textsf{[agent]}(Z) \wedge \textsf{equals}(Z, \text{'reader'}) \wedge \textsf{[object]}(T) \wedge \textsf{equals}(T, \text{'book'}) \wedge \textsf{interesting}(T, Z) \wedge [T \equiv X]))_+.$$

The predicate $\textsf{better}/2$ is a shorthand notation for the fact that the value of the first argument is higher than the value of the second, on a certain scale. The conjunct $[T \equiv X]$ is a procedure that states that T and X are identical variables.

In L_2's second turn, no indefinite construction is used, because the same topos τ as above is already present in L_1's stack of accessed knowledge ζ_{L_1} (see Section 2.1), as brought forth by L_2's first turn; hence, the situation corresponds to path (1.2) on the tree in Figure 1.

However, in its third turn, addressed to L_3, L_1 uses the FCI *any*, because the topos τ from above needs to be elicited again, as $\tau \notin \zeta_{L_3}$ yet (L_2's first turn was addressed to L_1 only, and we assume that if an utterance has not been addressed to an interlocutor, then the latter does not update its commitment store with the effects of this utterance).

L_4's use of *any* in its dialogue turn is not felicitous, because the NIC is violated. Indeed, the verb in the past ('has taken') entails that the concrete actions associated to that utterance are already present in L_4's commitment store:

$$CS_{L_4} \ni \exists X, Y : \textsf{[object]}(X) \wedge \textsf{[agent]}(Y) \wedge \textsf{equals}(X, \text{'book'}) \wedge \textsf{equals}(Y, \text{'customer'}) \wedge \textsf{borrow}(Y, X).$$

This situation thus corresponds to path (1.1.2.2) on the tree in Figure 1.

In the last turn of L_3, a similar argument as above entails that NIC is violated and hence, the situation cannot correspond to path (1.1.1) on the tree in Figure 1. However, since L_3's utterance is addressed to L_4, who needs the topos τ being elicited ($\tau \notin \zeta_{L_4}$), the utterance is felicitous by virtue of path (1.1.2.1), because it is dependent on a hard topos of the type: 'For an entity x that has a feature δ_x, the more δ_x is higher than a certain value δ, the more x is a better entity, on an appropriate scale'.

3 Discussion

In this paper we have proposed a framework for predicting the production of FCIs in multi-party dialogue. For this, we started from previous work of (Jayez and Tovena, 2004) on the interpretation of FCIs in monologue utterances. Thus, we extended this work to generation in multi-party dialogue situations. For this, several adjustments had to be made:

(i) establishing a reason for generating indefinite constructions (i.e., the need to determine the addressees to access certain topoï for deriving certain conclusions),

(ii) providing an interpretation for the concept of "world", *at* which a certain clause is true (i.e., assimilating such a world to the commitment stores of the speaker and the addressees),

(iii) restating the non-individuation constraint in terms of speakers commitments and of a model-theoretic entailment relation, instead of Blackburn's hybrid logic "at" operator (Blackburn, 2000), and

(iv) unifying the processing steps required to make the decision to generate a FCI, by using a lambda calculus-inspired formalism.

However, several points have been left untackled, with respect to the study of (Jayez and Tovena, 2004) concerning the interpretation of FCIs. Thus, the issue of the quantificational profile of FCIs has not been addressed: for instance, in French some FCIs are existential (such as 'n'importe quel' – lit. 'no matter which'), while others are universal (such as 'tout' – lit. 'any', as in 'Tout abus sera puni' – 'Any abuse will be punished').

Then, the thorny problem of FCIs applied on negative predicates has not been addressed either: for instance, constructions like 'I am sure John will refuse **any* book' (in French, 'Je suis sûr que Jean refusera **n'importe quel* livre') are not felicitous; investigating how one can know this in generation, without resorting to a bare list of negative predicates, remains a topic of further research.

In adapting Jayez's hybrid logic notion of truth at a world, we could have used a construction more akin to the original one in (Jayez and Tovena, 2004) by conflating λ-abstraction to "at" operators. Thus, in formalizing the fact that in a commitment store it is true that $\lambda X.\Phi(X)$ and that β-reduction is not possible in this expression, we could have written, for a speaker L_i, $@_{CS_{L_i}}[\lambda X.\Phi(X)\neg@]$, instead of $CS_{L_i} \models \lambda X.\Phi(X)\neg@$. But, if we had kept Jayez's account, we would have stated a stronger condition than one actually needs, namely that the λ-abstraction $\overline{\Phi}$ of Φ were actually already available as true in CS_{L_i}; however, we only need that $\overline{\Phi}$ be *entailed* from CS_{L_i}.

Concerning the differences between languages, for the English FCI 'any' one has two French rough translations, 'n'importe quel' and 'tout'. Jayez's study shows that the two French FCIs differ in that for 'tout', the set of potential alternative referents is not rigid (or a priori fixed, known), whereas for 'n'importe quel', the set of potential alternatives is fixed in advance, rigid. At a formal level, this situation could be captured by a logical form like:

$[\lambda X.(\textsf{[object]}(X) \wedge \textsf{equals}(X,...) \wedge ... \wedge \textsf{SubsetOf}(X, Set))\neg@]\wedge(...\wedge\textsf{value}(Set,\nu)\wedge...)$

for 'n'importe quel' (i.e., the λ-abstracted X belongs to a set Set that is a priori initialized with a value, ν). Consider for example: 'Prends *n'importe quel livre* [dans la bibliothèque – n.a.]' ('Take *no matter which / any* book [in the library]'), versus 'Prends **tout* livre [dans la bibliothèque]' ('Take *any* book [in the library]'). For 'tout', the conjunct concerning the properties of the set Set should be explicitly $\neg\textsf{value}(Set,\nu)$ or, in a Prolog-like environment, it would suffice that no restriction apply on Set. Take for example, 'Punis *tout* délit' ('Punish *any* misdemeanor') – unlike the set of possible books in the library, the set of misdemeanors is not a priori specified.

The framework presented in this paper can be applied in artificial agents as well, for endowing them with the capability of generating contextually-relevant answers in dialogues around a specified task (e.g., book reservation in a public library). Thus, dialogue modeling frameworks that explicitly address utterance generation as an important aspect (see, e.g., (Stent, 2001), or (Popescu, 2008)) could benefit from the proposal described in this paper for generating FCIs in dialogue. However, in order to do this, a series of adjustments might be appropriate, such as simplifying the computation of the commitment stores of the interlocutors. Indeed, keeping whole user-specific dialogue SDRSs in the commitment stores might be more than one needs. In the model-theoretic framework proposed in this paper, the entailment ($\models$) operation needs a model, i.e., a set of rules and facts in the left-hand side; the fine-grained SDRS representation (with scoping constraints over referents (Asher, 1993)) is not needed. We might thus adopt the strategy of computing the commitment stores in a manner akin to (Maudet et al., 2006).

Thus, we assume that the commitment store CS_{L_i} for each user L_i in a dialogue, contains the semantics of the utterances that L_i has produced, along with the semantics of the utterances from the other interlocutors, that L_i has agreed with (this is indicated by rhetorical relations between these utterances and utterances of L_i), and finally, along with the *negated* semantics[8] of the utterances of other speakers, that L_i did *not* agree with, along with the rhetorical relations that emphasize this fact (e.g. *P-Corr* (Plan Correction) or *Contrast* (Asher and Lascarides, 2003)).

For example, consider the following dialogue, between two speakers L_i and L_j, the former being a customer and the latter, a librarian:

L_j: You can still borrow three books!

L_i: So, I can take this one as well?

L_j: Yes, you can take it, sir.

This interaction contains a question of L_i, that is in an $Elab_q$ relation to the first utterance of L_j; the subsequent answer of L_j is in an *Elaboration* relation to the first utterance. The commitment store of L_i, after she had asked the question, is a set:

$CS_{L_i} = \{K(\pi_1), K(\pi_2), \Sigma_{Elab_q(\pi_1,\pi_2)}\}$, where π_1 and π_2 denote the first utterance of L_j and the first utterance of L_i (the question) respectively, and $\Sigma_{Elab_q(\pi_1,\pi_2)}$ denotes the SDRT semantics of the rhetorical relation $Elab_q(\pi_1, \pi_2)$, which specifies that utterance π_2 is a question such that any relevant answer elaborates on utterance π_1 (Asher and Lascarides, 2003).

References

Jean-Claude Anscombre. 1995. Topique or not topique: formes topiques intrinsèques et formes topiques extrinsèques. *Journal of Pragmatics*, 24:115–141.

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

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

Patrick Blackburn. 2000. Representation, reasoning and relational structures: A hybrid logic manifesto. *Logic Journal of the IGPL*, 8:339–365.

Jean Caelen and Anne Xuereb. 2007. *Interaction et pragmatique - jeux de dialogue et de langage*. Hermès Science, Paris.

Anastasia Giannakidou. 2001. The meaning of free choice. *Linguistics and Philosophy*, 24:659–735.

Jerry Hobbs, Mark Stickel, and Douglas Appelt. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Jacques Jayez and Lucia Tovena. 2004. Free choiceness and non-individuation. *Linguistics and Philosophy*, 28:1–71.

Rodger Kibble. 2006. Reasoning about propositional commitments in dialogue. *Research on Language and Computation*, 4(2-3):179–202.

Alex Lascarides and Nicholas Asher. 2009. Grounding and correcting commitments in dialogue. *Journal of Semantics*, (to appear).

Nicolas Maudet, Philippe Muller, and Laurent Prvot. 2006. Social constraints on rhetorical relations in dialogue. In *Proceedings of the 2nd SIGGen Workshop Constraints in Discourse*, pages 133–139, Maynooth, Ireland, July 7-9. ACL.

Vladimir Popescu and Jean Caelen. 2008. Argumentative ordering of utterances for language generation in multi-party human-computer dialogue. *Argumentation*, 23(2):205–237.

Vladimir Popescu. 2008. *Formalisation des contraintes pragmatiques pour la génération des énoncés en dialogue homme-machine multilocuteurs*. Ph D Thesis, Grenoble Institute of Technology.

Amanda Stent. 2001. *Dialogue Systems as Conversational Partners: Applying Conversational Acts Theory to Natural Language Generation for Task-Oriented Mixed-Initiative Spoken Dialogue*. Ph D Thesis, University of Rochester.

[8] The negation is defined in a special manner, for handling interrogative utterances as well. Let us consider for example a question as: 'Is this book OK for you?', labeled π. Since it is a question, the logical form $K(\pi)$ of the utterance contains a predicate which takes a non-initialized variable as argument ((Asher and Lascarides, 2003) use λ-abstracted variables in questions):

$\exists X, Y, Z \; : \; \mathsf{[patient]}(X) \wedge \mathsf{[object]}(Y) \wedge \mathsf{equals}(Y, \text{'book'}) \wedge \mathsf{want}(X, Y, Z) \wedge \mathsf{equals}(Z, \text{'?'})$.

Here, the non-initialized variable is the boolean Z that contains the truth value of the predicate $\mathsf{want}/3$, which is true if the entity designated by its first argument wants the entity designated by the second argument. The negation of such a question does not boil down to negating each predicate in the conjunction, and then substituting the conjunctions with disjunctions, but to assigning the value 0 to the boolean Z; hence, in our case, $\neg K(\pi)$ has the same form as $K(\pi)$, excepting the last predicate, which has the form $\mathsf{equals}(Z, 0)$.

A comparison of addressee detection methods for multiparty conversations

Rieks op den Akker
Human Media Interaction
University of Twente
Enschede, the Netherlands
infrieks@cs.utwente.nl

David Traum
Institute for Creative Technologies
University of Southern California
Marina Del Rey, CA 90292 USA
traum@ict.usc.edu

Abstract

Several algorithms have recently been proposed for recognizing addressees in a group conversational setting. These algorithms can rely on a variety of factors including previous conversational roles, gaze, and type of dialogue act. Both statistical supervised machine learning algorithms as well as rule based methods have been developed. In this paper, we compare several algorithms developed for several different genres of multiparty dialogue, and propose a new synthesis algorithm that matches the performance of machine learning algorithms while maintaining the transparency of semantically meaningful rule-based algorithms.

1 Introduction

Detecting who is being addressed, i.e. who the speaker is talking to, is non-trivial in multi-party conversations. How speakers make clear who they address depends on the conversational situation, knowledge about other participants, inter-personal relations, and the available communication channels.

In this paper we present rule based methods for automatic addressee classification in four-participant face-to-face meetings. A rule based method is more transparent than the statistical classifiers. It synthesizes empirical findings of addressing behavior in face-to-face conversations. We have analysed addressing behavior in small design group meetings, and we have evaluated our methods using the multi-layered multi-modal annotated AMI meeting corpus (Carletta, 2007). The same multi-modal corpus has been used for developing statistical addressee classifiers using (Dynamic) Bayesian Networks (Jovanovic, 2007). The (Dynamic) Bayesian Network classifiers have performances ranging from 68-77%, depending on the types of features used and whether it is a static network, using Gold Standard (i.e. the manual annotated) values for addressees of previous acts, or dynamic, using own predicted values for addressees of previous acts in the dialogue. Our best performing rule-based method has an accuracy of 65%, which is 11% over the baseline (always predict that the group is addressed).

Performance measures don't tell much about the confidence we can have in the outcome in particular cases. A reliability analysis of the manually annotated data that is used for training and testing the machine classifier can reveal in what cases the outcomes are less reliable. In specific situations, such as when the speaker uses *"you"*, or when the speaker performs an initiating act, supported by visual attention directed to the addressed partner, the method outperforms the statistical methods. Our method uses speaker's gaze behavior (focus of attention), dialogue history, usage of address terms as well as information about the type of dialogue act performed by the speaker to predict who is being addressed.

2 How do speakers address others?

Addressing occurs in a variety of flavors, more or less explicitly, verbally or non-verbally. Thus, sometimes deciding whether or not the speaker addresses some individual partner in particular is far from a trivial exercise. Within a single turn, speakers can perform different dialogue acts (i.e. they can express different intentions), and these dialogue acts can be addressed to different participants. In small group discussions, like those in the AMI meetings with 4 participants, most contributions are addressed to the whole group. But sometimes speakers direct themselves to one listener in particular. Some important motivations for individual addressing are that the group members bring in different expert knowledge and that they have different tasks in the design process. If someone says to a previous speaker *"can you clarify what you just said about ..."* it is clearly addressed to that previous speaker. This doesn't rule out that a non-addressed participant takes the next turn. But generally this will not happen in an unmarked way.

The basis of our concept of addressing originates from Goffman (Goffman, 1981). The addressee is the participant *"oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants"*. Thus, according to Goffman, the addressee is the listener the speaker has selected because he expects a response from that listener. The addressee coincides with the one the speaker has selected to take the next turn. But addressing an individual does not always imply turn-giving, such as can be seen in (1), a fragment of Alice's speech, in a conversation between Alice, Ben and Clara.

(1) Yes, but, as Clara already said earlier
gaze: < Ben >

correct me if I'm wrong,
gaze: < Clara >

the price of working out *your* proposal is too high for us, so ...
gaze: < Ben >

In (1), the main dialogue act performed by Alice is addressed to Ben. Although Alice's contribution is to the whole group, it is meant especially as a reaction to the preceding proposal made by Ben, and she directs herself to Ben more than to the others. That is why we say that in this case *the dialogue act is addressed to* Ben. Note that *"your"* refers to Ben as well, and also Alice's gaze is directed at Ben. Alice is especially interested to see how Ben picks up and validates the concern that she expresses. The dialogue act expressed by the embedded phrase is addressed to Clara. Although, Alice explicitly invites Clara to correct her, which is indicated by the gaze shift during this clause, after mentioning her name, she doesn't yield the turn, but continues speaking.[1]

Speakers use different procedures to make clear who they address. The selection of this procedure depends on (a) what the speaker believes of the attentiveness of the listener(s) to his talk, and (b) the speaker's expectation about the effect his speech has on the listener that he intends to address. For example if A just was just asked a question by B then A will assume that B is attending his answer. In a face-to-face meeting A will usually monitor how B takes up his answer and will now and then gaze at B as his visual focus of attention is not required for competing foci of interest. Lerner distinguished *explicit* addressing and *tacit* addressing. To characterize the latter he writes: *"When the requirements for responding to a sequence-initiating action limit eligible responders to a single participant, then that participant has been tacitly selected as next speaker. Tacit addressing is dependent on the situation and content."* (Lerner, 2003).

An example from our corpus is when a presenter says *"Next slide please"* during his presentation, a request that is clearly addressed to the one who operates the laptop. Tacit addressing is most difficult for a machine, since it requires to keep track of the parallel activities that participants are engaged in.

Explicit addressing is performed by the use of vocatives (*"John, what do you think?"*) or, when the addressee's attention need not be called, by a *deictic personal pronoun*: *"What do you think?"*. There is one form of address that always has the property of indicating addressing, but that does not itself uniquely specify *who* is being addressed: the *recipient reference term* "you" (Lerner, 2003). *The use of "you" as a form of person reference separates the action of "addressing a recipient" from the designation of just who is being addressed. In interactional terms, then, "you" might be termed a recipient indicator, but not a recipient designator. As such, it might be thought of as an incomplete form of address* (Lerner, 2003). Gaze or pointing gestures should complete this form of addressing. These analytical findings motivated the selection of rules for addressee detection.

3 Automatic Addressee Recognition

The starting point of our design of a rule based algorithm for addressee prediction was Traum's algorithm as presented in (Traum, 2004), shown in (2). This algorithm was meant to be used by virtual agents participating in a multi-party, multi-

[1] The rules for dialogue act segmentation used in the AMI corpus do not cover dialogue act units embedded in other units, as is the case in this made up example.

conversation environment (Traum and Rickel, 2002; Rickel et al., 2002), in which conversations could be fluid in terms of starting and stopping point and the participants that are included. The algorithm only uses information from the previous and the current utterance; thus no information about uptake of the act performed by the current speaker. The method doesn't use speaker gaze. In initial versions of the virtual world, the agents did not have access to human gaze. Even when gaze is available, it is non-trivial to use it for addressee-prediction, because there are many other gaze targets in this dynamic world other than the addressee, including monitoring for expected events in the world and objects of discussion (Kim et al., 2005; Lee et al., 2007).

(2) 1 If utterance specifies a specific addressee (e.g. a vocative or utterance of just a name when not expecting a short answer or clarification of type person) then Addressee = *specified addressee*.

2 else if speaker of current utterance is the same as the speaker of the immediately previous utterance then *Addressee = previous addressee*

3 else if previous speaker is different from current speaker then *Addressee = previous speaker*

4 else if unique other conversational participant (i.e. a 2-party conversation) then *Addressee = that other participant*

5 else *Addressee = unknown*

Traum's algorithm had good performance in the Mission Rehearsal Exercise domain. (Traum et al., 2004) reports F-scores of from 65% to 100% in actual dialogues, using noisy speech recognition and NLU as input). In this paper we will examine to what degree this algorithm generalizes to a different sort of multi-party corpus, and what can be done to improve it.

4 The AMI meeting corpus

The manually annotated conversations that we analysed are from the AMI meeting corpus; (Carletta, 2007). There are 14 four-participant face-to-face meetings, where participants are mostly sitting at a rectangular table. Twelve of the 14 meetings were recorded in one meeting room, the other two in two other rooms.

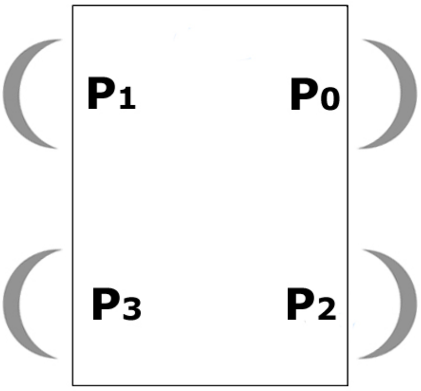


Figure 1: Fixed seating positions around a square table.

The 14 meetings were annotated with dialogue acts, addressee information as well as focus of attention of participants (FOA). Utterances are segmented in consecutive DA-segments. The segments are assigned a type. Dialogue acts types are: *Inform*, *Elicit-inform*, *Suggest*, *Offer*, *Elicit-offer-or-suggestion*, *Assess*, *Elicit-assessment*, *Comment-about-understanding*, *Elicit-comment-about-understanding*, *Be-positive*, and *Be-negative*. Other labels for DA-segments are *Backchannel*, *Stall*, and *Fragment*. The *Other* label was used when an utterance could not be labeled by one of the list of dialogue labels.

For important contentful dialogue acts (i.e. excluding Stall, Fragment and Backchannel[2] acts) the annotators have indicated whether the DA was addressed to the whole group (*G-addressed*), or to some individual (*I-addressed*), in which case they indicated who was being addressed. Annotators could also label the addressee as Unknown, but because there was very little reliability in this category, we combined it with the G-addressed category.

I-addressed acts are marked in terms of table position of the person being addressed: P0, P1, P2 or P3. Figure 1 shows the layout of the fixed seating positions in the meeting rooms.

Focus of attention (FOA) can be on one of these participants or at the white board, at the table, or at no specific target. Words and dialogue acts were time aligned, so that it can be computed what the

[2] Backchannel acts were assumed to be addressed to the "previous speaker" and were therefore not annotated in the AMI corpus.

focus of attention is of each of the participants during a specific time frame. Note that neither the addressee annotation, nor the FOA annotation allowes a multiple target label. This could be a possible cause of confusion between annotators in case a sub-group is addressed by the speaker. However, subgroup addressing hardly occurs in the data.

4.1 Reliability of the AMI annotations.

Since we based our models on the analysis of a human annotated corpus, and since we also tested them on manual annotated data, the question arises how much human annotators agree on the addressee labeling. How does the accuracy depend on the annotator? Are there specific situations in which results are more reliable than in others? (Jovanovic, 2007) (Chapter 3.4) contains a detailed examination of the inter-annotator agreement of the codings of the AMI corpus. We present some highlights here.

We compare three annotations of one and the same meeting in our corpus. Most confusions in the addressing labeling are between *I-addressed* and *G-addressed*, If annotators agree that the DA is I-addressed then they agree on the individual as well. We found that for both dialogue acts and addressee identification, reliability is higher for some decisions than others. Table 1 shows Krippendorff's alpha values (Krippendorff, 2004) for inter-annotator agreement for each pair of annotators. The statistics are computed on the subsets of pairwise agreed DA-segments: cases in which the annotators did not agree on the segmentation are left out of this analysis.[3]

Table 1 shows that annotators consistently agree more on the addressing of elicit acts (3rd column) than on DAs in general (2nd column). For the subset of elicit acts, when annotators agree that an elicit is *I-addressed* (which happens in 50-80% of the agreed elicit acts, depending on the annotators), than they agree on the individual that is addressed, without exception. Addressing is a complex phenomenon and we believe that the mediocre agreement between addressee annotations is due to this complexity. In particular, we observed that some annotators prefer to see a response act as I-addressed at the speaker of the initiating act, where for others the content is more decisive (does, for example, the question address an issue that is relevant for the whole group or does it only concern the speaker and his addressee?)

pair	adr	adr-eli	da	da-eli
a-b	0.56(412)	0.67(31)	0.62(756)	0.69
a-c	0.45(344)	0.58(32)	0.58(735)	0.64
b-c	0.46(430)	0.62(53)	0.55(795)	0.80

Table 1: Alpha values (and numbers of agreed DA segments) for the three pairs of annotators; for addressing, addressing of elicit acts only, dialog acts (all 15 DA classes), and elicit vs non-elicit acts (5th column).

As expected (because speakers FOA is an important indicator for addressing) annotators agree more on the addressee in situations with a clear speaker gaze at one person. We refer to (Reidsma et al., 2008) for more details.

Annotators agreed rather well in telling elicit acts from other types of dialogue acts, as is shown in Table 1, 5th column. This DA type information is thus quite reliable.

Focus of attention annotation was done with high agreement, so we can take gaze target information as reliable information, with a timing precision of about 0.5 sec. (See (Jovanovic, 2007) for a detailed reliability analysis of the FoA annotation.)

5 Dialog structure

Gupta et al. present experiments into the resolution of *"you"* in multi-party dialog, and they used the same part of the scenario based AMI meetings as we did. They distinguish between generic and referential uses of "you"; and, the referential uses, they try to classify automatically by identifying the referred-to addressee(s): either one of the participants, or the group. All results are achieved without the use of visual information. (Gupta et al., 2007). Gupta et. al. expected that *the structure of the dialog gives the most indicative cues to addressee: forward-looking dialog acts are likely to influence the addressee to speak next, while backward-looking acts might address a recent speaker*. In a similar way Galley et al. (Galley et al., 2004) also used the dialog structure present in adjacency pairs as indicative for ad-

[3] A better analysis of addressing (dis)agreements might be based on speaker turns or sequences of dialogue acts, because (a) many segmentation disagreements do not affect addressing, and (b) the distribution of DA types over the set of agreed segments is different from the distribution of DA types over the whole corpus (agreed segments are shorter in the mean)

dressees: the speaker of the a-part would likely be the addressee of the b-part and the addressee of the a-part would likely be the speaker of the b-part (dyadic pattern $ABBA$). In the one dimensional DA schema that we used on the AMI corpus there is no clear distinction between Backward Looking (BL) and Forward Looking (FL) "types" of dialogue acts. However, we may consider the *elicit types* as FL types of DAs. Typical BL DA types are *Comment about Understanding* and to a lesser extend *Assessments*. The other DA types can be assigned to BL as well as to FL utterances, but if an *Inform* act follows an *Elicit-Inform*, the last one more likely has a BL function. The AMI corpus is also annotated with dialog relation pairs, much like the classical adjacency pairs: they are typed relations (the type carries polarity information: is the response of the speaker positive or negative, or partial negative/positive the target act, or does the speaker express uncertainty), and related DAs need not be adjacent (i.e. there can be other DAs in between). In the AMI corpus the speaker addressee pattern $ABBA$ fits 60% of all adjacency pairs, which makes them a good feature for addressee prediction. We will however not use this adjacency pair information because this information is as hard to obtain automatically as addressee information.

The total number of DAs in our corpus is 9987, of which 6590 are contentful DAs (i.e. excluding *Stall*, *Fragment*, and *Backchannel*, which did not get an addressee label assigned). Of these, 2743 are addressed to some individual (*I-addressed*); the others are addressed to the Group (*G-addressed*).

In 1739 (i.e. 63%) cases of the 2743 *I-addressed* dialog acts, the addressed person is the next speaker (the current speaker might also perform additional dialogue acts before the next speaker's speech).

Forward looking DAs that are I-addressed are more selective for next speaker than I-addressed DAs in general. There are 652 elicit acts in our corpus. Of these, 387 are $I - addressed$. In 302 cases (78%) the addressee is the next speaker. This is indeed substantially more than the mean (63%) over all DA types.

Speaker's gaze is an important indication for whom they address their DA. (see (Kendon, 1967), (Kalma, 1992), (Vertegaal and Ding, 2002)). In our corpus, speakers gaze three times more at their addressee than at other listeners.

6 Algorithms for Addressee Identification in the AMI corpus

In this section, we compare several different algorithms for recognizing the addressee in the AMI corpus.

6.1 Jovanovich's DBN

In (Jovanovic, 2007), Dynamic Bayesian Networks, (D)BNs, were used to classify the addressee based on a number of features, including context (preceding addressee and dialogue acts, related dialogue acts), utterance features (personal pronouns, possessives, indefinite pronouns and proper names), gaze features, and the types of meeting actions, as well as topic and role information. The best performance for all features yielded roughly 77% accuracy on the AMI corpus. The best performing BNs uses "Gold Standard" values of addressees of previous and related DAs. The DBNs uses own predicted addressee values for these features. For comparison purposes, we recoded this approach using the Weka toolkit's implementation of BayesNets, using the same features as our other algorithms had available: no adjacency pair information, no topic role and role information. The BNs achieved accuracies of 62% and 67%.

6.2 Traum's algorithm

We re-implemented Traum's algorithm shown above in (2). While Traum's algorithm had good performance in the Mission Rehearsal Exercise domain it has very bad performance in the AMI domain, as shown in the next section. Why is this? Interaction styles are different across the two domains. Patterns of speaker turns are different and that is caused by the different scenarios. In the meeting scenario, there is a much more static environment, so gaze is a better predictor, which was not used in Traum's algorithm. More importantly, Traum's algorithm does not adequately account for speech addressed to a group rather than (primarily) to a single participant, while this formed the majority of the AMI data. Traum's algorithm indicates group, only if a group addressing term (e.g. "you all") is used , or the group is the previous addressee, or if the addressee is unknown. There are also more frequent uses of address terms in the Mission Rehearsal context than in the AMI meetings.

6.3 GazeAddress

The method *gazeAddress* predicts the addressee of a DA using only information about speaker's cumulative focus of attention over the time period of the utterance of the speech act. It predicts the addressee as follows. If there is an individual B such that the speaker A gazes for more than 80% of the duration of his dialogue act in the direction of B, it is assumed that the dialogue act performed by A is I-addressed to B. Otherwise, the speaker is assumed to address the group (G). To obtain the best threshold value, we ran several tests with different values for the threshold and computed recall and precision for the Group class as well as for the individual class values. Going up from 50% to 80%, the precision and recall of the single addressee and group addressee identification slowly improves. After that the precision of the single addressee does not improve nor decline much. But the recall and precision of the group identification gets a lot worse. We used 80% as threshold value in subsequent experiments.

6.4 The Addressee Prediction Algorithm

Our Addressee Prediction Algorithm (APA) that returns the addressee of the current dialogue act (DA) runs as follows. It returns ”G” when it predicts that DA is *G-addressed.* If it predicts that the DA is *I-addressed* it returns the table position of the individual participant.

```
(1) (address term used)
if (containsAddressTerm(DA)){
          return referredPerson;}

(2) (same speaker turn)
if (daSpeaker=prevDASpeaker) {
     if (gazeAddress=previousADR ){
          return previousADR;
     } else{
          return "G";}}

(3) (other speaker)
     if (daSpeaker=previousADR)
          return prevDASpeaker;
     if (gazeAddress!=null && you)
          return foa;
     if(gazeAddress=prevDASpeaker){
          return prevDASpeaker;}}
```

In (1) it is tested whether the speaker uses an address term (name or role name of a participant). If so, the referred person is returned as the addressee. Clause (2) fires when the current DA is by the same speaker as the previous one. If the *gazeAddress* method would return for an individual (the value of foa) and this is the same one as the person addressed in the previous act then this one is returned. Clause (3) fires when a speaker change occurred. If the previous speaker addressed the current speaker, then the previous speaker is the returned addressee. If not, when the DA contains “you” and the *gazeAddress* method returns some individual then this one is returned. If *gazeAddress* decided for an individual and this equals the previous speaker then this one is returned. Otherwise, the group is addressed. We experimented with some variations of this method. A slight improvement was obtained when we have a special treatment for forward looking DA types. Analyses of the corpus reveals that elicit acts are more frequently used as forward looking acts. In that case, the decision is based on *gazeAddress* not taking into account the previous speaker.

7 Results

Table 2 shows the performance of four methods from the previous section in terms of Recall, Precision, and F-score for group, participant P0 (the most challenging of the participants), and overall accuracy (i.e. percentage correct).

	Group			P_0			
Method	R	P	F	R	P	F	Acc
Traum's	12	92	22	70	31	44	36
BayesNet	65	73	69	62	45	52	62
GazeAdr	66	65	65	36	43	40	57
APA	89	65	75	26	62	36	65

Table 2: Performance table of the four methods for addressee prediction. N=6590 (DAs). Baseline (always Group) is 54%.

We can see from table 2 that APA has the highest overall accuracy for recognizing the addressees of each dialogue act in sequence. However it is the lowest of the four in recognizing P0. In table 3 we look at the importance of recognizing the previous addressee correctly, by supplying the Gold Standard value for this feature rather than the value calculated by the respective algorithms. Traum's algorithm shows the biggest improvement in this case, while APA improves the least.

Table 4 gives an overview of the performances of the two new methods - *gazeAddress* and APA - on various subclasses of the data set. ALL is the set of all contentful dialogue acts; ELI is the set of elicit acts; YOU is the set of acts containing

	Group			P_0			
Method	R	P	F	R	P	F	Acc
Traum	47	88	61	67	42	52	56
BayesNet	66	85	75	73	50	60	67
APA	86	68	76	34	61	44	67

Table 3: Performance table when using *Gold Standard* values for previous addressees of the three methods making use of previous addressee information.

	DA-SETS			
	ALL	ELI	YOU	ELI-Y
N	6590	652	1061	166
Gaze	57	62	62	68
APA	65	62	68	69

Table 4: Accuracy values of methods on various sets of dialogue acts

“you”; ELI-YOU is the subset of eliciting acts that contains *“you”*.

We see that for the subsets of dialogue acts that contains *“you”* as well as for the mostly forward looking elicit acts APA performs better than the mean performance of APA over all DAs, and even better than the Dynamic Bayesian Networks. The average accuracy of APA for DAs with *“you”* over all the meetings is 68%.

The results vary over the set of meetings and a factor that causes this is the percentage of G-addressed DAs in the meeting. In general, the performance raises with the percentage of G-addressed DAs.

How does the performance depend on the annotators? For the one meeting IS1003d that was annotated by all three annotators involved, the accuracies of method APA were 61, 75 and 60. For the method *gazeAddress* they were 58, 66, 57, respectively. Also here the data annotated by the annotator who had a preference for the G-label over one of the individual labels has a higher accuracy.

7.1 Further research

A more detailed analyses of the results of method *gazeAddress* reveals that the recall and precision values depend on the position of the speaker as well as on the *relative position* of the person gazed at most by the speaker. In future work, we will examine both the role of the meeting participant and the physical locations in terms of their effect on performance and possibly augmentations to the algorithms. Using the same part of the AMI corpus, (Frampton et al., 2009) classify referential uses of “you” in terms of relative position of addressees from the view point of the speaker. They achieve good results in finding the I-addressee of those speech acts that contain such a referential use of “you”. Note that our method does not identify if an occurrence of “you” is referential, so it is hard to compare the results.

8 Conclusion

We have seen that a rule based method can predict addressing with an accuracy that is comparable with that of the purely statistical methods using dynamic Bayesian networks. It is hard to obtain a high precision and recall for individual addressing. Although slight improvements can be expected if we take into account the relative positions of speakers and addressees when using gaze direction of speakers as indicator for who is being addressed, substantial improvements will likely be only possible when the system has more knowledge about what is going on in the meeting.

Knott and Vlugter implemented in their multi-agent language learning system a rule-based method for addressee detection which is similar to the one of Traum, see (Knott and Vlugter, 2008). In their system, agents make frequent use of address terms, and they do sub-group addressing, unlike the agents in the face-to-face meetings. Sub-group addressing remains a challenging issue for multi-agent dialogue systems.

Comparative analysis of various human annotations of the same data is very informative for clarifying such abstract and complex notions as addressing is. Such an analysis is important to improve our understanding of the phenomena and to sharpen the conceptual definitions that we use. Results inferred from statistics and patterns in relations between annotated data should take the difficulties that annotators have in applying the general notions in concrete new situations into account.

Acknowledgments

The authors are gratefull to the anonymous reviewers for their comments and suggestions. Traum’s effort described here 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. The work by R. op den Akker is supported by the EU IST Programme Project FP6-0033812 (AMIDA). This paper only reflects the authors views and funding agencies are not liable for any use that may be made of the information contained herein.

References

Jean C. Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, May.

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is "you"? combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 273–281, Athens, Greece, March. Association for Computational Linguistics.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain.

Erving Goffman. 1981. Footing. In *Forms of Talk*, pages 124–159. Philadelphia: University of Pennsylvania Press.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky. 2007. Resolving "you" in multiparty dialog. In *Proceedings of 8th SigDial Workshop*, pages 227–230.

N. Jovanovic. 2007. *To whom it may concern. Addressee identification in face-to-face meetings*. Ph.D. thesis, University of Twente, Enschede, The Netherlands, March.

A. Kalma. 1992. Gazing in triads: A powerful signal in floor apportionment. *British Journal of Social Psychology*, 31(1):21–39.

A. Kendon. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

Youngjun Kim, Randall W. Hill, and David R. Traum. 2005. Controlling the focus of perceptual attention in embodied conversational agents. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 1097–1098, New York, NY, USA. ACM.

Alistair Knott and Peter Vlugter. 2008. Multi-agent human-machine dialogue: issues in dialogue management and referring expression semantics. *Artif. Intell.*, 172(2-3):69–102.

K. Krippendorff. 2004. *Content analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage, 2nd edition.

Jina Lee, Stacy Marsella, David R. Traum, Jonathan Gratch, and Brent Lance. 2007. The rickel gaze model: A window on the mind of a virtual human. In Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé, editors, *IVA*, volume 4722 of *Lecture Notes in Computer Science*, pages 296–303. Springer.

Gene H. Lerner. 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, 32:177–201.

D. Reidsma, D. K. J. Heylen, and H. J. A. op den Akker. 2008. On the contextual analysis of agreement scores. In J-C. Martin, P. Paggio, M. Kipp, and D. K. J. Heylen, editors, *Proceedings of the LREC Workshop on Multimodal Corpora, Marrakech, Morrocco*, pages 52–55, Paris, France, May. ELRA.

J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and W. Swartout. 2002. Towards a new generation of virtual humans for interactive experiences. *Intelligent Systems*, 17:32–36.

David R. Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first International Joint conference on Autonomous Agents and Multiagent systems*, pages 766–773.

David R. Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1699–1702.

D. Traum. 2004. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, pages 201–211. Springer-Verlag.

Roel Vertegaal and Yaping Ding. 2002. Explaining effects of eye gaze on mediated group conversations:: amount or synchronization? In *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 41–48, New York, NY, USA. ACM.

A domain ontology based metric to evaluate spoken dialog systems

Jan Kleindienst, Jan Cuřín, Martin Labský
IBM Research
Prague, Czech Republic
{jankle, jan_curin, martin.labsky}@cz.ibm.com

Abstract

Current methods and techniques for measuring performance of spoken dialog systems are still very immature. They are either based on subjective evaluation (Wizard of Oz or other usability studies) or they are borrowing automatic measures used in speech recognition, machine translation or action classification, which provide only an incomplete picture of the performance of the system. We introduce a method for quantitative evaluation of spoken dialog systems that utilizes the domain knowledge encoded by a human expert. The evaluation results are described in the form of a comparison metric consisting of domain coverage and dialog efficiency scores allowing to compare relative as well as absolute performance of a system within a given domain. This approach has the advantage of comparing incremental improvements on an individual dialog system that the dialog designer may want to verify along the way. In addition, the method allows to cross-check the performance of third-party dialog systems operating on the same domain and understand the strong and weak points in the dialog design.

1 Introduction

Research in the field of conversational and dialog systems has a long tradition starting in 1966 with Weizenbaum's Eliza (Weizenbaum, 1966). More recently, research in spoken dialog systems has tackled more ambitious domains, such as problem solving (Allen et al., 2007), navigation (Cassell et al., 2002), or tutoring systems (Graesser et al., 2001). This paper is organized as follows: in the introduction we outline our motivation and the principle of the proposed method. Section 2 describes in detail the proposed dialog score and its computation. Section 3 presents a case study in the music management domain and demonstrates the application of the scoring to a real-world task. We discuss the correlation of the proposed metric with subjective evaluation in Section 4, and conclude by Section 5.

1.1 Rationale

Current methods and techniques for measuring performance of speech-enables user interfaces are still very immature. They are either based on subjective evaluation (Wizard of Oz or other usability studies) or they are borrowing automatic measures used in speech recognition, machine translation or action classification, which provide only incomplete picture of the performance of the system. Nowadays, dialog systems are evaluated by action classification error rate (Jurafsky and Martin, 2008), by techniques that measure primarily dialog coherence (Gandhe and Traum, 2008), by methods based on human judgment evaluation, such as PARADISE (Walker et al., 2000; Hajdinjak and Mihelific, 2006), or using reward function values (Rieser and Lemon, 2008; Singh et al., 1999). What is particularly missing in this area are (1) a measurement of performance for a particular domain, (2) possibility to compare one dialog system with others, and (3) evaluation of a progress during the development of dialog system. The score we present attempts to address all three issues.

2 The Proposed Method of Dialog System Evaluation

The proposed dialog score (DS) consists of two ingredients both of which range from 0 to 1:

- Domain Coverage (DC) score,
- Dialog Efficiency (DE) score.

The DC expresses how the evaluated system covers the set of tasks in the ontology for a particular domain, while the DE indicates the performance of the evaluated system on those tasks supported by the system over user test sessions.

We describe both scores in the following subsections. Note that the results of domain coverage and dialog efficiency may be combined into a single compound score to attain a single overall characteristic (the eigen value) of the assessed dialog system.

2.1 Scoring of Domain Coverage

The domain coverage (DC) is a sum of weights of the tasks supported by the system (S) over the sum of weights of all tasks from the ontology (O).

$$DC(S,O) = \frac{\sum_{t \in supported\_tasks(S,O)} w_t}{\sum_{t \in all\_tasks(O)} w_t} \qquad (1)$$

Table 1 shows a sample domain task ontology for the music management domain that shows the raw points assigned by a domain expert and their normalized versions that are used to assess the relative importance of individual tasks. The expert may control the weights of whole task groups (such as Playback control) as well as the weights of individual tasks that comprise these groups. Generally, the ontology can have more than two levels of sub-categorization that are shown in the example. So far our task ontologies have been limited to hierarchical sets of weighted tasks. We are however investigating whether introducing domain concepts, such as “song”, “album” or “playlist”, and relations among them, can help derive possible user tasks and their weights semi-automatically.

2.2 Scoring of Dialog Efficiency

The actual efficiency of a dialog is measured using the number of dialog turns (Le Bigot et al., 2008; Nielsen, 1994) needed to accomplish a chosen task. In spoken dialog systems, a dialog turn corresponds to a pattern of a user speech input followed by the system’s response. We introduce a generalized penalty turn count (PTC) that measures overall dialog efficiency by incorporating other considered factors: number of help requests, number of rejections, and user and system reaction times, and in the future possibly also others.

Table 1: Speech-enabled reference tasks for the music management domain. (Tasks are divided into groups. Both the group as well as tasks within the group are assigned relative importance points (weights) by an expert. These points are normalized to obtain per-task contribution to the domain’s functionality. ITC shows ideal turn count range for each task.)

Description	Points	Contr	ITC
Volume	2	15.50	-
relative	2	6.20	1
absolute	1	3.10	1
mute	2	6.20	1
Playback	4	31.01	-
play	3	7.75	1
stop	3	7.75	1
pause	1.5	3.88	1
resume	1.5	3.88	1
next, previous track	1	2.58	1
next, previous album	1	2.58	1
media selection	1	2.58	1
Play mode	0.5	3.88	-
shuffle	1	1.94	1
repeat	1	1.94	1
Media library	6	46.51	-
browse by criteria	2	3.93	1..2
play by criteria	4	7.85	1..2
search by genre	2	3.93	1
search by artist name			-
up to 100 artists	1	1.96	1..2
more then 100 artists	2	3.93	1..2
search by album name			-
up to 200 albums	1	1.96	1..2
more than 200 albums	2	3.93	1..2
search by song title			-
up to 250 songs	1	1.96	1..2
more than 2000 songs	2	3.93	1..2
search by partial names			-
words	1	1.96	2
spelled letters	1	1.96	2
ambiguous entries	2	3.93	2
query			-
item counts	0.5	0.98	1
favorites			-
browse and play	0.5	0.98	1..2
add items	0.3	0.59	1
media management			-
refresh from media	0.2	0.39	1
add or remove media	0.2	0.39	1..2
access online content	1	1.96	2..3
Menu	0.4	3.10	-
quit	0.5	1.03	1..2
switch among other apps	1	2.07	1..2
Sum	44.2	100	-

$$\begin{aligned} PTC(t) = \quad & TC(t) + \lambda_{hr} hr(t) + \lambda_{rj} rj(t) \\ & + \lambda_{srt} srt(t) \end{aligned} \qquad (2)$$

where TC is the actual dialog turn count, hr is the number of help requests, rj is the number of rejections, and srt is system response time and the coefficients represent weights of each contributor to the final penalty turn count (PTC)[1]. TC, hr, and rj are averaged over the number of trials. By trial we mean each attempt of the user to perform a specific task. The system response time (srt)

[1] In our experiments, we set $\lambda_{hr} = 0.5$, $\lambda_{rj} = 1$, and $\lambda_{srt} = 0.3$.

is the average of system reaction times (in seconds) exceeding a constant c_{asrt} over the number of turns in trials (t_i). Acceptable systems reaction time constant (c_{asrt}) is set to 0.1, i.e. the acceptable threshold is 100 ms.

$$srt(t) = \frac{\sum\limits_{\text{all turns } t_i \text{ for task } t} \max(st(t_i) - c_{asrt}, 0)}{|t|} \tag{3}$$

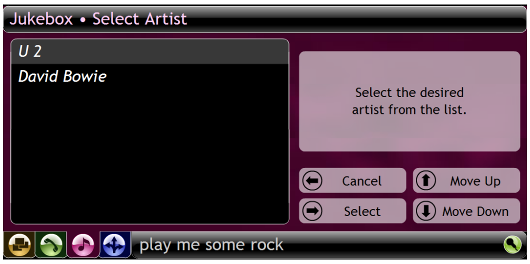
The obtained penalty turn count is then compared to an ideal number of turns for a particular task. The ideal turn count $ITC(t)$ for task t is the number of dialog turns needed to accomplish the task using an ideally efficient dialog system by a native user acquainted with the system.

Currently we determine $ITC(t)$ manually by human judgment. The $ITC(t)$ typically corresponds to the number of *coherent information blocks* that can be identified in the information that needs to be communicated by the user. For example, suppose a "date" value consisting of three information slots (day, month and year) needs to be entered. All slots however comprise a single coherent block of information that is typically communicated at once and thus we would set $ITC(t) = 1$ for this task. Table 2 shows a task in which the user selects a song whose title is ambiguous. The ideal system is expected to disambiguate in one extra turn and therefore we set $ITC(t) = 2$.

The actual score of the dialog efficiency (DE score) for an individual task is then counted as a fraction of the difference between ITC and PTC against current PTC, i.e.:

$$DE(t) = 1 - \max\left(\frac{PTC(t) - ITC(t)}{PTC(t)}, 0\right) \tag{4}$$

To avoid subjective scoring we typically use several human testers as well as several trials per one task. For example for the task "play by artist" the following set of trials can be used: "Play something by Patsy Cline", "Play some song from your favorite interpreter", or "Play some rock album, make the final selection by the artist name". Each of these trials is assigned its ideal number of turns (this is why ITCs for tasks in the ontology are given by ranges in Table 1.) The task dialog efficiency score is then computed as an average over all human testers and dialog efficiency scores for all their trials.


Jukebox • Select Artist

U 2
David Bowie

Select the desired artist from the list.

Figure 1: GUI of Jukebox application

Samples of trials used in the evaluation of the music management domain are given in Table 2. Figures of ITC and average turn count in this table are further discussed in Section 3.

The final dialog score is then counted as a sum of products of domain coverage and dialog efficiency for each task in the domain ontology, i.e.:

$$DS(S, O) = \frac{\sum_{t \in tasks(S,O)} w_t \, DE(t)}{\sum_{t \in all\_tasks(O)} w_t} \tag{5}$$

3 Example of Dialog Scoring on Music Management Domain

We applied the dialog scoring to our two dialog systems developed at different times and both partially covering the music management dialog domain. Both allow their users to play music by dynamically generating grammars based on meta tags found in users' mp3 files. The first one, named A-player, is simpler and covers a limited part of the music management domain. The second, named Jukebox, covers a larger part of the domain and also allows free-form input using a combination of statistical language models and maximum entropy based action classifiers. Figure 1 shows the GUI of the Jukebox application.

For both applications, we collected input from a group of 15 speakers who were asked to accomplish tasks listed in Table 2. Each of these user tasks corresponded to a task in the domain task ontology and there was at least one user task per each ontology task that was supported by either A-player or Jukebox. The subjects were given general guidance but no sample English phrases were suggested to them that could be used to control the system. In order not to guide users even by the wording of the user tasks, the tasks were described to them in their native language. All subjects were non-native but fluent English speakers.

Table 2: Specific tasks to be accomplished by personas using A-player and Jukebox with ideal number of turns (ITC) and average turn count (TC). Tasks which appeared to be more hard than expected are indicated in bold, easier than expected are in italic.

		Aplayer		Jukebox	
Task	ITC	TC	TC/ITC	TC	TC/ITC
Start playback of arbitrary music	1	1.5	1.5	3.1	**3.1**
Increase the volume	1	-	-	1.4	1.4
Set volume to level 10	1	-	-	1.4	1.4
Mute on	1	-	-	1.2	1.2
Mute off	1	-	-	1.5	1.5
Pause	1	-	-	2.1	**2.1**
Resume	1	-	-	2.5	**2.5**
Next track	1	1.4	1.4	1.1	1.1
Previous track	1	1.5	1.5	1.3	1.3
Shuffle	1	1.0	1.0	1.3	1.3
Play some jazz song	1	-	-	1.4	1.4
Play a song from Patsy Cline	1	1.5	1.5	2.0	**2.0**
Play Iron Man from Black Sabbath	1	1.9	**1.9**	2.8	**2.8**
Play the album The Best of Beethoven	1	1.1	1.1	1.7	**1.7**
Play song Where the Streets Have No Name	1	1.4	1.4	1.3	1.3
Play song Sonata no. 11 (ambiguous)	2	1.1	*0.6*	3.7	**1.8**
Play a rock song by your favorite artist	3	2.6	*0.9*	4.4	1.5
Reload songs from media	1	1.5	1.5	-	-

3.1 Domain Coverage for Music Management Domain

This restricted ontology represents the human expert knowledge of the domain and is encoded as a set of tasks with two kinds of relations between the tasks: task generalization and aggregation. Individual tasks are defined as sequences of parametrized actions. Actions are separable units of domain functionality, such as volume control, song browsing or playback.

Parameters are categories of named entities, such as album or track title, artist name or genre. Tasks are labeled by weights, which express the relative importance of a particular task with respect to other tasks. The ontology may also define task aggregations which explicitly state that a complex task can be realized by sequencing several simpler tasks. Table 1 shows a sample task ontology for the music control domain. For example, the task volume control/relative with weight of 2 (e.g. “louder, please”) is considered more important in evaluation than its absolute sibling (e.g. “set volume to 5”). This may be highly subjective if scored by a single human judge and thus a consensus of domain experts may be required to converge to a generally acceptable ontology for the domain. Once acknowledged by the community, this ontology could be used as the common etalon for scoring third-party dialog systems.

Table 3: Computation of domain coverage, dialog efficiency and dialog score for A-player

Task	DC	DE	final DS
play	7.75	0.67	0.052
stop	7.75	1.00	0.078
next, prev. track	2.58	0.73	0.019
play by criteria	7.85	0.71	0.055
search by artist			
$\leq$ 100 artists	1.96	0.60	0.012
$>$ 100 artists	3.93	0.60	0.024
search by album			
$\leq$ 200 albums	1.96	0.89	0.017
$>$ 200 albums	3.93	0.89	0.035
search by song			
$\leq$ 250 songs	1.96	0.86	0.017
$>$ 2000 songs	3.93	0.86	0.04
media refresh	0.39	0.67	0.003
Total (in %)	47.92	71.14	36.11

3.2 Computing Dialog Scores for Music Management Domain

Tables 3 and 4 show the computation of the final dialog system score (DS) and its components: domain coverage (DC) and domain efficiency (DE). For A-player, which is limited in functionality, the weighted domain coverage reached only 47.92%, whereas for Jukebox it was 83.17%. On the other hand, A-player allowed its users to accomplish the tasks it supported faster than Jukebox; this is documented by the weighted dialog efficiency score reaching 71.14% for A-player and 64.62% for Jukebox. This was mainly due to Jukebox being more interactive (e.g. asking questions, presenting choices) and due to a slightly higher error

Table 4: Computation of domain coverage, dialog efficiency and dialog score for Jukebox

Task	DC	DE	final DS
volume relative	6.20	0.74	0.046
volume absolute	3.10	0.74	0.023
mute	6.20	0.82	0.051
play	7.75	0.33	0.025
stop	7.75	0.82	0.064
pause	3.88	0.48	0.019
resume	3.88	0.41	0.016
next, prev. track	2.58	0.93	0.024
next, prev. album	2.58	0.76	0.020
shuffle	1.94	0.76	0.015
browse by criteria	1.97	0.53	0.010
play by criteria	7.85	0.68	0.054
search by genre	3.93	0.74	0.029
search by artist			
$\leq$ 100 artists	1.96	0.50	0.010
$>$ 100 artists	3.93	0.60	0.024
search by album			
$\leq$ 200 albums	1.96	0.35	0.007
$>$ 200 albums	3.93	0.75	0.029
search by song			
$\leq$ 250 songs	1.96	0.65	0.013
$>$ 2000 songs	3.93	0.93	0.036
word part. search	1.96	0.51	0.010
ambiguous entries	3.93	0.54	0.021
Total (in %)	83.17	64.62	54.45

rate of a free-form system (language model-based) as opposed to a grammar-based one. The overall dialog score was higher for Jukebox (54.45%) than it was for A-player (36.11%). This was in accord with the feedback we received from users, who claimed they had better experience with the Jukebox application, see Section 4.

4 Towards Correlation between Proposed Metrics and Subjective Evaluation

The HCI methodology (Nielsen, 1994) advocates several factors that human judges collect in the process of dialog system evaluation. These key indicators include accuracy, intuitiveness, reaction time, and efficiency. When designing the evaluation method we attempted to incorporate the core of these indicators into the scoring method to ensure good correlation of the proposed metric with human judgment.

After performing the case study for DE scoring, we asked the evaluators to fill in a questionnaire with their subjective feedback. There were three sets of questions: (1) speech suitability, (2) application-specific evaluation, and (3) question about location where they would be willing to use such applications.

The human evaluators were asked to rate each question (listed in Table 5), for both applications, with a score of 0 points (worst) to 5 points (best). The meaning of the points is shown below:

0 . . . worst, the system is not usable at all by anyone
1 . . . not sufficient for real usage, only good as a toy
2 . . . reasonable, but I would not consider using it
3 . . . reasonable, I would consider using it
4 . . . good understanding and behavior, I would use it
5 . . . excellent understanding and behavior

Generally, the evaluators were pretty positive in scoring speech suitability for music management domain in Question 1. In the application evaluation group of questions, the more advanced Jukebox application was perceived better (63.2% vs. 50.7% for A-player). Support of free-form commands by the Jukebox application and its broader functionality was reflected in Jukebox's score of 72.9% for Question 4 (vs. 54.3% for A-player) and influenced also answers to Questions 2 and 3. A-player's slightly higher score for Question 5 (65.7% vs. 62.9% for Jukebox) corresponds to the fact that the restricted set of commands and functionality makes the speech recognition task easier and therefore the users feel the system obeys their commands better. Results for the last two questions about location, where the evaluator would be willing to use the voice driven system, are less positive for home usage (54.3% and 57.1%) but the evaluators foresee an added value in using speech modality in environments when other input devices (such as keyboard, buttons, or touch screens) can be disturbing, i.e. in cars.

Statistically speaking, the average correlation between the vector of dialog scores, assembled for each individual speaker, and the vector of averaged points received from his/her subjective evaluation, was 0.67.

5 Conclusion

The objective of our approach is to evaluate spoken and multi-modal dialog systems within a predefined, well-known (and typically narrow) domain. In our labs we have used heterogeneous technologies such as grammars, language models and natural language understanding techniques to develop many speech and multimodal applications for various domains, such as music selection, TV remote control, in-car navigation and phone control. In order to compare two spoken dialog systems that deal with the same domain, we first describe the domain using a task ontology which defines user tasks relevant for the chosen domain as

Table 5: Questionnaire filled by the human evaluators after the test. The figures are given in percentage of “satisfaction” calculated from averaged points (between 0 and 5) given by the human evaluators.

Question	Aplayer	Jukebox
A. Speech suitability		
1. Do you think the concept of voice control makes sense for the jukebox domain?	71.4	
B. Application evaluation		
2. Would you use the system?	37.1	**55.7**
3. Do you think someone else could use the system?	45.7	**61.4**
4. Did you know what to say at each point of interaction?	54.3	**72.9**
5. Did the system obey your commands?	**65.7**	62.9
Application evaluation results (questions 2-5 averaged)	*50.7*	*63.2*
C. Where to use the application		
6. Would you use the system at home?	54.3	**57.1**
7. Would you use the system in car?	62.9	**71.4**

well as their relative importance. This enables us to compare two dialog systems against each other (1) by comparing their coverage of the ontology tasks, and (2) by contrasting their dialog efficiency over the supported tasks. A single dialog score statistic can be produced by combining the dialog coverage and dialog efficiency components.

The presented approach is suitable for comparing different dialog systems of third parties as well as successive versions of a single system being developed. Human evaluations are currently conducted to estimate the correlation between the dialog score and human judgment. The subjectivity of human scoring and consensus on the ontology coverage are subject of further investigation.

Acknowledgments

We thank to the Jukebox application creators, esp. to Luboš Ureš, Ladislav Seredi and Mark Epstein. We would like to acknowledge support of this work by the European Commission under IST FP6 integrated project Netcarity, contract number IST-2006-045508.

References

Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Swift, M., Taysom, W. 2007. PLOW: A Collaborative Task Learning Agent. *Twenty-Second Conference on Artificial Intelligence (AAAI-07).*

Le Bigot, L., Bretier, P., Terrier, P. 2008. Detecting and exploiting user familiarity in natural language human-computer dialogue. *Human Computer Interaction: New Developments. Kikuo Asai (Eds), InTech Education and Publishing*, ISBN: 978-953-7619-14-5, 269-382.

Carroll, J. 2001. Human Computer Interaction in the New Millennium. *New York: ACM Press.*

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K. 2002. Mack: Media lab autonomous conversational kiosk. *Imagina02.*

Gandhe, S., Traum, D. 2008. Evaluation understudy for dialogue coherence models. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, Ohio, Association for Computational Linguistics*, 172-181.

Graesser, A.C., VanLehn, K., Rosfie, C.P., Jordan, P.W., Harter, D. 2001. Intelligent tutoring systems with conversational dialogue. *AI Mag. 22(4)*, 39-51.

Hajdinjak, M., Mihelific, F. 2006. The paradise evaluation framework: Issues and findings. *Comput. Linguist. 32(2)*, 263-272.

Jurafsky, D., Martin, J.H. 2008. Speech and Language Processing (2nd Edition): An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. *(Prentice Hall Series in Artificial Intelligence)*

Nielsen, J. 1994. Heuristic evaluation. *Usability Inspection Methods, J. Nielsen and R.L. Mack, R.L. (Eds), John Wiley and Sons: New York*, ISBN: 0-471-01877-5, 25-64.

Rieser, V., Lemon, O. 2008. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz data: Bootstrapping and Evaluation. *Proceedings of ACL-08: HLT*, pages 638-646, Columbus, Ohio, USA, June 2008.

Singh, S., Kearns, M. S., Litman, D. J., Walker, M. A. 1999. Reinforcement learning for spoken dialogue systems. *In Proc. NIPS99.*

Walker, M., Kamm, C., Litman, D. 2000. Towards developing general models of usability with paradise. *Nat. Lang. Eng. 6(3-4)*, 363-377.

Weizenbaum, J. 1972. ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the Association for Computing Machinery 9*, 36-45.

Agency & Information State in Situated Dialogues: Analysis & Computational Modelling

Robert J. Ross
SFB/TR8 Spatial Cognition
Universität Bremen, Germany
robertr@informatik.uni-bremen.de

John Bateman
SFB/TR8 Spatial Cognition
Universität Bremen, Germany
bateman@uni-bremen.de

Abstract

Spatially situated applications present notable challenges and unique opportunities for the dialogue modelling community. In light of this, we report on our experiences developing information-state dialogue management models for the situated domain, and present a dialogue management model that fuses information-state update theory with a light-weight rational agency model. We describe the model, report on its implementation, and comment on its application in concrete spatial language processing applications.

1 Introduction

Our work is concerned with the development of language and dialogue processing for the class of situated systems. Examples of situated systems include in-vehicle information technologies, spatially aware assistance applications, and cognitive robots. In all of these situated applications, user-system interaction through standard graphical, textual, or tactile modes of communication is either insufficient or simply not feasible for various reasons. As such, the language interface presents a highly appealing interaction mode for such applications.

Situated systems do however present notable research challenges for the dialogue community. While one noteworthy issue concerns the context-sensitive interpretation and production of spatial language that is seen frequently in the situated domain (Ross, Forthcoming), a second issue, and one which we directly address in this paper, is the *agentive* nature of situated applications. Specifically, situated applications have complex internal mental states, operate in a semi-autonomous manner, and perform actions that have clear temporal extent. Such agency features minimally require mixed-initiative and multi-threading in dialogues, but also a coupling of dialogue management with rational agency that recognizes the disparate, yet tightly coupled, nature of these elements.

We see the Information State Update (ISU) theory of dialogue management (Traum and Larsson, 2003) as being well placed to provide a basis for situated dialogue. Specifically, due to a shared lineage, ISU is a natural bridge between dialogue processes and the models of rational agency that continue to be applied within current cognitive robotics and situated systems models. But, arguably more importantly, it has now been well shown that the ISU approach is highly suited to the production of mixed-initiative and multi-threaded dialogue (Lemon et al., 2002; Larsson, 2002).

The class of classical ISU models, and in particular their realization through toolkits like TrindiKit (Traum and Larsson, 2003) and DIPPER (Bos et al., 2003) do however present some challenges when applied in the situated domain. One issue concerns the relationship between dialogue policy and the contextualization of user contributions. Within many classical ISU-based models, dialogue plans are first processed to collect mandated frame information from a user before this information is sent to a domain model for contextualization, update or query. This *collect-then-contextualize* policy favours explicit constraint gathering for complex frames, but can, if applied directly in the situated domain, lead to unnecessary clarifications and hence unnatural dialogue. To illustrate, consider the application of a *collect-then-contextualize* policy to a simple command-oriented dialogue in which the user of a robotic wheelchair attempts to direct the system to turn when the situational context makes the direction of turning clear:

(1) a. *User:* turn here
left direction is only obvious direction

b. *System:* should I turn left or right?
c. *User:* left

In such a case the clarification dialogue is superfluous and can be avoided through immediate contextualization of user contributions prior to dialogue planning policy invocation.

More significantly, due to an intended flexibility, the relationship between dialogue plans and the operations of mental state update applied within intentional systems is highly underspecified. Namely, following dialogue plan completion, domain model update information is typically flattened into a proposition set which has no epistemological form or persistence in of itself, and which must be interpreted by the domain application in an unspecified manner (Larsson, 2002). We, on the other hand, argue that scalable intelligent systems require more transparent links between the constructs of information state and the units of epistemological and intentional state.

In light of such issues, in the remainder of this paper we introduce a dialogue management model that we have developed for use in mobile robot applications. This *Agent-Oriented Dialogue Management* (AODM) model is cast within ISU theory, but (a) establishes a link between models of rationality and classical information state; and (b) applies an explicit function-based model of domain contextualization. We proceed by introducing the model's main components, followed by a description of the assumed dialogue processes, and, finally, an overview of the dialogue model's realization and application.

2 The AODM Model Components

While rejecting intractable, monolithic agent-based dialogue management models, we argue that the properties of the situated domain necessitate the inclusion of the intelligent agent metaphor in domain modelling. Thus, we apply agency models to domain organization, but capture dialogue management as meta-behaviours which operate over these cognitive constructs. In particular, we draw on techniques from the so-called *agent-oriented programing language* community (Shoham, 1993). While agent-oriented frameworks provide very rich rational agency models, here we limit ourselves to only their most salient aspects that necessarily interact with dialogue modelling and management constructs.

Taking an agent-oriented view of a domain application suggests the use of speech-act wrapped domain action and state definitions as the natural units of communication between system and user. Such a construct is essentially equivalent to a speech act in artificial agent communication languages, e.g., (FIPA, 1998). However, in natural communication, such a *dialogue move* is the result of a complex grounding process rather than a direct product of perception. Thus, following the approaches to dialogue structure originally proposed by Butler (1985) and later Poesio and Traum (1998), we assume the *dialogue act* as the primary unit of exchange at the surface/semantics interface, while assuming the *dialogue move* as the coarse grained unit of interaction established through grounding and contextualization at the semantics/pragmatics interface.

As we will see below, the *move* in classical ISU terminology corresponds more closely to our notion of dialogue act rather than dialogue move. While clearly in conflict with classical ISU terminology, our use of these terms is intended to capture two distinct levels of communicative action with meaningful terms. Moreover, this usage is derived from earlier models of Exchange Structure description used in the discourse analysis community (Berry, 1981).

In the following we flesh out these principles by detailing, first, the assumed agent components, and then the dialogue components and information state model.

2.1 Agentive Components

The main non-dialogic mental state modelling types assumed by the AODM model are briefly summarised below.

2.1.1 Capabilities

The AODM model assumes a domain agent to be endowed with one or more action definitions and zero or more plan definitions. We use the term *Capability* to generalize over actions and plans, and thus assume the agent to have a Capability Library that defines an inventory of available plans and actions. It should be noted that plan bodies can be composed dynamically outside the scope of named plan types, thus allowing a user to conjoin action and plan types arbitrarily.

We define the signatures of all capabilities, i.e., actions and plans, to have certain shared properties. First, we assume all capabilities to be performed by an agent - in our case either the dialogue agent itself or the user. Second, we assume that all capabilities have a certain earliest time at which a parametrised capability may be invoked. We may express these constructs from an ontological perspective, and assume these units to be defined in terms of the agent's conceptual ontology. Individual domains extend such signature properties into a capability hierarchy.

2.1.2 Intentions

An intention can be defined in the usual way in terms of the capability to be performed, when it is to be performed, whether there are any child or parent intentions, the state of the intention, and so forth. The intention-like corollary of a plan is an intention structure, and the agent can at any time have any number of planned or active intentions – which may be either single intentions or more complex intention structures.

The use of intentions and intention structures is of course common in both formal pragmatics and in agent-oriented applications, but for the language processing domain we minimally extend the notion of the agent's intention structure with an *Intention Salience List (ISL)*. The ISL is a stack of atomic intentions used to explicitly track the most prominent intentions within the agent's mental state. We define an atomic intention to be most salient based on recent state transitions of that intention. The ISL facilitates process resolution as required for interpreting highly elliptical process resolving commands such as "stop".

2.1.3 Beliefs & Domain State

In line with the prevalent view in the dialogue management community, we assume the details of belief state organization to be highly domain dependent. Thus, the AODM model requires only an abstract query interface over the agent's belief state. Moreover, due to the highly complex and detailed nature of spatial state, we eschew the existence of simplistic *addBelief* and similar mental state manipulation primitives in favour of specific capabilities for addressing task-specific user questions or additions of information by a user. We do however assume that unlike physical capabilities, such *cognitive* capabilities are effectively instantaneous from a user's perspective.

2.2 Dialogue Components

The AODM model also assumes a number of core dialogue components.

2.2.1 Dialogue Acts

The Dialogue Act (DA) is a conceptual-level description of a dialogue contribution made by an interlocutor. The dialogue act thus captures the semantics of individual utterances, and reflects a traditional pragmatic view of communicative function. The dialogue act may thus be informally defined as an entity which: (a) is performed by some agent; (b) potentially takes a propositional content defined in terms of the agent's domain ontology; (c) is performed at a particular time; and (d) has an associated speech function type.

2.2.2 Dialogue Moves

The Dialogue Move (DM) on the other hand is a frame-like construct that acts as the main interface between dialogue management and rational agency processes. The dialogue move is thus a more complex construct than a dialogue act – although one-to-one correspondences between dialogue acts and dialogue moves may also occur. The use of a dialogue move rather than a more complete dialogue frame was motivated by the necessity of taking an agent-oriented perspective on dialogue processing, yet building on the frame metaphor as a staging ground for meaningful unit composition.

The licensed content of a DM is directly coupled to the agent's range of capabilities and potential mental states. More specifically, user DMs and the intentions an agent may adopt are coupled in the usual way in terms of classical illocutionary logic rules which dictate that if the system is requested to perform some capability, and the system can perform that capability in the current state, then the system should adopt the intention to perform that capability.

Due to the DM's role as a construct that sits between the language interface and the agent's intentional state, we model the DM as a dynamic frame-like structure with three components:

- **The Move Template:** defines the DM type and content potential in terms of concept and role definitions extracted from the agent's conceptual ontology.
- **The Move Filler:** is the set of shallow descriptions provided by the user to fill out the

Role	Type	Filler	Solution A	Solution B
actor	Agent	nil	1.0, system	1.0, system
placement	Place	nil	1.0, here	1.0, here
earliestTime	Time	nil	1.0, now	1.0, now
direction	GenDir	GenDir modality Left	0.5, GenDir modality Left_{Ego} extent 90	0.5, GenDir modality $\text{Left}_{\text{Allo}}$ extent 90
speed	Speed	nil	1.0, normalSpeed	1.0, normalSpeed

Table 1: Instance of an `Instruct-Reorient` dialogue move for the interpretation of "turn left" in a context where both an allocentric or egocentric interpretation of "left" are possible. The first two columns define the parameter types applicable to the dialogue move in terms of concept and role restrictions. The filler column denotes the unresolved content derived from the instantiating dialogue act. The final two columns show contextualization solutions denoting alternative but equally likely interpretations of the move are denoted.

roles in the move template.

- **The Solution Set:** is the set of possible interpretations of the move filler following contextualization. While solution contents are defined in terms of the agent's application ontology, solution contents also have associated interpretation likelihoods, and typically includes content which was not directly provided by the speaker.

For illustration, Table 1 depicts a move instance which includes the Move Template, Move Filler, and Solution Set information for an instruction to make a turning, or *Reorientation*. It should be noted that for this example, the speaker provided only direction information, and that all other parameters in the presented solutions were filled through contextualization.

Though somewhat similar in nature, there are a number of notable distinctions between the DA and the DM. Unlike DAs, which can be instantiated for a broad number of speech function types, DMs may only be instantiated for task-relevant speech function types. This distinction is due to the level of non-task exchange elements being handled by the dialogue management processes without any need for explicit domain contextualization. Also, although the contents of both DAs and DMs are defined in terms of the agent's conceptual ontology, the content of a DA can be any consistent selection from this ontology, whereas the content of a DM must be headed by an application state or capability. Thus, a DM is assumed to constitute a 'meaningful' update of the agent's state rather than a fragmentary piece of information. It is then the responsibility of the dialogue process as a whole to make the mapping from fragmentary acts to complete moves.

The AODM model also applies the DM to the modelling of system initiated dialogue goals – albeit with some differences to account for the initial certainty in system rather than user dialogue moves. Essentially, unlike user dialogue moves, system dialogue moves only have a single contextualized interpretation as there is no ambiguity in system generated content.

2.2.3 Complex Components

Just as actions can be complexed into plans, and intentions into intention structures, we assume both DAs and DMs can be complexed together via semantic relations. Such modelling is necessary to capture the conjunction, disjunction, or sequencing of instructions and statements as seen frequently in situated task-oriented dialogue. We thus introduce the notion of both *Dialogue Act Complexes* and *Dialogue Move Complexes* as reified constructions of individual dialogue acts and moves. However, for the remainder of this paper we generalize the two complex sorts to their atomic constituents for the sake of brevity.

2.3 The Information State Structure

To conclude the discussion of the AODM's components, Table 2 depicts the AODM's Information State Structure. Most slot types are self explanatory, therefore we will not detail the contents of these slots here.

Slot	Type
Input Abstractions	
Latest-User-Utterance	{String,float}
Latest-User-Act	Act
User Act Containers	
Non-Integrated-User-Acts	Set(Act)
User Move Containers	
Open-User-Moves	Stack(Move)
Closed-User-Moves	Stack(Move)
System Moves Containers	
Planned-System-Moves	Stack(Move)
Raised-System-Moves	Stack(Move)
Closed-System-Moves	Stack(Move)
System Act Containers	
Planned-System-Acts	Set(Act)
Open-System-Acts	Set(Act)
Output Abstractions	
Next-System-Act	Act
Next-System-Contribution	String
Error types	
Input-Error	ErrorType

Table 2: The Information State Structure

3 Dialogue Process Models

The AODM process models and update approach follow broadly from an ISU perspective, but have been modified both due to the more action-oriented dialogues with which we deal, and to provide a more efficient implementation strategy. First, in light of the highly context-sensitive nature of situated language, we reject a strict *collect-then-contextualize* dialogue policy, and instead invoke a *contextualize-then-collect* perspective that makes use of an explicit contextualization process called immediately following the integration of user dialogue acts into the information state. This contextualization process aims to augment and resolve any resultant open user moves prior to dialogue planning. Second, to achieve a tighter coupling of dialogue and intentional behaviour, intention adoption and management strategies are integrated directly into the ISU process model. Specifically, the intention adoption strategy is integrated with dialogue planning in a single planning module, while an intention management process is invoked between response planning and the planning of concrete system messages.

Ignoring dialogue act recognition and language realization processes, the AODM control cycle can thus be summarized in terms of the following processes called in sequence:

- **Act Integration**
- **Move Contextualization**
- **Response Planning**
- **Intention Management**

Details of these processes, as well as the discourse model, are presented by Ross (Forthcoming). In the following we given a brief overview of these processes.

3.1 Act Integration

The language integration process is responsible for taking user speech acts (possibly complex) and integrating them into the information state. Successful integration of task-specific acts involves the update of open user or system dialogue moves, or the creation of new user dialogue moves. The integration process follows closely with the ISU methodology, and in particular with the general features of the model outlined by Larsson (2002) – including support for multi-threading in dialogue. As such we will not detail the model further here except to note that rather than assume a rule-based model of update, we apply in the AODM model, and in its implementation described later, a procedural approach to update specification. More specifically, while we acknowledge the importance of clear, strongly-typed, declarative models of information state, we argue that a procedural model of the update process, which is equivalent to a rule-based specification, provides a more transparent view on the update decision process, and is thus both easier to debug and extend. Moreover, we would argue that a procedural approach is in fact closer to the original view of update strategies in the *Questions-Under-Discussion* model as proposed by Ginzburg (1998).

3.2 Move Contextualization

Directly following integration, all open user moves are contextualized against the current situational model. Contextualization requires the resolution of anaphoric (in the general sense) references, elided content, and ambiguous features such as reference frame use. Due to domain complexity, we cannot view contextualization in the situated domain as simply partial unification of a dialogue move with a context model. Instead, we have developed a situated contextualization approach where functions, associated with individual semantic constituents, are used to compose concrete resolved meanings.

Rather than relying only on a set of resolution functions, i.e., functions dedicated for the contextualization of user specified content, our contextualization approach also relies on a second set of augmentation functions. Thus, each semantic role in a dialogue move type has both a resolution and augmentation function associated with it. Augmentation functions are applied in the case of a user completely omitting a move role, and typically apply default information based on situational norms – including the affordances offered by a physical context. For any given semantic role, the triggered augmentation or resolution function may produce multiple possible interpretations for that semantic role. These multiple interpretations thus result in the addition of possible solutions to a move specification as was described in Section 2.2.2. The solution set associated with a given move can both decrease and increase in size over the course of the contextualization process, and if, at the end of contextualization, more than one solution is available, the reduction of the solution set becomes the responsibility of response planning.

3.3 Response Planning

Following contextualization, the response planning process is triggered to review the information state and determine what new actions, if any, should be performed. In order to maintain synchronization between backward looking system dialogue moves and adopted intentions, the dialogue planning process is tightly coupled to the agent's intention adoption strategy. Moreover, as with language integration, we have developed our dialogue planning processes in a procedural rather than update-rule based methodology to provide greater transparency in design.

The response planning process is designed in multiple stages. The first level of response planning include the determination of what intentions and dialogue goals – if any – should be adopted. Intentions to perform requested capabilities are adopted if a requesting open user move has a single associated complete solution. Based on whether an intention is to be adopted or not, the system may also adopt an explicit dialogue move goal to signal the acceptance or rejection of particular user requests. The second level of response planning involves a lower-level choice of which dialogue acts should be assembled to pursue either new system goals or open user moves.

3.4 Intention Management

Although not a linguistic process, the AODM model also directly includes an intention management process that is responsible for sequencing adopted intentions. The justification for directly including what is usually considered a domain specific process is to ensure that sufficiently developed models of intention management, which includes the notion of the Intention Salience List as introduced earlier, are available to specific applications.

3.4.1 Illustration

To illustrate the properties of the AODM model, and in particular the relationship between units of mental state, Figure 1 depicts a dialogue example along with a partial discourse structure typical of the dialogue types that the AODM model has been designed to handle. Note that this exchange consists of two moves – the first move being a user move which requests a concrete action, and the second move being the system's response to the user move.

4 Realizing AODM with Daisie

The AODM model has been implemented within an information-state update based dialogue framework which grew out of our earlier attempts at dialogue system construction based directly on agent-oriented programming solutions. The dialogue framework, named Daisie (**D**iaspace's **A**daptive **I**nformation **S**tate **I**nteraction **E**xecutive), is a dialogue systems framework, written in Java, which provides a tightly coupled dialogue systems integration approach based on the use of a plugin architecture. An important part of our motivation in developing the system was to support a more rigorous approach to ontology definition and modularization within the description of linguistic resources and mental state. As such, the content of individual Information State slots is captured in terms of a Description Logic based representation and reasoning system.

Following earlier experiments with highly decentralized, middleware based, integration solutions, we have opted instead for a far more tightly coupled integration strategy. We argue that the future of spoken dialogue systems shall head towards ever more tight integration between ele-

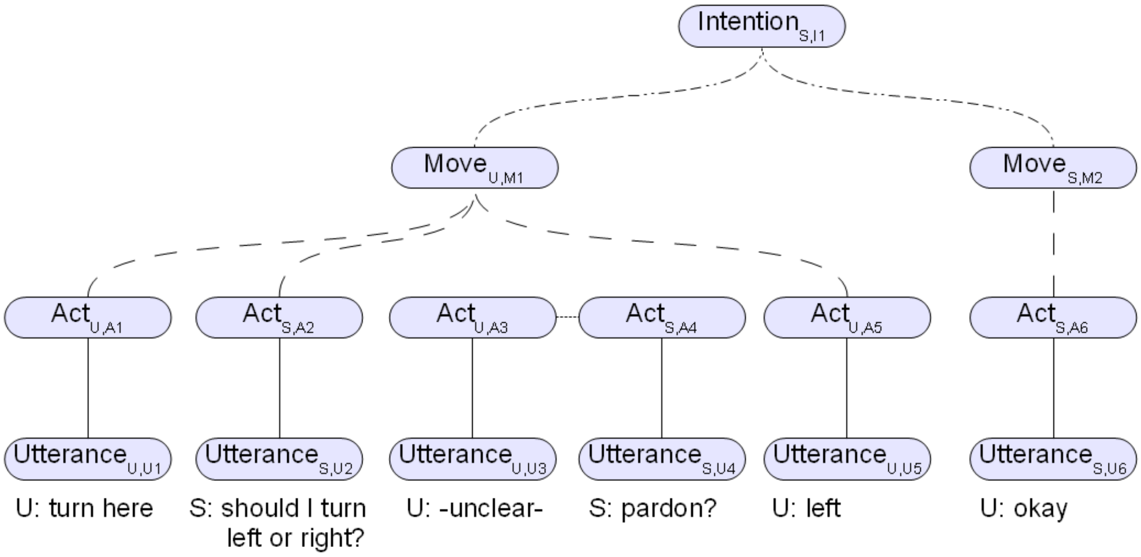


Figure 1: Inter-stratal relationships in the AODM model. Full lines denote correspondence relationships, coarse dashed lines express a constituency relationship between dialogue moves and dialogue acts, and fine dashed lines express a loose causal relationship. For each unit instance, the first subscript indicates ownership, i.e., u=user, s=system. The second subscript in turn indicates the unit instance name, e.g., u2=Utterance2.

ments where contextual information is applied at increasingly early stages to resolve ambiguity in input. While constant communication between components could be achieved through a distributed architecture, we argue that a tighter coupling between components both improves efficiency at runtime, and also improves the development process since programming interface based design rather than composing and interpreting messages is in practice easier to implement. Moreover, we argue that although a multi-agent based approach to software integration is very useful in the case of dynamic systems, typical spoken dialogue systems are very static in component design, and thus little is actually gained from a fully distributed architecture.

Ross (2008) reports on the application of an early version of the AODM model and Daisie framework to the dialogic interpretation of spatial route instructions. In this *Navspace* application, a user plays the role of a Route Giver in directing a mobile robot around a simulated office environment. The example given in Figure 1 is typical of the dialogues handled by this application. User study based evaluation of this application demonstrated that the AODM model - and in particular the contextualization process applied - led to an 86% task completion rate over 58 experimental trails conducted by 6 participants (Ross, 2008). However, this task completion rate belies the fact that participants invariable moved towards communicating their intents through very simplistic language. Integrating the AODM model with strategies that provide better context-sensitive feedback to users is thus a focus of current work.

5 Relation to Other Work

From a core modelling perspective, our treatment of dialogue moves and dialogue acts as the central representation units in a discourse representation can be considered a partial realization of Poesio & Traum Theory (PTT) (Poesio and Traum, 1998). However, whereas PTT focused on the basic tenets of the grounding process, the AODM model has been developed to explore the relationship between dialogue processes, agency and contextualization. Our consideration of the grounding process, the information state, and the problems of situated contextualization also distance the AODM model both from classical agent-based dialogue management models and also *neo-* agent-based dialogue management models such as Sadek et al. (1997)'s ARTIMIS system, or Egges et al. (2001)'s BDP dialogue agents. Within the ISU school, the AODM approach and its implementation is probably closest to Gruenstein and Lemon's Conversational Intelligence Architecture (Gruenstein, 2002; Lemon and Gruenstein, 2004). Specifically, both models advocate a tight coupling between dialogue management and agency features – although in our work we have attempted to push towards issues of representation and function-based language resolution and augmentation in an ontologically modular architecture. Finally, the function-based approach to contextualization shares motivations with recent work by Tellex and Roy (2007) in the interpretation of spatial language. However, whereas Tellex & Roy focused on the resolution of explicit language in a monologue setting, we have applied a function-based strategy to both resolution *and augmentation* in a full dialogue setting.

6 Future Work & Conclusions

We have developed and applied the AODM model to investigate the relationship between models of discourse, physical context and agency models. As such, the dialogue management model has necessarily focused on the handling of simple action-oriented dialogues. Thus, interactions typical of more complex frame structures such as booking flights cannot be handled by the current model. Instead we see frame-filling as a higher order dialogue process which operates directly on ground moves rather than un-contextualized dialogue acts. Amongst other issues, in future work we hope to investigate these relationships, and develop a frame-filling process which effectively sits above the AODM model.

Acknowledgements

We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition – Subproject I5-[DiaSpace].

References

Margaret Berry. 1981. Towards layers of exchange structure for directive exchanges. *Network: news, views and reviews in systemic lingustics and related areas*, 2.

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*.

Christopher S. Butler. 1985. Discourse systems and structures and their place within an overall systemic model. In James D. Benson and William S. Greaves, editors, *Systemic perspectives on discourse: selected theoretical papers from the 9th International Systemic Workshop*. Ablex, Norwood, NJ.

A. Egges, A. Nijholt, and R. op den Akker. 2001. Dialogs with BDP agents in virtual environments. In *Knowledge and Reasoning in Practical Dialogue Systems. Working Notes, IJCAI-2001*.

FIPA. 1998. FIPA 97 specification part 2. Technical report, Foundation for Intelligent Physical Agents, June 24.

Jonathan Ginzburg. 1998. Clarrifying utterances. In J. Hulstijn and A. Niholt, editors, *In Proceedings of the Twente Workshop on the Formal Semantics and Prgmatics of Dialogue*, pages 11–30.

Alexander Gruenstein. 2002. Conversational Interfaces: A Domain-Independent Architecture for Task-Oriented Dialogues. Master's thesis, Stanford University.

Staffan Larsson. 2002. *Issue-Based Dialogue Management*. Ph.d. dissertation, Department of Linguistics, Göteborg University, Göteborg.

Oliver Lemon and Alexander Gruenstein. 2004. Multithreaded Context for Robust Conversational Interfaces: Context-sensitive speech recognition and interpretation of corrective fragments. *ACM Transactions on Computer-Human Interaction*, 11(3):241–267, September.

Oliver Lemon, Alexander Gruenstein, and Stanley Peters. 2002. Collaborative Activities and Multitasking in Dialogue Systems. *Traitement Automatique des Langues (TAL)*, 43(2):131–154. Special issue on dialogue.

Massimo Poesio and David Traum. 1998. Towards an axiomatization of dialogue acts. In *Processings of TWENDIAL, the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*.

Robert J. Ross. 2008. Tiered models of spatial language interpretation. In *Proceedings of Spatial Cognition 08*, Freiburg, Germany.

Robert J. Ross. Forthcoming. Situated Dialogue Systems: Agency & Spatial Meaning in Task-Oriented Dialogue.

M. David Sadek, Philippe Bretier, and E. Panaget. 1997. ARTIMIS: Natural dialogue meets rational agency. In *IJCAI (2)*, pages 1030–1035.

Yoav Shoham. 1993. Agent Oriented Programming. *Artificial Intelligence*, 60:51–92.

Stefanie Tellex and Deb Roy. 2007. Grounding language in spatial routines. In *AAAI Spring Symposia on Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems*.

David Traum and Staffan Larsson. 2003. The Information State Approach to Dialogue Management. In Ronnie Smith and Jan van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer Academic Publishers, Dordrecht.

TRIK: A Talking and Drawing Robot for Children with Communication Disabilities

Peter Ljunglöf
DART: Centre for AAC and AT, Gothenburg, Sweden

Abstract

We will demonstrate a setup involving a communication board (for manual sign communication) and a drawing robot, which can communicate with each other via spoken language. The purpose is to help children with severe communication disabilities to learn language, language use and cooperation, in a playful and inspiring way. The communication board speaks and the robot is able to understand and talk back. This encourages the child to use the language and learn to cooperate to reach a common goal, which in this case is to get the robot to draw figures on a paper.

1 Introduction

1.1 Dialogue systems

Most existing dialogue systems are meant to be used by competent language users without physical or cognitive language disabilities – either they are supposed to be spoken to (e.g., phone based systems), or one has to be able to type the utterances (e.g., the interactive agents that can be found on the web). The few dialogue systems which are developed with disabled people in mind are targeted at persons with physical disabilities, who need help in performing common acts.

Dialogue systems have also been used for second language learning; i.e., learning a new language for already language competent people. However, we are not aware of any examples where a dialogue system has been used for improving first language learning.

1.2 Target audience

Our intended target group are children with severe communication disabilities, who needs help to learn and practice linguistic communication. One example can be children with autism spectrum disorders, having extensive difficulties with representational thinking and who therefore will have problems in learning linguistic communication. Our dialogue system will give an opportunity to explore spoken language – content as well as expression. Another target audience are children whose physical disabilities are very extensive, usually as a consequence of Cerebral Palsy (CP). The ablility to control a robot gives a fantastic opportunity to play, draw and express oneself in spoken language, which otherwise would be very difficult or even impossible.

1.3 Language development

To be able to learn a language one must have practice in using it, especially in interplay with other language competent people. For the communication to be as natural as possible, all participants should use the same language. For that reason there is a point in being able to express oneself in spoken language, even if one does not have the physical or cognitive ability. If one usually expresses oneself by pointing at a communication board, it is thus important that the board can express in words what is meant by the pointing act. This is even more important when learning a language, and its expressions and conventions (Sevcik and Romski, 2002; Thunberg, 2007).

When it comes to children with autism, learning appears to be simpler in cooperation with a technical product (e.g., a computer), since the interaction in that case is not as complex as with another human (Heimann and Tjus, 1997). Autistic persons have difficulties in coordinating impressions from several different senses and different focuses of attention. When one is expected to listen to, look at and interpret a number of small signals, all at the same time, such as facial expressions and gazes, human communication can become very difficult.

2 TRIK: A talking and drawing robot

Our basic idea is to use a dialogue system to support language development for children with severe communicative disabilities. There are already communication boards connected to speech synthesis in the form of communication software on computers. The main values that this project add to existing systems are that: *i)* the child can explore language on her own and in stimulating cooperation with the robot; *ii)* it can be relieving and stimulating at the same time, with a common focus on the dialogue together with a robot; and *iii)* the child is offered an exciting, creative and fun activity.

In our setup the child has a communication board which can talk; i.e., when the child points at some symbols they are translated to an utterance which the board expresses via speech synthesis, and in grammatically correct Swedish. This is recognized by a robot which can move around on a paper and draw at the same time. The robot executes the commands that was expressed by the communication board; e.g., if the child points at the symbol for *"draw a figure"*, and the symbol with a flower, the utterance might be *"draw a flower, please"*, which the robot then performs.

The dialogue system comes into play when the robot is given too little information. E.g., if the child only points at the symbol for *"draw a figure"*, the robot does not get enough information. This is noticed by the dialogue system and the robot asks a follow-up question, such as *"what figure do you want me to draw?"*.

2.1 Pedagogical advantages

By having the communication board and the robot talking to each other there is a possibility for users in an early stage of language development to understand and learn basic linguistic principles.

As discussed in section 2.3 later, the setup works without the robot and the communication board actually listening to each others' speech – instead, they communicate wirelessly. However, there is an important pedagogical point in having them (apparently) communicate using spoken language. It provides the child with an experience of participating in a spoken dialogue, even though the child does not speak.

2.2 The robot and the communication board

The robot itself is built using LEGO Mindstorms NXT, a kind of technical lego which can be controlled and programmed via a computer. Apart from being cheap, this technology makes it easy to build a prototype and to modify it during the course of the project.

The communication board is a computer touchscreen. The computer also controls the robot, both movements and speech. Every utterance by the robot will be executed by the speech synthesizer, and then sent to the robot via radio.

2.3 Perfect speech recognition

Typically, the most error-prone component of a spoken dialogue system is speech recognition; i.e., the component responsible for correctly interpreting speech. An advantage of the TRIK setup is that we will, in a sense, have "perfect speech recognition", since we are cheating a bit. The (dialogue system connected to the) robot does not actually have to listen for the speech generated by the (computer connected to the) communication board; the information is instead transferred wirelessly.

2.4 The dialogue system

The dialogue system is implemented using the GoDiS dialogue manager (Larsson, 2002), which is designed to be easily adaptable to new domains, but is nevertheless able to handle a variety of simpler or more complex dialogues.

The grammars of the dialogue system are implemented in Grammatical Framework (GF) (Ranta, 2004), which makes it easy to quickly design the language interpretation and generation components of a dialogue system.

3 Evaluation

During April–June 2009, the system is evaluated by a small number of users with linguistic communication disorders. The users are children with a diagnose within the autism spectrum, or with Cerebral Palsy. The evalation process is designed as a case study with data being collected before and after interventions. The children are also video recorded when playing with the robot, to enable analysis of common interaction patterns.

Both before and after the two month trial period, the parents answer a survey about how they perceive their interaction with their children. They also estimate the communicative abilities of their children. During the trial period, the children are filmed while interacting with the robot. Furthermore, all interaction between the communication board and the robot will be logged by the system. The logs and videos will be analysed after the trial period using suitable methods.

Acknowledgements

The TRIK project is financed by the Promobilia foundation and Magn. Bergvall's foundation.

References

Mikael Heimann and Tomas Tjus. 1997. *Datorer och barn med autism*. Natur och Kultur.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Department of Linguistics, University of Gothenburg.

Aarne Ranta. 2004. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189.

Rose Sevcik and Mary Ann Romski. 2002. The role of language comprehension in establishing early augmented conversations. In I. J. Reichle, D. Beukelman, and J. Light, editors, *Exemplary Practices for Beginning Communicators*, pages 453–475. Paul H. Brookes Publishing.

Gunilla Thunberg. 2007. *Using speech-generating devices at home: A study of children with autism spectrum disorders at different stages of communication development*. Ph.D. thesis, Department of Linguistics, University of Gothenburg.

Multimodal Menu-based Dialogue in Dico II

Staffan Larsson
Department of Linguistics
Göteborg University
Sweden
sl@ling.gu.se

Jessica Villing
Department of Linguistics
Göteborg University
Sweden
jessica@ling.gu.se

Abstract

We describe Dico II, a multimodal in-vehicle dialogue system implementing the concept of Multimodal Menu-based Dialogue. Dico II is based on the GoDiS dialogue system platform, enabling flexible dialogue interaction with menu-based in-vehicle applications.

1 Introduction

Dico II is a multimodal in-car dialogue system application. DICO (with capital letters) is a research project involving both industry and academia[1]. Dico II is built on top of the GoDiS dialogue system platform (Larsson, 2002), which in turn is implemented using TrindiKit (Traum and Larsson, 2003). In the original Dico application (Olsson and Villing, 2005), (Villing and Larsson, 2006), the dialogue system was able to control a cellphone. The main goal was to develop an interface that is less distracting the driver, and thus both safer and easier to use than existing interfaces. A subsequent version (Larsson and Villing, 2007) included also a Driver Information Display (DID) and a radio in order to o show how a multimodal dialogue system can help when controlling several devices. This paper describes the Dico II system resulting from work in the DICO project.

2 In-vehicle dialogue systems

An obvious advantage of spoken dialogue in the vehicle environment is thath the driver does not have to take her eyes - and the attention - off the road. In an in-vehicle environment, it is cruicial that the system is intuitive and easy to use in order to minimize the cognitive load imposed on the driver by the interaction. The GoDiS dialogue manager allows the user to interact more flexibly and naturally with menu-based interfaces to devices.

3 GoDiS and TrindiKit

In GoDiS, 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.

4 GoDiS features in Dico

To enable flexible dialogue interaction, GoDiS supports (among other things) accommodation, task switching and grounding.

4.1 Accommodation

The applications in Dico II are based on existing menu interfaces, so it is possible for the novice user to let the system take initiative and guide the user through menus. For expert users, *accommodation* strategies enables skipping through the menus and getting right to the point.

Ex. 1: "Call Lisa's home number"

4.2 Multiple simultaneous tasks and task switching

GoDiS enables arbitrarily nested subdialogues. 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, and explicitly signals this. 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 twentythree fourtyfive sixtyseven."
U: "Okay. Changing Lisa's home number to oh three one two three four five six seven."

[1] www.dicoproject.org

4.3 Feedback and grounding

The GoDiS dialogue manager provides general feedback strategies 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 (semantic understanding), understands the meaning of the utterance (pragmatical understanding) and accepts the dialogue moves performed in utterances.

As an example, the single user utterance "Lisa" may result in positive grounding on the semantic level but negative on the pragmatic, resulting in a system utterance consisting of two feedback 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 Multimodal menu-based dialogue in Dico II

While previous versions of Dico did include some multimodal interaction, Dico II is our most ambitious attempt yet at implementing fully the concept of multimodal menu-based dialogue (MMD). Technologies for MMD in menu-based applications have already been developed for other GoDiS applications (Hjelm et al., 2005) and the ideas behind these solutions have been re-implemented and significantly improved in Dico II.

The idea behind MMD is that the user should be able to switch betweem amd combine modalities freely across and within utterances. This should ideally make it possible to use the system using speech only, using traditional GUI interaction only, or using a combination of the two.

MMD enables *integrated multimodality* for user input, meaning that a single contribution can use several input modalities, e.g. *"Call this contact [click]"* where the [click] symbolises haptic input (e.g. amouse click) which in this case selects a specific contact. For output, MMD uses *parallel multimodality*, i.e., output is generally rendered both as speech and as GUI output. To use speech only, the user can merely ignore the graphical output and not use the haptic input device. To enable interaction using GUI only, speech input and output can be controlled using a "push-to-talk" button which toggles between "speech on" and "speech off" mode.

Acknowledgments

The work reported here was funded DICO, Vinnova project P28536-1. Dico II was implemented by the authors, Alex Berman, Fredrik Kronlid, Peter Ljunglöf and Andreas Wiberg. The authors gratefully acknowledge the assistance of Talkamatic AB, Volvo Technology AB, and the DICO project group. The GoDiS system is the property of Talkamatic AB.

References

David Hjelm, Ann-Charlotte Forslund, Staffan Larsson, and Andreas Wallentin. 2005. DJ GoDiS: Multimodal menu-based dialogue in an asychronous isu system. In Claire Gardent and Bertrand Gaiffe, editors, *Proceedings of the ninth workshop on the semantics and pragmatics of dialogue*.

Staffan Larsson and Jessica Villing. 2007. The dico project: A multimodal menu-based in-vehicle dialogue system. In H. C. Bunt and E. C. G. Thijsse, editors, *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*.

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

Anna Olsson and Jessica Villing. 2005. Dico - a dialogue system for a cell phone. Master's thesis, Department of Linguistics, Goteborg University.

David Traum and Staffan Larsson, 2003. *Current and New Directions in Discourse & Dialogue*, chapter The Information State Approach to Dialogue Management, pages 325–353, 28 pages. Kluwer Academic Publishers.

Jessica Villing and Staffan Larsson. 2006. Dico - a multimodal in-vehicle dialogue system. In D. Schlangen and R. Fernandez, editors, *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue*.

Demonstration of the Amani Tactical Questioning Dialogue System

Ron Artstein Sudeep Gandhe Michael Rushforth Nicolle Whitman
Sarrah Ali Jillian Gerten Anton Leuski Antonio Roque
David DeVault David Traum
Institute for Creative Technologies, University of Southern California
13274 Fiji way, Marina del Rey, CA 90292, USA
<lastname>@ict.usc.edu

Amani is a character implemented in a third-generation tactical questioning dialogue system, intended to train students in extracting information through interview. Amani responds to user speech with synthesized voice and gestures. She employs a robust statistical classifier to map user utterances to a limited set of dialogue acts, which she uses to reason about the conversation. Amani's dialogue manager includes the ability to answer a question either truthfully or falsely (lying), withhold information until certain demands are met, respond to compliments and insults, offers and threats, and build rapport with the user. The dialogue act representation is intentionally kept minimalistic, allowing much faster creation and adjustment of scenarios than in a full-fledged virtual human. The system and dialogue act representation are described in a full paper in this volume.

The project or effort described here 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.

Multimodal interaction control in the MonAMI Reminder

Gabriel Skantze
Dept. of Speech Music and Hearing
KTH, Stockholm, Sweden
gabriel@speech.kth.se

Joakim Gustafson
Dept. of Speech Music and Hearing
KTH, Stockholm, Sweden
jocke@speech.kth.se

Abstract

In this demo, we show how attention and interaction in multimodal dialogue systems can be managed using head tracking and an animated talking head. This allows the user to switch attention between the system and other humans. A preliminary evaluation in a tutoring setting shows that the user's attention can be effectively monitored with this approach.

1 Introduction

Most spoken dialogue systems are based on the assumption that there is a clear beginning and ending of the dialogue, during which the user pays attention to the system constantly. However, as the use of dialogue systems is extended to settings where several humans are involved, or where the user needs to attend to other things during the dialogue, this assumption is obviously too simplistic (Horvitz et al., 2003). When it comes to interaction, a strict turn-taking protocol is often assumed, where user and system wait for their turn and deliver their contributions in whole utterance-sized chunks. If system utterances are interrupted, they are treated as either fully delivered or basically unsaid.

In this demo, we show how attention and interaction in multimodal dialogue systems can be managed using head tracking and an animated talking head. This allows the user to switch attention between the system and other humans, and for the system to pause and resume speaking.

2 The MonAMI Reminder

This study is part of the 6th framework IP project MonAMI. The goal of the MonAMI project is to develop and evaluate services for elderly and disabled people. Based on interviews with potential users in the target group, we have developed the MonAMI Reminder, a multimodal spoken dialogue system which can assist elderly and disabled people in organising and initiating their daily activities (Beskow et al., submitted). Information in their personal calendars can be added using digital pen and paper, allowing the user to continue using a paper calendar, while the written events are automatically transferred to a backbone (Google Calendar). The dialogue system is then used to get reminders, as well as to query and discuss the content of the calendar.

The system architecture is shown in Figure 1. A microphone and a camera are used for system input (speech recognition and head tracking), and a speaker and a display are used for system output (an animated talking head). As can be seen in the figure, all system input and output is monitored and controlled by an Attention and Interaction Controller (AIC). The purpose of the AIC is to act as a low level monitor and controller of the system's speaking and attentional behaviour. The AIC uses a state-based model to track the attentional and interactional state of the user and the system. The system is initially in a non-attentive state, in which the animated head looks down. As the user starts to look at the system, the animated talking head looks up and the system may react to what the user is saying. If the user looks away while the system is speaking, the system will pause and resume when the user looks back. If the user starts to speak while the system is speaking, the controller will make sure that the system pauses. The system may then decide to answer the new request, simply ignore it and resume speaking (e.g., if the confidence is too low), or abort speaking (e.g., if the user told the system to shut up).

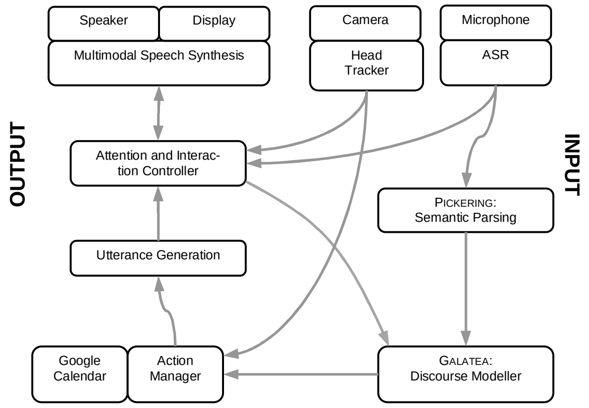

Speaker
Display
Multimodal Speech Synthesis
Camera
Head Tracker
Microphone
ASR
OUTPUT
INPUT
Attention and Interaction Controller
PICKERING: Semantic Parsing
Utterance Generation
Google Calendar
Action Manager
GALATEA: Discourse Modeller

Figure 1. The system architecture in the MonAMI Reminder.

3 Preliminary evaluation

In the evaluation, we not only wanted to check whether the AIC model worked, but also to understand whether user attention could be effectively modelled using head tracking. Similarly to Oh et al. (2002), we wanted to compare “look-to-talk” with “push-to-talk”. To do this, we used a human-human-computer dialogue setting, where a tutor was explaining the system to a subject (shown in Figure 2). Thus, the subject needed to frequently switch between speaking to the tutor and the system. A second version of the system was also implemented where the head tracker was not used, but where the subject instead pushed a button to switch between the attentional states (a sort-of push-to-talk). 8 subjects were used in the evaluation, 4 lab members and 4 elderly persons in the target group (recruited by the Swedish Handicap Institute).

An analysis of the recorded conversations showed that the head tracking version was clearly more successful in terms of number of misdirected utterances. The subjects almost always looked at the addressee in the head tracking condition, and did not start to speak before the animated head looked up. When using the push-to-talk version, however, they often forgot to “turn it off”, which resulted in the system interpreting utterances directed to the tutor and started to speak when it shouldn’t. The addressee of the utterances in the push-to-talk condition was correctly classified in 86.9% of the cases, as compared with 97.6% in the look-to-talk condition.

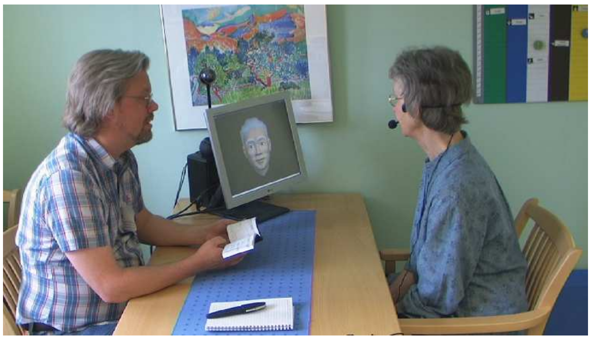


Figure 2. The human-human-computer dialogue setting used in the evaluation. The tutor is sitting on the left side and the subject on the right side

These finding partly contradict findings from previous studies, where head pose has not been that successful as a sole indicator for the addressee (cf. Bakx et al., 2003; Katzenmaier et al., 2004). One explanation for this might be that the subjects were explicitly instructed about how the system worked. Another explanation is the clear feedback (and entrainment) that the agent’s head pose provided.

Acknowledgements

This research is supported by MonAMI, an Integrated Project under the European Commission’s 6th Framework Program (IP-035147), and the Swedish research council project GENDIAL (VR #2007-6431).

References

Bakx, I., van Turnhout, K., & Terken, J. (2003). Facial orientation during multi-party interaction with information kiosks. In *Proceedings of the Interact 2003*.

Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., & Tobiasson, H. (submitted). The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. Submitted to *Interspeech 2009*.

Horvitz, E., Kadie, C. M., Paek, T., & Hovel, D. (2003). Models of attention in computing and communication: from principles to applications. *Communications of the ACM*, 46(3), 52-59.

Katzenmaier, M., Stiefelhagen, R., Schultz, T., Rogina, I., & Waibel, A. (2004). Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech. In *Proceedings of ICMI 2004*.

Oh, A., Fox, H., Van Kleek, M., Adler, A., Gajos, K., Morency, L-P., & Darrell, T. (2002). Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment. In *Proceedings of CHI 2002*.

Spontal – first glimpses of a Swedish database of spontaneous dialogue

Jens Edlund
KTH
Stockholm, Sweden
edlund@speech.kth.se

Abstract

This demonstration provides a first glimpse of a large multimodal database of Swedish spontaneous dialogue that is currently being collected within the ongoing project *Spontal: Multimodal database of spontaneous speech in dialog*. This accompanying paper briefly gives background and motivation for the project.

1 Introduction

This demonstration provides a first glimpse of *Spontal: Multimodal database of spontaneous speech in dialog*. The demonstration will touch upon annotation, audio, video and motion capture data, the recording studio, and some initial analyses.

2 Background

Spontal is an ongoing data collection project aimed at gathering multimodal data on spontaneous spoken dialogues. The project, which began in 2007 and will be concluded in 2010 and is funded by the Swedish Research Council, KFI - Grant for large databases (VR 2006-7482), It takes as its point of departure the fact that both vocal signals and gesture involving the face and body are key components in everyday face-to-face interaction – arguably the context in which speech was borne – and focuses in particular on spontaneous conversation.

There is a lack of data with which we can make more precise measurements of many aspects of spoken dialogue. We have for example an increasing understanding of the vocal and visual aspects of conversation, but there is little data with which we can measure with precision multimodal aspects such as the timing relationships between vocal signals and facial and body gestures. Furthermore, we need data to gauge acoustic properties that are specific to conversation, as opposed to read speech or monologue, such as those involved in floor negotiation, feedback and grounding, and resolution of misunderstandings. As a final example, there is a current surge in research on the related topics of incremental processing in dialogue on the on hand, and synchronous and converging behavior of interlocutors on the other – studies that are also hampered by a lack of data.

3 Scope

120 half-hour dialogues, resulting in a total in excess of 60 hours, will be recorded in the project. Sessions consist of three consecutive 10 minute blocks. All subjects are native speakers of Swedish and balanced (1) for gender, (2) as to whether the interlocutors are of opposing gender and (3) as to whether they know each other or not. The balancing results in 15 dialogues of each configuration: 15x2x2x2 for a total of 120 dialogues. Currently (May, 2009), about 45% of the database has been recorded. The remainder is scheduled for recording during 2009. Subjects permit, in writing, that (1) the recordings are used for scientific analysis, that (2) the analyses are published in scientific writings and that (3) the recordings can be replayed in front of audiences at scientific conferences and suchlike.

The recordings are comprised of high-quality audio and high-definition video. In addition, a motion capture system is used on virtually all recordings to capture body and head gestures, although the treatment and annotation of this data are outside the scope of the project and for this, resources have yet to be allocated.

4 Instruction and scenarios

Subjects are told that they are allowed to talk about absolutely anything they want at any point

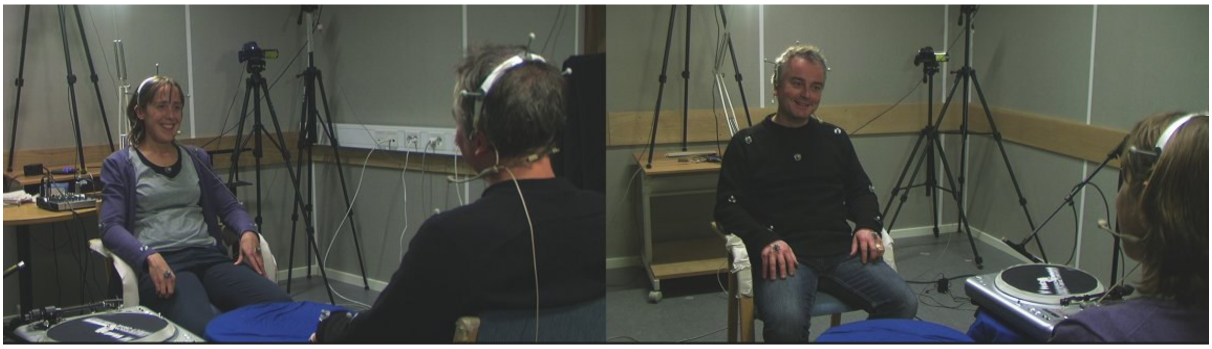


Figure 1. Example showing one frame from the two video cameras taken from the Spontal database.

in the session, including meta-comments on the recording environment and suchlike, with the intention to relieve subjects from feeling forced to behave in any particular manner. The recordings are formally divided into three 10 minute blocks, although the conversation is allowed to continue seamlessly over the blocks, with the exception that subjects are informed, briefly, about the time after each 10 minute block. After 20 minutes, they are also asked to open a wooden box which has been placed on the floor beneath them prior to the recording. The box contains objects whose identity or function is not immediately obvious.

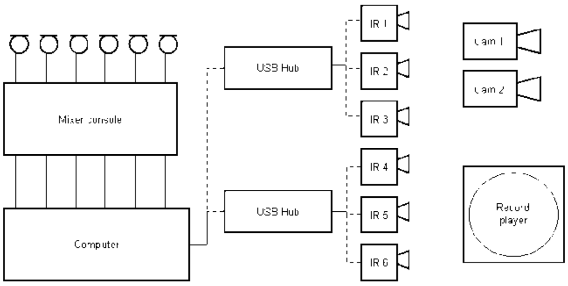


Figure 2. Schematic representation of the recording setup.

5 Technical specifications

The recording setup is illustrated in Figure 2. The audio is recorded on four channels using a matched pair of omni-directional microphones for high audio quality, and two headset microphones to facilitate subject separation for transcription and dialogue analysis. Two high definition video cameras are placed to obtain a good view of each subject from a height that is approximately the same as the heads of the subjects. The cameras work at 1920x1080 resolution at a bitrate of 26.6 Mbps. Audio, video and motion-capture are synchronized during post-processing with the help of a turntable placed between the subjects and a bit to the side, in full view of the motion capture cameras. A motion capture marker is placed near the edge on the turntable which rotates with a constant speed (33 rpm), enabling high-accuracy synchronization.

Figure 1 shows a frame from each of the two video cameras next to each other, so that both dialogue partners are visible. The opposing video camera can be seen centrally in the images, and a number of tripods with motion capture cameras are visible. Figure 3, finally, shows a 3D representation of motion-capture data. Each of the dots correspond to a reflective marker placed on the interlocutors' hands, arms, shoulders, trunks and heads, as can be seen in Figure 1.

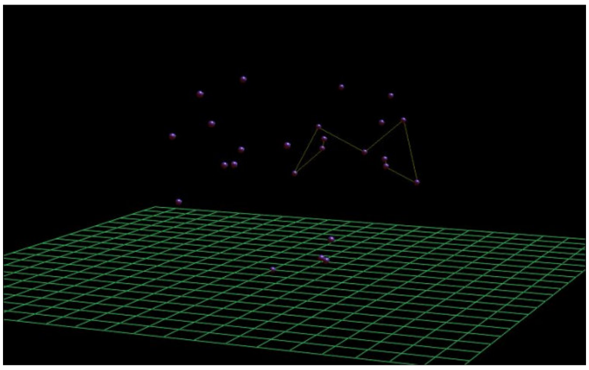


Figure 3. A single frame of motion-capture data from a Spontal dialogue.

The Spontal database will be made available to the research community after project completion. When recorded in its entirety, the Spontal database will be the largest of its kind in the world, and one of the richest dialogue data resources in Sweden.

Acknowledgments

Thanks are due to the Swedish Research Council for funding and support, the Spontal team at KTH Speech Music and Hearing for everything, and last but not least to all the volunteering subjects for making the data collection possible in the first place.

Prosodic Disambiguation in Spoken Systems Output

Samer Al Moubayed
KTH Centre for Speech Technology, Stockholm, Sweden
sameram@kth.se

Abstract

This paper presents work on using prosody in the output of spoken dialogue systems to resolve possible structural ambiguity of output utterances. An algorithm is proposed to discover ambiguous parses of an utterance and to add prosodic disambiguation events to deliver the intended structure. By conducting a pilot experiment, the automatic prosodic grouping applied to ambiguous sentences shows the ability to deliver the intended interpretation of the sentences.

1 Introduction

In using natural language in human computer interfaces, we expose ourselves to the risk of producing ambiguity – a property of natural language that distinguishes it from artificial languages. We may divide linguistic ambiguity broadly into lexical ambiguity involving single linguistic units and structural ambiguity – when an utterance can be parsed in more than one way as in:

"I ate the chocolate on the desk." (1)

In many cases, structurally ambiguous utterances are not communicatively ambiguous as in:

"I drank the water from the bottle" (2)

The sentence in (2) has the same syntactic structure as in (1) but is not communicatively ambiguous as common knowledge resolves the ambiguity. In some cases, the structural ambiguity can lead to communicative ambiguity that needs to be resolved.

A growing body of research demonstrates that listeners are sensitive to prosodic information in the comprehension of spoken sentences. Rowles & Huang (1992) show how prosody can aid the syntactic parsing of spoken English in automatic speech recognition systems. Others have also associated pitch with prosodic grouping and disambiguation (e.g. Schafer et al., 2000), as well as pauses (e.g. Kahn et al., 2005). Allbritton, McKoon & Ratcliff (1996) conclude that speakers do not *always* use prosody to resolve ambiguity simply due to unawareness of its existence. There is also a great body of work on the use of prosody in computer generated speech, but to our knowledge there is no study to date on using prosody as a disambiguation tool in computer generated speech.

In this paper, we explore the possibility of automating prosodic disambiguation of computer generated speech in spoken dialogue systems to avoid communicating ambiguity. We assume that the system has access to the syntactic structure of the utterances it generates.

2 Placement of prosodic disambiguation

For the present purposes, we will assume a system modeling its possible utterances with binary CFG grammars, noting that any CFG grammar can be transformed into a binary one. A miniature grammar is provided in Figure 1, which generates a simple PP-attachment ambiguity. If a system produces such a potential communicative ambiguity, we need to know exactly where the ambiguity takes place in order to group the relevant sequence of words more clearly and prevent unintended interpretations. Figure 2 shows the parsing of the sentence: *"I ate <the chocolate on the desk>"* generated by the grammar in Figure 1, using chart parser style representation. In the chart, black trajectories are rules shared by all parses, green ones exist in the required parse tree only, and red ones are not part of the required parse trajectories while they exist in other parses. We see that the green trajectory must be grouped, as the words covered by this trajectory could otherwise be grouped in other ways, according to the grammar rules. Grouping them along the green trajectory distinguishes the intended parse from other parses. The trajectory is defined by its start and end nodes, hence the green trajectory is unique in that it is the only one starting and ending at those nodes.

R1: S → NP VP
R2: VP → V
R3: NP → Noun | Noun PP
R4: Noun → N | Det N
R5: PP → Prep NP
R6: V → Verb NP
R7: VP → V PP

Figure 1: Simple CFG grammar for English. The grammar generates PP-attachment ambiguity.

The disambiguation strategy suggested here, then, is to prosodically group a set of words only when not grouping them could result in a different parse, and ultimately a different interpretation.

3 A Pilot Experiment

As a listening test of the interpretation enhancement of the automatic disambiguation grouping of the previous algorithm, 15 sentences with coordination or PP-attachment ambiguities were generated using an in-house TTS. This system has a phrasing property implemented. 5 sentences of these were communicatively unambiguous but structurally ambiguous, and the rest were communicatively ambiguous. To ensure that the preferred meaning of these sentences is not taken into account, one subject had listened to these computer generated sentences without any grouping and gave her interpretation, we will call this subject "Subject A". Subsequently these sentences were introduced to two subjects after disambiguating them using prosodic grouping. These sentences contained PP-attachment and coordination ambiguity, and generated only two possible interpretations.

The results of these two subjects are grouped into two groups. The first one is the result of these subjects for sentences disambiguated to deliver the interpretation of the sentences which matched the one given by "Subject A", that is when the sentences do not receive any disambiguation. The other group is the results for the sentences delivering the opposite interpretation.

The result shows that 95% of the sentences received the correct interpretation after disambiguation when the desired interpretation matched this of "Subject A", while 75% of the sentences received the correct interpretation when the sentences disambiguated to deliver the other interpretation than "subject A" interpretation. In addition, the results show that the grouping using the proposed algorithm, as hoped for, did not affect the interpretation of the communicatively unambiguous sentences regardless of the prosodic disambiguation.

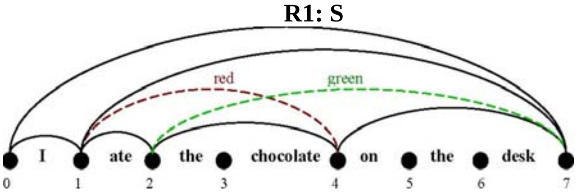


Figure 2: Possible parses of an example sentence. The black arcs are shared by all possible parses of the sentence. The green arcs exist only in the required parse and the red ones do not exist in the required parse by in other possible parses.

4 Conclusions

In this work, we presented an algorithm for spotting ambiguity in synthesized sentences with known syntactic structure. By conducting a small experiment, prosodic grouping (phrasing) is used by the disambiguation algorithm, and the results show high recognition rate by the subjects of the required interpretation of the disambiguation algorithm.

Future studies should focus on testing prosodic disambiguation using large scale grammar, or other types of grammars like PCFG, when disambiguation takes place depending on the probabilities of the multiple parses of the same utterance.

Acknowledgments

This work is supported by MonAMI, an Integrated Project under the European Commission's Sixth Framework Program (IP-035147).

Many Thanks to Jens Edlund for the interesting and helpful discussions.

References

Allbritton, D.W., McKoon, G., Ratcliff, R. "Reliability of prosodic cues for resolving syntactic ambiguity. Journal of Experimental Psychology: Learning, Memory, & Cognition, 22, 714-135, 1996.

Chris, Rowles., Xiuming, Huang. "Prosodic Aid to Syntactic and Semantic Analysis of Spoken English", in Proceedings of the 30st Annual Meeting of the Association for Computational Linguistic pages 112-119.ACL, 1992.

Jermy G. Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, Mari Osterdorf. "Effective use of prosody in parsing conversational speech", Proceedings of the Conference in Human Language Technology and Empirical Methods in Natural Language Processing, p.233-240, Vancouver, British Columbia, Canada. 2005.

Schafer, Amy J., Shari R. Speer, Paul Warren & S. David White. "Intonational Disambiguation in sentence production and comprehension". Journal of Psycholinguistic Research, 29, 169-182. 2000.

Learning Adaptive Referring Expression Generation Policies for Spoken Dialogue Systems using Reinforcement Learning

Srinivasan Janarthanam
School of Informatics
University of Edinburgh
s.janarthanam@ed.ac.uk

Oliver Lemon
School of Informatics
University of Edinburgh
olemon@inf.ed.ac.uk

Abstract

Adaptive generation of referring expressions in dialogues is beneficial in terms of grounding between the dialogue partners. However, handcoding adaptive REG policies is hard. We present a reinforcement learning framework to automatically learn an adaptive referring expression generation policy for spoken dialogue systems.

1 Introduction

Referring expression generation (REG) is the natural language generation (NLG) problem of choosing the referring expressions for use in utterances to refer to various domain objects. Adaptive REG could help in efficient grounding between dialogue partners (Issacs and Clark, 1987), improve task success rates or even increase learning gain. For instance, in a technical support task, the dialogue agent could use technical jargon with experts, descriptive expressions with beginners and a mixture of the two with intermediate users. Similarly, in a city navigation task, the dialogue agent could use proper names for landmarks with locals but descriptive expressions with foreign tourists. Although adapting to users seems beneficial, adapting to an unknown user is tricky and hand coding such adaptive REG policies is a cumbersome work. (Lemon, 2008) first presented the case for treating NLG as a reinforcement learning problem. In this paper, we extend the framework to automatically learn an adaptive REG policy for spoken dialogue systems.

2 Related work

Reinforcement Learning (RL) has been successfully used for learning dialogue management policies (Levin et al., 1997). The learned policies allow the dialogue manager to optimally choose appropriate instructions, confirmation requests, etc. In contrast, we present an RL framework to learn REG policies.

3 Reinforcement Learning Framework

A basic RL setup consists of a learning agent, its environment and a reward model (Sutton and Barto, 1998). The learning agent explores by taking different possible actions in different states and exploits the actions for which the environmental rewards are high. In our model, the learning agent is the NLG module of the dialogue system, whose objective is to learn an REG policy. The environment consists of a user who interacts with the dialogue system. Since learning occurs over thousands of interaction cycles, real users are replaced by user simulations that simulate real user's dialogue behaviour. In the following sections, we discuss the salient features of the important components of the architecture in the context of a technical support task (Janarthanam and Lemon, 2009a).

3.1 Dialogue Manager

The dialogue manager is the central component of the dialogue system. Given the dialogue state, it identifies the next dialogue act to give to the user. The dialogue management policy is modelled on a simple handcoded finite state automaton. It issues step by step instructions to complete the task and also issues clarifications on REs used when requested by the user.

3.2 NLG module

The task of the NLG module is to translate the dialogue act into a system utterance. It identifies the REs to use in the utterance to refer to the domain objects. As a learning agent in our model, it has three choices - jargon, descriptive and tutorial. Jargon expressions are technical terms like 'broadband filter', 'ethernet cable', etc. Descriptive expressions contain attributes like size, shape and color. e.g. 'small white box', 'thick cable with red ends', etc. Tutorial expressions are a combination of the two. The decision to choose one expression over the other is taken based on the user's domain knowledge, which is updated progressively in a user model (state) during the conversation.

3.3 User Simulation

In order to enable the NLG module to evaluate the REG choices, our user simulation model is responsive to the system's choice of REs. For every dialogue session, a new domain knowledge profile is sampled. Therefore, for instance, a novice profile will produce novice dialogue behaviour with lots of clarification requests. For user action selection, we propose a two-tiered model. First, the system's choice of referring expressions ($REC_{s,t}$) is examined based on the domain knowledge profile (DK_u) and the dialogue history (H). This step is more likely to produce a clarification request ($CR_{u,t}$) if the REs are unknown to the user and have not be clarified earlier.

$$P(CR_{u,t}|REC_{s,t}, DK_u, H)$$

If there are no clarification requests, then issue an appropriate user action ($A_{u,t}$) based on the system's instruction ($A_{s,t}$) and if there is one, the user action will be the clarification request itself.

$$P(A_{u,t}|A_{s,t}, CR_{u,t})$$

These parameters are set empirically by collecting real user dialogue data using wizard-of-Oz experiments (Janarthanam and Lemon, 2009b).

4 Learning REG policies

REG policies are learned by the NLG module by interacting with the user simulation in the learning mode. The module explores different possible state-action combinations by choosing different REs in different states. At the end of each dialogue session, the learning agent is rewarded based on parameters like dialogue length, number of clarification requests, etc. The magnitude of the reward allows the agent to reinforce the optimal moves in different states. Ideally, the agent gets less reward if it chooses the inappropriate REs, which in turn results in clarfication requests from the user. The reward model parameters can be set empirically using wizard-of-Oz data (Janarthanam and Lemon, 2009b). The learned policies predict optimal REs based on the patterns in knowledge. For instance, a user who knows 'broadband cable' will most likely know 'ethernet cable'.

5 Evaluation

Learned policies can be evaluated using the user simulation and real users. Policies are tested to see if they produce optimal moves for the given knowledge profiles. Learned policies can be compared to hand-coded baseline policies based on parameters like dialogue length, learning gain, etc. Real users are asked to rate the system based in its adaptive features after their interaction with the dialogue system.

6 Conclusion

A framework to automatically learn adaptive REG policies in spoken dialogue systems using reinforcement learning has been presented. Essential features to learn an adaptive REG policy have been highlighted. Although the framework is presented in the context of a technical support task, the same is suitable for many other domains.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework (FP7) under grant agreement no. 216594 (CLASSiC Project www.classic-project.org), EPSRC project no. EP/E019501/1, and the British Council (UKIERI PhD Scholarships 2007-08).

References

E. A. Issacs and H. H. Clark 1987. *References in conversations between experts and novices.* Journal of Experimental Psychology: General, 116(26-37).

S. Janarthanam and O. Lemon. 2009a. *Learning Lexical Alignment Policies for Generating Referring Expressions for Spoken Dialogue Systems.* Proc. ENLG'09.

S. Janarthanam and O. Lemon. 2009b. *A Wizard-of-Oz environment to study Referring Expression Generation in a Situated Spoken Dialogue Task.* Proc. ENLG'09.

O. Lemon. 2008. *Adaptive Natural Language Generation in Dialogue using Reinforcement Learning.* Proc. SEMdial'08.

E. Levin, R. Pieraccini and W. Eckert. 1997. *Learning Dialogue Strategies within the Markov Decision Process Framework.* Proc. ASRU97.

R. Sutton and A. Barto. 1998. *Reinforcement Learning.* MIT Press.

Ontology-Based Information States for an Artificial Sales Agent

Núria Bertomeu
Centre for General Linguistics (ZAS)
Berlin, Germany
bertomeu@zas.gwz-berlin.de

Anton Benz
Centre for General Linguistics (ZAS)
Berlin, Germany
benz@zas.gwz-berlin.de

Abstract

This paper presents an approach to the representation of dialogue states in terms of information states and joint projects, on the basis of which we are modelling a non-player character (NPC) with natural dialogue capabilities for virtual environments.

1 Introduction

In order to gather data on how humans interact with NPCs we collected a corpus by means of a Wizard-of-Oz experiment consisting of 18 dialogues of one hour of duration (Bertomeu and Benz, 2009). We simulated a scenario where the NPC played the role of an interior designer and helped the customer furnishing a living-room. The following dialogue gives a glimpse of the data:

(1) USR.1: And do we have a little sideboard for the TV?
NPC.3: What about this one?
USR.5: Is there a black or white sideboard?
NPC.6: No I'm afraid not, they are all of light or dark wood.
USR.6: Ok, then I'll take this one.
NPC.7: All right.

Our investigation of the data aims at addressing questions relevant for the development of dialogue models for NPCs, e.g. which action should an NPC carry out given a particular context. For this, we need to annotate not only the actions performed by the dialogue participants (DPs), but also the changes that these actions produce in the information state (IS) shared by them. As the dialogue is oriented to the task of furnishing a room, the ISs must contain a partial domain model which keeps track of the objects selected so far, and of the topics under discussion. We will use here the term information state (IS) to denote the information which has been established during the dialogue: concretely, the parameter values already fixed and the parameter values under discussion and under consideration, similar e.g. to Ginzburg's Dialogue Gameboard[1] (Ginzburg, 1995).

We developed an annotation scheme from which the ISs and their updates can be automatically generated. Interestingly, the ISs are closely related to the ontology used for representing the domain objects, i.e. rooms, furniture, wall-covers, etc. The ontology-based domain model allows the NPC to change the order in which topics are addressed at any time according to the user initiatives, resulting thus in a more flexible and natural dialogue.

Regarding the annotation of ISs, Poesio et al. (1999) have carried out a pilot study for the annotation of ISs, concluding that these are not suitable for large-scale annotation, because the task is time-consuming and difficult. Georgila et al. (2005) have automatically annotated ISUs in the COMMUNICATOR corpus. However, since the content of ISs is domain and task-specific such a procedure is not easily transferable to our corpus.

2 Projects and information states

We took a bottom-up approach to the analysis by choosing as our annotation unit minimal joint projects (Clark, 1996). Minimal joint projects are adjacency pairs which have a purpose and carry out an update of the IS. Each adjacency pair divides into an *initiating* and an *completing* act. A joint project is annotated for its function, its goal, whether it contains embedded projects, the common IS, and the initiating and completing actions. The actions are further specified ac-

[1] It should be noted that the information states in the Information State Update (ISU) framework, e.g. (Poesio et al., 1999), are richer in content than our representations, since they contain information on the individual dialogue moves and representations of goals and agendas.

cording to the act they perform and their role in the project, among other information. An example of an *initiating* act can be found in Fig. 1. The representation shows that the PARAMETER_UNDER_DISCUSSION addressed by the act is the location *l1* of a shelves item[2].

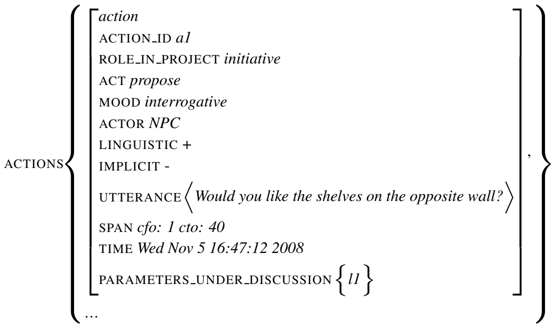

ACTIONS
action
ACTION_ID a1
ROLE_IN_PROJECT initiative
ACT propose
MOOD interrogative
ACTOR NPC
LINGUISTIC +
IMPLICIT -
UTTERANCE ⟨Would you like the shelves on the opposite wall?⟩
SPAN cfo: 1 cto: 40
TIME Wed Nov 5 16:47:12 2008
PARAMETERS_UNDER_DISCUSSION {l1}
...

Figure 1: The initiating act related to the utterance: *Would you like the shelves on the opposite wall?*

The common IS will only be updated after a joint project has been completed. If the completing act of the addressee accepts the proposed location, the IS will be updated as shown in Fig. 2.

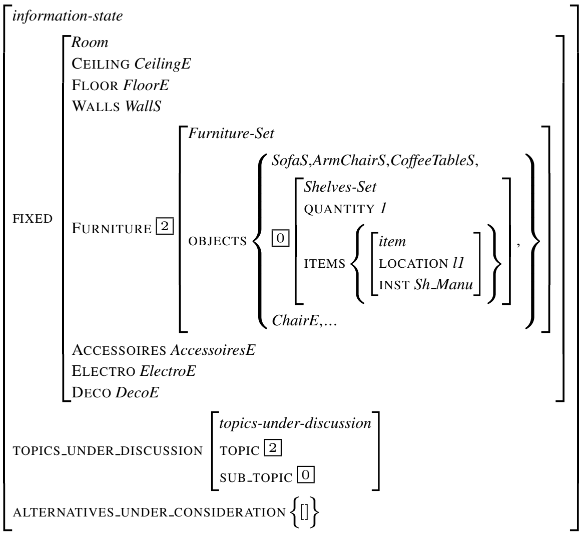


Figure 2: An information state

The value of FIXED is the feature-structure (FS) representation of a room as specified in the ontology. A room consists of different types of objects, such as furniture, decoration, etc. Furniture in turn includes sofas, arm-chairs, shelves, etc. The representation shows that *Shelves*, and thus *Furniture*, are currently under discussion. It also shows that the location of the chosen shelves has been fixed to be *l1*. *Fixing* information means agreeing on a value for a parameter. It may happen, though, that several values for a parameter are entertained simultaneously. This occurs e.g. if the user asks for another item of the same type without rejecting the item which has been under discussion before. Therefore, a set of ALTERNATIVES_UNDER_CONSIDERATION must be represented. Whenever an agreement is reached, this set is emptied.

The ISs are not annotated directly. They are automatically extracted from the annotation of the individual parameters addressed by the actions and the dialogue acts performed by those, and encoded in FSs following the TEI-P5 guidelines[3]. This procedure makes their annotation feasible.

3 Conclusion

For developing an artificial sales agent, we need a fine-grained representation of ISs and their updates. In particular, the topics under discussion and their discourse status as *open*, *fixed*, or *under consideration* are an essential aspect for planning a discourse strategy. We managed to develop an ontology-based format for representing ISs which is rich enough to fulfil these tasks, and came up with an annotation methodology which makes hand-coding feasible. For the future, an automatic extraction of a finite state description of the sales scenario is planned.

References

N. Bertomeu and A. Benz. 2009. Annotation of Joint Projects and Information States in Human-NPC Dialogue In *Proceedings of CILC-09*. Murcia.

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

J. Ginzburg. 1995. Resolving questions. *Linguistics and Philosophy*, 18:5, 459 – 527.

K. Georgila, O. Lemon and J. Henderson. 2005. Automatic annotation of COMMUNICATOR dialogue data for learning dialogue strategies and user simulations. In *Ninth Workshop on the Semantics and Pragmatics of Dialogue DIALOR*.

M. Poesio, R. Cooper, C. Matheson, D. Traum. 1999. Annotating conversations for Information State Update. *Dialogue*. Amsterdam University.

[2] *l1* is the id of the location referred to by '*on the opposite wall*'.

[3] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

Alignment and Priming of Spatial Perspective

Elena Andonova
Bremen University
Bremen, Germany
andonova@uni-bremen.de

Kenny R. Coventry
Northumbria University
Newcastle, United Kingdom
kenny.coventry@northumbria.ac.uk

Abstract

Research on interactive alignment has provided evidence for lexical and syntactic priming but little is known about alignment at the conceptual level. In this study we tested for effects of priming (and alignment) of spatial perspective in a route description task within a confederate design which consisted of an early and a later experimental block. Indeed, participants' choice of spatial perspective was affected by the preceding perspective choice in confederates' descriptions on both the early and the later experimental blocks but there was no interaction between early and later priming. Furthermore, individual differences in spatial ability as measured by a mental rotation task did not play a significant role in degree of priming.

1 Introduction

The interactive alignment framework (Pickering & Garrod, 2004) posits that much of language production in a dialogic situation can be explained via automatic priming mechanisms. Whereas previous confederate paradigm studies have shown lexical priming and syntactic priming effects across interlocutors, conceptual alignment has not been addressed in this way. Here a first confederate paradigm study examined alignment of spatial perspective in a task where participants took turns in describing routes on schematic maps.

A route and an environment can be described from an external or allocentric view in a survey perspective, and from an embedded, or egocentric view in a route perspective. The results revealed that participants' responses on an early experimental block were indeed affected by confederate priming. The alignment effect, however, did not emerge on the later experimental block. Thus, spatial perspective alignment appears to occur when a speaker encounters consistent perspective choices by the interlocutor and is weakened if the interlocutor lacks a stable preference or switches perspective.

2 Experimental Method

In a second study, we examined to what extent our findings on alignment of perspective in a confederate paradigm task can be replicated in a perspective priming task which would be indicative of common mechanisms at play. Participants took turns with confederates in describing routes on a series of maps in an early and in a later block of four maps each. Confederates started first and their descriptions followed a script that manipulated spatial perspective systematically: it was either consistently route, consistently survey, route switching to survey, or survey switching to route. We also included a measure of spatial ability in order to establish how much of the alignment and/or priming performance can be modulated by individual differences. It may be easier for participants with higher spatial ability to adopt a certain perspective choice than other participants given the underlying cognitive demands for switching between alternative views in priming in the route describing task in our experiments and the mental rotation task used as a measure of spatial ability.

3 Results

In this study, participants' choices of spatial perspective were affected by confederate priming—after hearing a confederate use a route perspective, an average of 66% of the descriptions (SD=38%) on the early block of trials were in the same route perspective while only 24% (SD=30%) were in the survey perspective, F=19.627, p<.001, η_p^2=.282. Priming of a similar magnitude also occurred on the later block where route perspective descriptions by the confederates were followed by 69% (SD=38%) route perspective descriptions and 25% (SD=39%) survey perspective descriptions by the participants, F=16.957, p<.001, η_p^2=.253. Furthermore, there was an effect of early priming on the later participants' responses as well, F=.8.128, p=.006, η_p^2=.145, and no interaction between early prime and later prime. These two priming sources have an additive effect.

We were particularly interested in examining the role that individual differences in spatial ability may play in priming participants' choices. Participants were divided into two groups on the basis of their Mental Rotation Test (MRT) scores: low MRT (M=5.85, range 3-8) and high MRT (M=11.80, range 9-17) performance. Although participants in the high MRT group were primed more by confederate use of the survey perspective on the early block (85% survey vs. 33.5% route perspective descriptions) than those in the low MRT group (74% survey vs. 34% route perspective, respectively), the interaction between spatial ability and primed perspective did not reach significance. There was even less difference across low and high MRT groups on the degree of priming in the later block.

4 Conclusion

We conclude that both priming of spatial perspective and alignment in spatial perspective can occur in highly similar tasks, which implicates common underlying components of the two phenomena. On the other hand, although individual differences in priming and alignment deserve further exploration, individual spatial ability appears to play a minor role here.

Acknowledgments

We gratefully acknowledge the support of the DFG to the I5-[DiaSpace] project as part of the SFB/TR 8 Spatial Cognition and the assistance of our colleagues and student assistants.

Reference

Pickering, M., and Garrod, S. 2004. The Interactive Alignment Model. *Behavioral and Brain Sciences,* 27(2):169-189.

Using Screenplays as Corpus for Modeling Gossip in Game Dialogues

Jenny Brusk
Dept of Game Design, Narrative and Time-based Media
Gotland University
Graduate School of Language Technology
Sweden
jenny.brusk@hgo.se

Abstract

We present a dialogue model for handling gossip conversations in games. The model has been constructed by analyzing excerpts from sitcom scripts using Eggins and Slade's conversational structure schema of the gossip and opinion genre. We mean that there are several advantages in using screenplays rather than transcriptions of human dialogues for creating game dialogues: First, they have been tailored to suit the role characters. Second, they are based on fiction, just like games. Third, they reflect an "ideal" human conversation. The model is expressed using Harel statecharts and an example of an analysis of one script excerpt is given.

1 Introduction

In this paper we will argue that game dialogues have more in common with dialogues between role characters in a screenplay than dialogues between humans in a natural setting. The arguments we have found motivates using screenplays as corpora rather than transcriptions of ordinary conversations between humans. As an example, we will show how Eggins and Slade's (1997) and Horvath's and Eggins (1995) schema for analyzing gossip and opinions can be applied on a given excerpt from one famous sitcom, Desperate Housewives (2004). Eggins and Slade define gossip as a conversation in which the speakers make pejorative statements about and absent third person, so we have chosen an excerpt that fills this criterion.

The reason why we primarily have taken an interest in gossip and opinion is that we think that a game character that can engage in these types of activities will be more interesting to interact with. Gossip can then for instance be used to get informal information about other characters in the game, and furthermore, since gossip can be potentially face threatening (see e.g. Brown and Levinson, 1987), it can also be used to create characters that appear to have a social awareness and social skills.

1.1 Motivation

There are some significant similarities between dialogues in screenplays and game dialogues: They are both scripted and based on fiction, and they are tailored to fit a particular scene, which means that they have a natural beginning and end, as well as a language use that is consistent with both the role characters as well as the overall theme. One could say that they reflect "ideal" conversations, i.e. conversations in which all uninteresting and unnecessary parts have been removed; hence they are already distilled (Larsson et al, 2000).

There is however one prominent difference between the two: the level of engagement on behalf of the audience. A player of a game is actively engaged in performing actions that affect how the story progresses, whereas the story in a movie is remained unchanged independently of the audience's interferences. In this sense, interacting with a game character is similar to interacting with a traditional conversational agent (CA), also because they both serve as an interface to an underlying system. But when a CA typically is used as a substitute for a human, to which the user communicates using his *real identity* (Gee, 2003), a game character has been given a role. And when the player interacts with the game character, he too is expected to play his part, i.e. to use a *projective identity* (ibid).

Speaker	Utterance	Gossip	Opinion
Gabrielle	Can I say something? I'm glad Paul's moving	Third person focus	Opinion
Bree	Gaby!	Probe	Seek evidence
Gabrielle	I'm sorry, but he's just always given me the creeps. Haven't you guys noticed?	Substantiating behavior	Provide evidence
Gabrielle	He has this dark thing going on. There's something about him that just feels…	Pejorative evaluation	Provide evidence
Lynette	Malignant?	Pejorative evaluation	Agree
Gab	Yes	Acknowledgement	
Susan	We've all sorta felt it	Agree	Agree
Bree	That being said, I do love what He's done with the lawn	Wrap-up	Wrap-up

Table 1. Analysis of excerpt from Desperate Housewives

2 A Model of Gossip

Eggins and Slade (1997) have found that gossip has a generic structure that includes the obligatory elements of *Third person focus*, *Substantiating behavior*, and *Pejorative evaluation*. In the *Substantiating behavior* stage the speaker justifies the negative evaluation, which also serves to express the appropriate way to behave.

The opinion genre shows several familiarities with gossip, where opinion is an expression of an attitude towards some person, event or thing (Horvath and Eggins, 1995; Eggins and Slade, 1997). The obligatory elements of opinion are however less than those constituting gossip, and consists solely of an *Opinion* followed by a *Reaction*. When a reaction involves a request for evidence, the structure however becomes more complex. In this case, the conversation might have elements of evidence and finally a resolution (given that the hearer accepts the evidence). An analysis of a scene from Desperate housewives (2004) based on their structure is presented in table 1, above.

From the analysis, we have created a dialogue model using statecharts (Harel, 1987), which really are extended finite state machines, see figure 1, below.

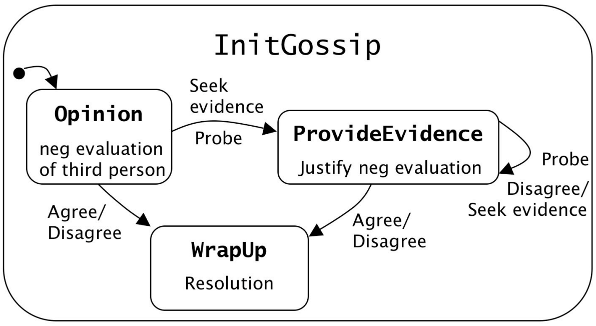


Figure 1. Dialogue model of gossip

The boxes illustrate states that in turn represent the system's (game character) actions. The labeled arcs represent transitions between states that can be triggered by user input (events) and/or conditions that have been satisfied.

To present this as a generic model for gossip, we have to think of the actual function a certain dialogue move has. For instance, when Bree says "Gaby" (line 2 of table 1), it could easily be exchange by a more typical probe, such as "why?" or "How so?". Even if its surface function is to make Gabrielle aware of the inappropriateness of her statement, it also serves to encourage her to continue. If Bree instead would have said "me too", in a dialogue between just the two of them, the gossip could be completed immediately and Gabrielle would not have to substantiate her statement (as in line 3), instead the dialogue could be wrapped up. Worth noticing is that the provide evidence stage can be iterated.

References

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Suzanne Eggins and Diana Slade. 1997. *Analysing Casual Conversation*. Equinox Publishing Ltd.

David Harel. 1987. Statecharts: A Visual Formalism for Complex Systems. *Science of Computer Programming*, 8:231-274.

Horvath and Suzanne Eggins. 1995. Opinion Texts in Conversation. In Peter H. Fries and Michael Gregory (eds) *Discourse in Society: Systemic Functional Perspectives*. Ablex Norwood NJ, pp 29-46.

Staffan Larsson, Arne Jönsson and Lena Santamarta 2000. Using the process of distilling dialogues to understand dialogue systems. In *Proceedings of ICSLP 2000*, Beijing, China, pp. 374-377.

Tom Spezialy and Mark Cherry. 2004. Who's That Woman?. *Desperate Housewives*, season 1, episode 4. Touchstone Television.

The Acquisition of a Dialog Corpus with a Prototype and two WOz[*]

L.F. Hurtado, E. Segarra, E. Sanchis, F. García
Dept. de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera sn, 46022 València
{lhurtado, esegarra, esanchis, fgarcia}@dsic.upv.es

D. Griol
Departamento de Informática
Universidad Carlos III de Madrid
Av. Universidad 30, 28911 Leganés
dgriol@inf.uc3m.es

Abstract

In this paper, we present our approach to simplify the dialog corpus acquisition task. This approach is based on the use of a prototype of the dialog manager and two Wizards of Oz.

1 Introduction

The development of spoken dialog systems is a complex process that involves the design, implementation and evaluation of a set of modules that deal with different knowledge sources. Currently, one of the most successful approaches is based on statistical models, which represent the probabilistic processes involved in each module, whose corresponding models are estimated by means of corpora of human-machine dialogs (Williams and Young, 2007; Griol et al., 2008). The success of statistical approaches highly depends on the quality of such models and, therefore, on the quality and size of the corpora from which they are trained. That is the reason why the acquisition of adequate corpora is a key process.

With the objective of facilitating the acquisition of a dialog corpus for the EDECAN-SPORT task for the booking of sports facilities within the framework of the EDECAN project (Lleida et al., 2006), we followed the process described below. Firstly, we analyzed human-human dialogs provided by the sports area of our university, which have the same domain defined for the EDECAN-SPORT task. From these dialogs we defined the semantics of the task in terms of dialog acts for both the user utterances and system prompts, and labeled these dialogs. Thus, we have a very low initial corpus for the EDECAN-SPORT task. From this small corpus we learned a preliminary version of the dialog manager (Griol et al., 2008). This dialog manager was used as a prototype in the supervised process of acquiring a larger corpus by means of the Wizard of Oz technique.

Secondly, as the initial corpus is not large enough to train a suitable model for the speech understanding module, we do not have a preliminary version of this module for the acquisition process with the Wizard of Oz. Our proposal is based on using a specific Wizard of Oz to play the role of the natural language understanding module and a second Wizard of Oz to supervise the dialog manager. Using these two WOz allows us to obtain after the acquisition process not only the dialog corpus, but also the dialog acts corresponding to the labeling of the user and system turns (avoiding the subsequent process of manual labeling).

2 Architecture of the acquisition

Following the main contributions in the literature in the area of spoken dialog systems, we used the Wizard of Oz technique to acquire a dialog corpus for the EDECAN-SPORT task. The main difference of our proposal (Garcia et al., 2007) consists of using two Wizards of Oz: a simulator of the speech understanding process and a supervisor of the dialog manager. The first wizard listens to the user utterances, simulates the behavior of the automatic speech recognition and speech understanding modules for recognizing and understanding speech, and provides a semantic representation of the user utterance. From that representation, the second wizard supervises the behavior of the dialog manager. Figure 1 shows the architecture defined for the acquisition.

2.1 The understanding simulator

The Wizard of Oz that deals with the understanding process generates the semantic representation of the user utterances. To achieve the most similar result to a real statistical understanding module, the representation generated by the first wizard is

[*] This work has been partially funded by Spanish MEC and FEDER under project TIN2008-06856-C05-02.

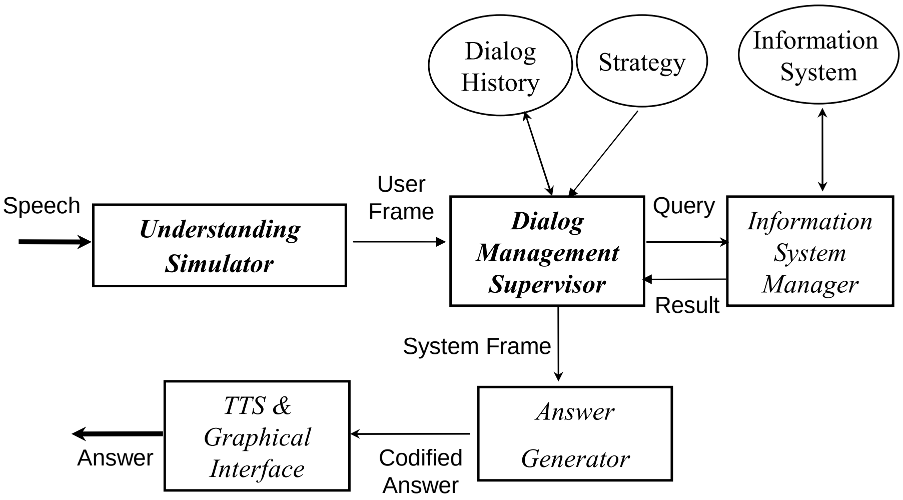


Figure 1: The proposed acquisition schema for the EDECAN corpus

passed though a module that simulates errors. This simulation (Garcia et al., 2007) is based on the analysis of the errors in the recognition and understanding processes generated when our models were applied to a corpus of similar characteristics.

2.2 The dialog manager

We have developed an approach to dialog management using a statistical model that is estimated from a dialog corpus (Griol et al., 2008). This model is automatically learned from a dialog corpus labeled in terms of dialog acts. This approach, which was originally developed to be used in a dialog system that provides train schedules and prices, was adapted for its use in the task of the booking sports facilities. This adaptation takes into account the new requirements introduced in this task, which involves using an application manager that interacts with the information servers and verifies if the user queries fulfill the regulations defined for the booking service. The actions taken by the application manager can affect the decision made by the dialog manager, aspect which was not considered in the previous task.

From the human-human corpus, a prototype of the dialog manager module was implemented to be included in our acquisition system. The second wizard supervised its behavior. This supervision is carried out by means of two applications. The first one is used to supervise the response automatically generated by the dialog manager (the wizard corrects this response when he considers that it is inadequate). The second application is used to supervise the operation of the application manager.

3 The acquisition

Using the approach described in this article, a set of 240 dialogs has been acquired for our task. A total of 18 different speakers from different origins (the headquarters of the research teams of the EDECAN consortium). The languages involved in the acquisition have been Spanish, Catalan and Basque. A set of 15 types of scenarios was defined in order to cover all the possible use cases of the task.

The information available for each dialog consists of four audio channels, the transcription of the user utterances (with an average of 5.1 user turns per dialog and 6.7 words per user turn) and the semantic labeling of the user and system turns.

References

F. Garcia, L.F. Hurtado, D. Griol, M. Castro, E. Segarra, and E. Sanchis. 2007. Recognition and Understanding Simulation for a Spoken Dialog Corpus Acquisition. In *TSD 2007*, volume 4629 of *LNAI*, pages 574–581. Springer.

D. Griol, L. F. Hurtado, E. Segarra, and E. Sanchis. 2008. A statistical approach to spoken dialog systems design and evaluation. In *Speech Communication*, volume 50, pages 666–682.

E. Lleida, E. Segarra, M. I. Torres, and J. Macías-Guarasa. 2006. EDECN: sistEma de Dilogo multidominio con adaptacin al contExto aCstico y de AplicaciN. In *IV Jornadas en Tecnologia del Habla*, pages 291–296, Zaragoza, Spain.

J. Williams and S. Young. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. In *Computer Speech and Language 21(2)*, pages 393–422.

Integrating Prosodic Modelling with Incremental Speech Recognition

Timo Baumann
Department for Linguistics
University of Potsdam
Germany
timo@ling.uni-potsdam.de

Abstract

We describe ongoing and proposed work concerning incremental prosody extraction and classification for a spoken dialogue system. The system described will be tightly integrated with the SDS's speech recogntion which also works incrementally. The proposed architecture should allow for more control over the user interaction experience, for example allowing more precise and timely end-of-utterance vs. hesitation distinction, and auditive or visual back-channel generation.

1 Introduction

Incremental Spoken Dialogue Systems start processing input immediately, while the user is still speaking. Thus they can respond more quickly after the user has finished, and can even back-channel to signal understanding. In order for this to work, all components of the SDS have to be incremental and interchange their partial results. While both incremental ASR (Baumann et al., 2009) and incremental prosody extraction (Edlund and Heldner, 2006) exist, we here describe work to join both for better processing results.

2 Related Work

Skantze and Schlangen (2009) present an incremental spoken dialogue system for a micro-domain, which uses prosody extraction for better end-of-utterance detection, reducing response time for affirmatives to 200 ms (Skantze and Schlangen, 2009). Their prosody extraction is rather crude though, and relies on the words in their number-domain being of equal length and type. We extend their work by implementing a theory-based prosody model, which should be applicable for a variety of purposes.

3 Prosody Modelling

The main prosodic features are pitch, loudness and duration. A combination of their contours over time determine whether syllables are *stressed* or not and whether there are intonational boundaries between adjacent words (Pierrehumbert, 1980). Stress and boundary information can then be used to further determine syntactic and semantic status of words and phrases.

Phonemes and their durations are directly available from ASR and syllables can either be reconstructed from a dictionary or computed on the fly.[1] Fundamental frequency and RMSE are calculated on the incoming audio stream. Prosodic features must be normalized by speaker (mostly pitch) and channel (mostly loudness), and phoneme identity from ASR may help with this. Also, we look into FFV (Laskowski et al., 2008) and advanced loudness metering (ITU-R, 2006) for robust pitch and loudness estimation, respectively.

In order to derive features per syllable, contours have to be parameterized. Both TILT (Taylor, 1998) and PaIntE (Möhler, 1998) require right context, which is unavailable in incremental processing, so their methods must be adapted.

Finally, the feature vectors for syllables and word boundaries should be reduced in dimensionality in order to be more useful for higher-level processing. It might also be possible to train classifiers for specific upcoming events. (like end-of-utterance prediction (Baumann, 2008)).

The dataflow through the module is shown in Figure 1. Output is generated for both prosody and word events. The frequency of these events can be different (e. g. several juncture measures could follow each other, indicating juncture growing as time proceeds) and filtering techniques similar to those by Baumann et al. (2009) will be used.

[1] The first approach allows predictions into the future, while the second is more flexible.

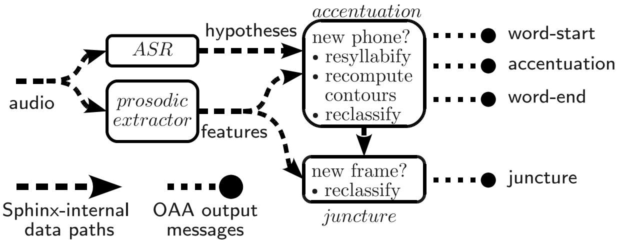


Figure 1: Dataflow diagram for the combined ASR and prosody processing.

4 A Prototype System

We construct a micro-domain (Edlund et al., 2008) exposing select problems we try to resolve with our system, and simplifying other problems that are outside of our focus.

The user's task is to order a robot hand to move (glowing) waste above a recycle-bin and to drop it there. In other words, the user controls a 1-dimensional motion and a final stop signal.

A data collection on user behaviour in this domain has been caried out in a Wizard-of-Oz setting with 12 subjects, comprising 40 minutes of audio and 1500 transcribed words.

The data shows the expected phenomena: sequences of directions ("left, left, left, ok; drop"), or use of lengthening ("leeeeft") to express distance. Marking of corrections (of purposeful misunderstandings by the wizard) using prosody, and stress on content words.

Another property of the domain are the consequences for different system actions: going right can easily be undone by going left, but dropping cannot be corrected. Thus, there are different levels of certainty that must be reached for the system to take different actions. Prosody should help in identifying confidences and finality of utterances.

5 Possible Extensions, Future Work

The model presented in Section 3 probably exceeds what would be strictly necessary for implementing the system proposed in Section 4. This is by purpose, as it allows for a basis for future extensions:

- Juncture could be calculated for all frames considered word-boundaries by the ASR and this information could be used in addition to the language model's transition probability.
- The syllable stress measure could be used in ASR rescoring to favor likely stress patterns.
- The juncture measure could be easily used in a stochastic parser.
- An obvious extension is a more complex positioning task in a 2D or 3D environment with multiple named entities in them. This would show whether the proposed system scales and introduces reference resolution problems in which prosody might be help.

Acknowledgments

This work is funded by a DFG grant in the Emmy Noether programme, and a short-term grant by the German Academic Exchange Service.

The author would like to thank David Schlangen and Jens Edlund for valuable input on the ideas and plans presented.

References

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA.

Timo Baumann. 2008. Simulating Spoken Dialogue With a Focus on Realistic Turn-Taking. In *Proceedings of the 13th ESSLLI Student Session*, Hamburg, Germany.

Jens Edlund and Mattias Heldner. 2006. /nailon/ - Software for Online Analysis of Prosody. In *Ninth International Conference on Spoken Language Processing*. ISCA.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50:630–645.

ITU-R. 2006. *ITU-R BS. 1770-1. Algorithm to measure audio programme loudness and true-peak audio level.* International Telecommunication Union.

Kornel Laskowski, Mattias Heldner, and Jens Edlund. 2008. The fundamental frequency variation spectrum. In *Proceedings of FONETIK 2008*.

Gregor Möhler. 1998. *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Ph.D. thesis, Universität Stuttgart.

Janet B. Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, MIT.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of EACL 2009*.

Paul Taylor. 1998. The TILT Intonation Model. In *Proceedings of the ICSLP 1998*, pages 1383–1386.

Author Index