Catalog ´04

Proceedings
of the Eighth Workshop
on the Semantics and Pragmatics
of Dialogue

Edited by
Jonathan Ginzburg and Enric Vallduví

Department of Translation and Philology
Universitat Pompeu Fabra, Barcelona

Catalog ´04

Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue

Jonathan Ginzburg and Enric Vallduví (editors)

19-21 July 2004
Barcelona

Organized by
Department of Translation and Philology
Universitat Pompeu Fabra, Barcelona

Catalog '04
Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue
(Barcelona, 19-21 July 2004)
Edited by Jonathan Ginzburg and Enric Vallduví

Printed in Barcelona: 2004
ISBN: 84-609-2205-7
Dipòsit Legal: B-48 755-2004

Catalog '04 is endorsed by:

SIGdial: http://www.sigdial.org
SIGSEM: http://www.sigsem.org

Foreword

Catalog '04, hosted by the Department of Translation and Philology at Universitat Pompeu Fabra in Barcelona, July 19-21, 2004, is the eighth in the SEMDIAL series of workshops. SEMDIAL Workshops bring together researchers working on the semantics and pragmatics of dialogue in fields such as artificial intelligence, formal semantics and pragmatics, computational linguistics, philosophy, and psychology. The series was founded by Gerhard Jäger and Anton Benz, who were students at the CIS, University of Munich, at the time of the first meeting, MunDial 1997. MunDial 1997 was followed by Twendial 1998 (Enschede, The Netherlands), Amstelogue 1999 (Amsterdam, The Netherlands), Götalog 2000 (Gothenburg, Sweden), Bi-Dialog 2001 (Bielefeld, Germany), Edilog 2002 (Edinburgh, Scotland), and DiaBruck 2003 (Saarbrücken, Germany).

One of the toughest tasks for the organizers of a SEMDIAL workshop is choosing a name for the event. Following a by now time-honored tradition, one has to toy around with the name **dialog** and the name of the city or area that hosts the workshop. Fortunately, however, this year someone did the thinking for us. After a few unsuccessful tosses with *Barnalog*, *Dial-a-barn*, and *Dialona*, Massimo Poesio came up with the splendid *Catalog*, an obvious yet camouflaged blend of *Cata*lonia, host country, and dia*log*. Hence, **Catalog '04**. Thanks, Massimo.

Catalog '04 received 39 submissions. Each was reviewed by at least two peers. Nineteen papers were accepted as regular contributions and a further 13 were accepted as poster presentations, ten of which will be present at Catalog '04. The 19 regular contributions and 2-page abstracts for the 10 poster presentations appear in these proceedings, along with abstracts for the four Catalog '04 keynote speakers: Robin Cooper (Göteborgs Universitet), Massimo Poesio (University of Essex), Alexander Rudnicky (Carnegie Mellon University), and Michael Tanenhaus (University of Rochester). We are extremely grateful to the keynote speakers for accepting the Catalog '04 invitation. Overall, thanks to all speakers and poster presenters, Catalog '04 succeeds in providing a good overview of the state of the art in dialog research from which we hope new powerful ideas and avenues of research evolve.

Jonathan Ginzburg (King's College, London) acted as chair of the Program Committee for Catalog '04 and oversaw the processes having to do with calls for papers, paper submission and paper reviewing. His role was truly vital through the early stages of the Catalog '04 preparations. The Program Comittee included Anton Benz (Syddansk Universitet), Johan Bos (University of Edinburgh), Justine Cassell (Northwestern University), Lawrence Cavedon (CSLI, Stanford), Robin Cooper (Göteborgs Universitet), Paul Dekker (University of Amsterdam), Claire Gardent (CNRS, Loria), Simon Garrod (University of Glasgow), Jonathan Ginzburg (King's College, London, Chair), Pat Healey (Queen Mary, University of London), Ivana Kruijff-Korbayová (Universität des Saarlandes), Staffan Larsson (Göteborgs Universitet), Colin Matheson (University of Edinburgh), David Milward (Linguamatics, Cambridge), Massimo Poesio (University of Essex), Hannes Rieser (Universität Bielefeld), and David Traum (University of Southern California). Many thanks to all of them for an excellent job.

Putting together an event like Catalog '04 entails worrying about lots of little details you never imagined you would have to worry about some day. Local committee members Stefan Bott, Judith Domingo, Laia Mayol, Stella Puig-Waldmüller, and Ana Ruggia have helped out with these. Also, Susi Bolós, head secretary of the Department of Translation and Philology, has provided a great deal of help in sorting out all kinds of logistic issues. Catalog '04 has been made possible by financial support from the Ministry of Universities of the Catalan Government (2004 ARCS1 00056), the Ministry of Education of the Spanish Government, and the Department of Translation and Philology and the School of Translation and Interpreting at Universitat Pompeu Fabra.

Enric Vallduví
Catalog ´04 Local Committee chair

Programme committee and referees

Jonathan Ginzburg (chair)	[King's College, London]
Anton Benz	[Syddansk Universitet]
Johan Bos	[University of Edinburgh]
Justine Cassell	[Northwestern University]
Lawrence Cavedon	[CSLI, Stanford]
Robin Cooper	[Göteborgs Universitet]
Paul Dekker	[Universiteit van Amsterdam]
Claire Gardent	[CNRS, Loria]
Simon Garrod	[University of Glasgow]
Pat Healey	[Queen Mary, University of London]
Ivana Kruijff-Korbayová	[Universität des Saarlandes]
Staffan Larsson	[Göteborgs Universitet]
Colin Matheson	[University of Edinburgh]
David Milward	[Linguamatics, Cambridge]
Massimo Poesio	[University of Essex]
Hannes Rieser	[Universität Bielefeld]
David Traum	[University of Southern California]

Invited Speakers

Robin Cooper	[Göteborgs Universitet]
Massimo Poesio	[University of Essex]
Alex Rudnicky	[Carnegie Mellon University]
Michael Tanenhaus	[University of Rochester]

Local Committee [Universitat Pompeu Fabra]

Enric Vallduví (chair)
Stefan Bott
Judith Domingo
Laia Mayol
Stella Puig-Waldmüller
Ana Ruggia

Table of Contents

Invited Talks

Papers

Poster Abstracts

Type theoretic approach to information state update in issue based dialogue management

Robin Cooper
Department of Linguistics
Göteborgs Universitet
Box 200, S-405 30 Göteborg, Sweden
cooper@ling.gu.se

For several years the research group at our Dialogue Systems Lab has been involved in the development of the information state update approach to the building of dialogue systems and in particular Issue based dialogue management developed in Staffan Larsson's PhD thesis and based on Jonathan Ginzburg's gameboard approach to dialogue, focussing on the notion of questions (or issues) under discussion.

Larsson's computational approach to information state updates involves a large collection of update rules which fire when certain conditions in the information state are met in a regime determined by a general control algorithm. An utterance by a dialogue participant will in general unleash a whole chain of such updates and part of the power of the approach lies in the fact that we can define very general update rules which have small effects on the information state and which are not necessarily linked to any particular form of utterance. It gives us a much finer grain on update rules than thinking in terms of single monolithic updates associated with speaker utterances.

Larsson's formulation of update rules is based on a Prolog implementation and exploits some aspects of Prolog: logic programming variables, backtracking to deal with non determinism, ordering of update actions within an update rule. In this paper we will show how notions of record and record type in type theory can be used to formulate update rules without relying on these aspects of Prolog. This allows us to give an abstract characterisation of update rules independent of programming language which points to a general theory of updates as well as the possibility of implementation in any programming language. The tools we are using can also be used in an account of compositional semantics and this points to the possibility of an integrated formal theory of information state update and compositional semantics.

This work is related to other work on computational approaches to information state update by Johan Bos on the Dipper system and to type theoretical approaches to semantics and dialogue which exploit the notion of context in type theory, e.g. work by Ranta, Ahn et al. and Piwek.

Completions and Continuations in Dialogue: Preliminary Observations

Massimo Poesio, University of Essex, United Kingdom
(Joint work with Hannes Rieser, University of Bielefeld, Germany)

I will report about work in progress on *completions* and *continuations*, two fundamental strategies for agents' alignment in dialog. (An example of completion is 1.2 in the following example.)

1.1	Inst	So, jetzt nimmst Du
		OK, now you take
1.2	Const	eine Schraube
		a screw
1.2	Inst	eine orangene mit einem Schlitz
		an orange one, with a slit

I'll start by reviewing the characteristics of completions and continuations occurring in the Bielefeld Toy Airplane Corpus (Skuplik 1999, Rieser and Skuplik 2000). I'll then discuss an 'intentional' account of how completions and continuations may be produced, building upon work by Clark (1996), Bratman (1993), Tuomela (2000), and Grosz and Kraus (1996). In the next part of the talk, I will propose an analysis of completions and continuations in an intentionally based version of the PTT framework (Poesio and Traum 1997, 1998, Matheson et al. 2000). Finally, I will consider a non-intentional explanation of completions, taking up Pickering and Garrod's suggestions concerning dialogue description (Pickering and Garrod 2003).

Learning to talk by listening

Alex Rudnicky
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
air@cs.cmu.edu

Real-time studies of comprehension and production in dialogue: Insights from eye movements

Michael K. Tanenhaus and Sarah Brown-Schmidt
Meliora 420
Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627-0268, USA
{mtan,sschmidt}@bcs.rochester.edu

Much of what we know about the cognitive processes by which speakers generate, and listeners comprehend, utterances come from on-line studies that measure real-time processing. However, the experimental methods used to generate these data are difficult to apply to natural interactive dialogue. In recent work, we have been examining the feasibility of using eye movements to study task-oriented dialog in variations of referential communication tasks. In this talk, I will first outline some reasons for why it is important to pursue such studies, and why monitoring eye movements is a promising approach. I will then summarize results from two lines of investigation. The first line of work uses a referential communication task that examines how referential domains of interlocutors align along task-relevant dimensions, allowing referential expressions to be linguistically underspecified and reducing competition from alternative potential referents. The second line of work investigates the eye movements of speakers as they plan and generate referential expressions in domains where modification might or might not be necessary. The timing of looks to potential contrast members predicts whether a referring expression will be produced fluently or not, and whether modification will be pre or post-nominal. Finally, I will briefly discuss "in progress" work that aims to determine if and when interlocutors monitor each other's likely intentions and knowledge when planning and interpreting utterances.

Contextual Reasoning in Multimodal Dialogue Systems: Two Case Studies

Johan Boye, Mats Wirén and Joakim Gustafson
Voice Technologies
TeliaSonera Sweden
{Johan.Boye|Mats.Wiren|Joakim.Gustafson}@teliasonera.com

Abstract

This paper describes an approach to contextual reasoning for interpretation of spoken multimodal dialogue. The approach is based on combining recency-based search for antecedents with an object-oriented domain representation in such a way that the search is highly constrained by the type information of the antecedents. By furthermore representing candidate antecedents from the dialogue history and visual context in a uniform way, a single machinery (based on β-reduction in lambda calculus) can be used for resolving many kinds of underspecified utterances. The approach has been implemented in two highly different domains.

1 Introduction

This paper describes an approach to contextual reasoning and its application to two radically different domains, both of which make use of spoken multimodal dialogue. The first system is ADAPT, which allows the user to look for apartments for sale in central Stockholm (Bell et al. 2001). Apartments are represented in a relational database and are displayed as icons on an interactive map. The second system is the NICE fairytale game, in which the user collaborates with an animated character to solve a problem in an immersive 3D world (Gustafson et al. 2004a).

Both domains are, each in its own way, sufficiently restricted that no serious problems are posed by lexical or structural ambiguity. Likewise, the use of quantification is limited and only rarely leads to ambiguity problems. In contrast, interaction in both domains abound with deictic and anaphoric expressions such as pronouns, definite descriptions and ellipses. These expressions refer to things in the visual surrounding as well as to objects that have been mentioned in the previous dialogue. Thus, all utterances have to be interpreted by way of reasoning about the objects in the combined dialogue and visual context.

Although naïve reference resolution methods — such as preferring the most recent grammatically compatible antecedent — may perform remarkably well (see Hobbs 1978, Mitkov 1998), dialogue applications typically must bring more knowledge into play in order to perform well. Often some logic-based reasoning using a representation of the task and domain is adopted. An early example of this is the focus representation of Grosz (1977), based on the partitioned semantic networks of Hendrix (1975). Another example is the resolution component of the CLE (Alshawi 1992). However, there is a computational price to be paid for the general semantic reasoning-based methods. Particularly for limited domains, it seems that it would be useful to find a less complex approach.

The approach described here replaces general-purpose semantic reasoning with a much more restricted type system. By combining this with the recency principle — thus looking for the most recently referred object of a compatible *type* — we obtain a highly efficient but still accurate search strategy. This approach works well in two fairly complex domains to be further described below. Moreover, by representing candidate antecedents from the dialogue history and visual

context in a uniform way, the same machinery can be used for resolving many kinds of underspecified utterances.

2 Systems and domains

2.1 The ADAPT system

The graphical user interface of the apartment database system ADAPT is shown in Figure 1. The system features an animated talking agent providing the user with information about apartments currently for sale. The system also displays the geographic locations of apartments under discussion as clickable icons with distinct colours on a map. Graphical input and the textual result from the speech recognizer are jointly interpreted by the system.

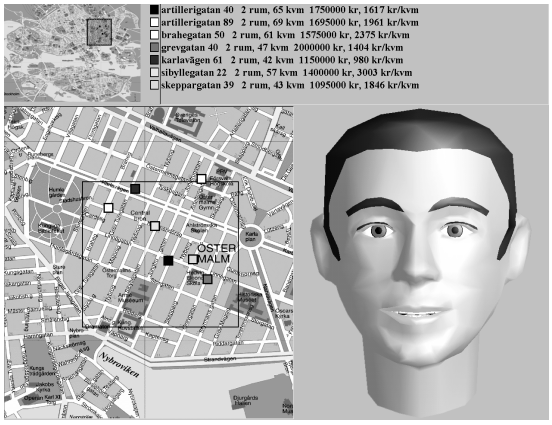


Figure 1. The graphical interface of the ADAPT system.

The following is a typical dialogue fragment with the ADAPT system.

A1. **User:** Are there any two-room apartments on the South Side that cost less than 2 million?
A2. **System:** Do you have any other requirements?
A3. **User:** A balcony would be nice.
A4. **System:** [Displays colored icons on the map] There are seven such apartments. They are shown on the map.
A5. **User:** I see… The green apartment… how much does it cost?
A6. **System:** The green apartment costs 1.5 million.
A7. **User:** What about the red one?
A8. **System:** The red apartment costs 1.9 million.
A9. **User:** Okay… Is there anything cheaper?

Note that even this short dialogue contains a fair variety of anaphora: pronouns, definite descriptions (often using the distinct colours of the displayed icons), elliptic questions (utterance A7) and elliptic answers to questions (A3). Also, graphical pointing actions (of which there are no instances in the particular dialogue here) are often accompanied by deictic noun phrases.

2.2 The NICE fairy-tale game

The NICE fairy-tale game system lets the user enter into an animated 3D world and communicate with a number of animated characters by means of spoken dialogue. So far, the first scene of the game has been implemented, featuring a single character, called Cloddy Hans (inspired by one of the characters of an H C Andersen fairytale). This first scene takes place in Andersen's authoring laboratory, where the key device is a fairytale machine. The machine has a magic quality, namely, that if one inserts objects into its designated slots (for "magic things", "dangerous things", and so on) and pulls a lever, it creates a new fairytale in which the objects come to life. Since Cloddy Hans would like to create a fairytale of his own, he needs help from the user in understanding where to put different things. The first scene therefore develops into a collaborative "put that there" task, where Cloddy Hans and the user have to agree on what to do. Figure 2 shows a screen shot from the game.

Figure 2. A screen shot from the NICE system. Cloddy Hans is standing in front of the fairy-tale machine, next to the shelf with objects.

The following fragment is from a dialogue with an eleven-year old user.

N1. **User:** I want you to go to the shelf.
N2. **Cloddy Hans**: Okay. [Walks over to the shelf.] What do you want me to do now?
N3. **User:** I want you to pick up the bag.
N4. **Cloddy Hans:** Do you want me to pick up the money sachet?
N5. **User:** Yes, pick up the sachet. [Clicks on the money sachet.]
N6. **Cloddy Hans:** [Picks up the sachet.] That went well!
N7. **User:** Then I want you to go to the slots.
N8. **Cloddy Hans:** Okay. [Walks over to the machine.]
N9. **User:** Now I want you to put the money sachet in the farthest slot.

Both systems allow users to refer to objects graphically by a mouse-compatible device. However, while users of ADAPT can scroll the map as they please, users of the NICE system cannot manipulate the camera at will. Rather, the camera follows Cloddy Hans as he moves around, and hence situations will arise where graphical reference is not possible.

2.3 Corpora

The ADAPT corpus is based on one Wizard-of-Oz collection with 16 subjects (Bell et al 2001), and one independent data collection with 24 subjects using the fully functional system (Edlund and Nordstrand 2002). The user tasks were to find an apartment obeying certain constraints (in the former case) and to find ones that the subject might want to live in (in the latter case).

The current NICE corpus is based on a semi-automated Wizard-of-Oz collection with 10 children aged 11–15 (Gustafson et al. 2004). The subjects were informed about the scenario described in the previous section and were instructed to collaborate with Cloddy Hans to put some (unspecified) things into the machine. (We have not yet collected any data using the existing, fully functional system corresponding to this scenario.)

All corpora examples in this paper have been translated from Swedish to English by the authors.

3 Referential phenomena

3.1 Knowledge sources

One way of studying referring expressions for the purpose of developing focus management and reference resolution is to look at what knowledge sources are needed to interpret them.

Perhaps the most obvious knowledge source in a graphics-based multimodal system is the *visual context*. Two examples of this are utterances A5 (”the green apartment”) and N9 (”the slot furthest away”) in the dialogue fragments in Section 2. As can be seen, definite descriptions include both visually salient properties and (in NICE) the relative position of 3D objects. Descriptions of the latter often include complex ordinal and directional expressions, like “the third tube from the left” or “the hole which is second from the right”. (Whereas currently each object in the shelf is unique, the four slots in the machine have to be distinguished by means of some other property.)

Another obvious knowledge source is the *preceding dialogue*, for example, utterance A9 (”anything cheaper”). Here the user expresses a desire which refers to the price of a previous apartment (presumably the green one).

Sometimes a record of *past events* is also needed to resolve a reference, as shown by the following example from the NICE corpus:

N10. **User:** Where we put the magic wand… there you can put it.

Here the clause ”Where we put the magic wand” is referring to a slot of the fairytale machine via its relation to a previous action.

Finally, a model of the *domain* is needed, as shown in the following example:

N11. **User:** I want you to take the hammer.
N12. **Cloddy Hans**: Okay. [Takes the hammer.]
N13. **User:** Then I want you to go to the machine… And put it in the first tube.

Here, it is obvious that “it” in utterance N13 corresponds to the hammer because of the way the particular objects and actions are related in this domain. However, a naive recency-based

model without this information would rather associate ”it” with the machine.[1]

Summing up, all of the above knowledge sources frequently come into play in ADAPT and NICE, with the exception of past events that are only rarely used.

3.2 Referential usage

A problem which is complementary to the one above is how referential expressions are constructed depending on which knowledge source is involved. In particular, how do people refer to objects that are present in the visual display but that have not yet been referred to in the dialogue? Also, to what extent are objects *outside* of the current visual display referred to? This is important for the purpose of determining how the current focus should be updated with respect to objects from the visual environment.

As for the first question, people frequently use definite descriptions to refer to visually displayed objects right from the first turn of the dialogues, without the objects ever having been mentioned. Both in ADAPT and NICE, there is a variety of characteristic properties that can be combined to describe objects — for example, in ADAPT the colour of the icon, the number of rooms and the street of the apartment, etc.

Pronouns are sometimes used without previous mentioning of the referred object in the dialogue, but then only in combination with a graphical pointing action:

N14. **User:** Go to the shelf.
[Cloddy Hans confirms and walks up to the shelf.]
N15. **User:** [Graphical pointing at diamond.] Take it.

As for objects outside of the visual display (the second question above), and looking first at ADAPT, it is clear that the set of apartment icons provides an extremely strong cue for the mutually grounded context: Although users frequently change their desired apartment constraints back and forth as they explore the search space, there is no instance of a user going back and referring verbally to a particular apartment that is no longer displayed on the map. Thus, in our data the objects under discussion are always those that are shown on the map. Similarly, there are no references to previous events (“Go back to the area where we were previously”).

In NICE, the situation is different because of the moving camera and the fact that the set of objects remains constant except when something is put into the machine. Here, users do refer to things currently outside of the visual display, like the fairytale machine and the shelf. Even objects that are no longer physically present in the scene may be referred to, as in utterance N10 above.

4 Contextual interpretation

4.1 The problem

The problem of contextual interpretation can be divided into three subproblems. First, expressions that refer to the context must be recognized in the input. For spoken input, this is not trivial, since state-of-the-art speech recognizers often fail in recognizing short (function) words, such as pronouns. Secondly, there is the issue of finding the set of candidate objects on which the interpretation of the input can be based — that is, computing the right context in which to interpret the utterance. We call this focus management. Thirdly, there is the issue of combining the contextual information with the information conveyed in the utterance to produce the final interpretation. It is well-known that the two last steps can be arbitrarily difficult (see e.g. Hobbs 1978).

As for the first subproblem, we have shown in a previous paper how spoken input exhibiting a large amount of anaphoric and deictic expressions can be efficiently parsed in a limited domain (Boye and Wirén 2003a). As for the third subproblem, our semantic representation is designed so as to let all contextual interpretation be realized by a uniform process of β-reduction in lambda calculus. This representation is described in detail in Section 5. The rest of this section deals mainly with the second subproblem: how to determine the set of objects that, at each moment, constitute the possible targets for interpretation of referring expressions. We call such objects *salient* objects.

[1] In Swedish, “hammer” and “machine” have identical gender, and hence the pronoun agrees grammatically with both of them.

4.2 Apartment domain

To begin with, the ADAPT system must distinguish internally between *intensional* and *extensional* objects. Whenever the user starts over by giving new constraints (as in utterance A1), a new intensional object is created. This object is considered to be salient until a set of concrete apartments is presented to the user (as in utterance A4). These apartments are represented internally as extensional objects and are considered salient as long as their icons are displayed on the map (i.e. until the user has asked for a new set of apartments). Then a new intensional object is created, and the whole cycle is repeated. In general, this kind of distinction must be made by any system in which the dialogue begins by specifying an "ideal" object before matching it with real ones.

As for displayed objects and their relation to the current context, it turns out that the basic mechanism for updating the set of salient object can be made very simple: As discussed in Section 3.2, the set of displayed apartment icons provides both necessary and sufficient information to determine the set of salient apartments. This approach is the same as the one taken in Cheyer and Julia (1995).

To handle implicitly focused items, objects of the domain are represented by means of a type hierarchy, motivated by the characteristics of the domain (this is further described in Section 5). This allows us to handle utterances like A12 below:

A10. **User:** How many rooms does the green apartment have?
A11. **System**: Three rooms.
A12. **User:** What is the monthly fee?

Here, "monthly fee" will be associated with an attribute of the relevant apartment object.

In many cases, referential expressions in the ADAPT domain turn out to be unambiguous (as in utterances A5 and A7 in Section 2.1). In those cases which remain ambiguous, a straightforward recency principle works in the vast majority of cases (that is, preferring type-compatible antecedents that appear at shorter linear distance backwards in the dialogue).

In some easily distinguishable special cases, other rules apply. An example of this is utterance A9, where the desired price should be less than all the previously discussed prices of the apartments in focus.

4.3 Fairy-tale game domain

Looking at the introductory fairy-tale scenario described in Section 2.2, our data so far indicate that it is sufficiently restricted to be amenable to the same basic methods as those used in ADAPT. First, the set of objects that can be referred to is limited and can be kept constant from the point of view of the visual context in which the user's utterance is to be understood. (Although objects disappear from the physical environment when they are put into the machine, they may still be referred to as exemplified by N10 above.) Secondly, the limited amount of moving of the camera also does not require any corresponding shifting of the visual context, as discussed in Section 2.2.

For these reasons (and in contrast to ADAPT), we do not make use of any mechanism for updating the visual context, but rather keep all objects from the scene constantly in the current context.[2] This includes the objects initially situated in the shelf, the shelf itself, the fairy-tale machine as well as relevant parts and properties of these, like the slots of the machine and the symbolic labellings of each slot.

Clearly, however, this simple strategy will not be tenable in the succeeding scenes of the game (currently under implementation). Here, the changing scenes will require updating of the visual context as the user freely moves about in the large 3D world. We will return to this scenario in a later paper.

5 Representation and implementation

As mentioned above, objects in both the ADAPT and NICE domains are represented by means of a type hierarchy in a standard object-oriented fashion, much the same way one would represent the domain in the Java™ programming language.

[2] This seems to be the approach taken also by Lemon et al (2001), whose system does not use an explicit internal representation of the visual context.

Specifically, this means:

- Every object belongs to exactly one type.
- A type may be a direct subtype of exactly one type.
- An object may have any number of attributes, whose values are objects of the appropriate types.[3]

For instance, in the fairy-tale system, objects that can be moved about belong to the type thing. Things have an attribute position whose values should belong to the type location. So, supposing that hammer and axe are things, and onShelf is a location, the fact that the hammer is lying on the shelf is representable, whereas the fact that the hammer is lying on the axe is not (since the equation hammer.position = onShelf obeys the type constraints whereas hammer.position=axe does not).

A slot (in the fairy-tale machine) is a special kind of location; hence the type slot is a subtype of location. This means that hammer.position can be given values also of type slot.

This object-oriented approach to coding the domain extends also to actions, events, dialogue acts, and so on. For instance, the action of picking up something is represented by an object of type pickUp having two attributes; agent of type (fairy-tale) character, and patient of type thing.

5.1 Representation of user utterances

During execution, user utterances are translated by a parser into typed combinators[4] over the domain model (see further Boye and Wirén 2003a, 2003b). In the NICE system, utterances are translated into expressions of type dialogue_act (request, ask, tell, and so on). As an example, utterance N3 would be translated into

$$\text{request(user, cloddy, pickUp(cloddy, bag))}$$

whereas the utterance "Pick it up" would be translated into

$$\lambda x^{\text{thing}}.\text{request(user, cloddy, pickUp(cloddy, x))}$$

Here, superscripts indicate the types of variables. Thus, the expression above denotes a function taking a thing as the argument, returning the fully instantiated request as the result. Here the domain model is used to infer that the missing information (the object x being picked up) is of type thing.

Resolution of the reference "it" now corresponds to applying[5] the function above to an expression of the appropriate type, e.g.

$$(\lambda x^{\text{thing}}.\text{request(user, cloddy, pickUp(cloddy, x))}\ \text{bag}) \rightarrow \text{request(user, cloddy, pickUp(cloddy, bag))}$$

Thus, the type constraints in the domain model help ruling out undesired interpretations of references in user utterances.

The ADAPT apartment system seeks to translate all user utterances into the form $?x^t(P)$, which can be paraphrased as "Give me x of type t such that P is true". Again, lambda abstractions are used to represent missing information. For instance, utterance A5 would be:

$$\lambda x^{\text{apartment}}\ ?p^{\text{money}}\ (x.\text{price} = p\ \&\ x.\text{color} = \text{green})$$

Supposing that apt1 denotes the apartment the user is referring to in utterance 5, then contextual interpretation of this utterance amounts to applying the functional expression to apt1:

$$(\lambda x^{\text{apartment}}\ ?p^{\text{money}}\ (x.\text{price} = p\ \&\ x.\text{color} = \text{green})\ \text{apt1}) \rightarrow ?p^{\text{money}}\ (\text{apt1.price} = p\ \&\ \text{apt1.color} = \text{green})$$

The resulting expression is paraphrased ``Give me the price of apt1", and can be translated straightforwardly into a database search command.

One of the nice features of this representation scheme is that various kinds of anaphora and ellipses can be handled the same way. For example, in utterance A7 it is evident that the user wants to know something about the red apartment, but it is not clear (before consulting the context) exactly what he wants to know. Such

[3] This representation scheme is thus significantly less expressive than e.g. the *partitioned semantic networks* by Hendrix (1975) (used by Grosz, 1977) , which are equivalent to first-order logic.

[4] A *combinator* is a lambda-expression without free variables (see e.g. Hindley and Seldin 1986). For an approach to natural language semantics based on combinators, see Jacobson (1999).

[5] Application of the lambda expression f to the argument a, so-called *β-reduction,* is denoted (f a). Another commonly used notation is f@a.

elliptic utterances are represented using higher-order lambda expressions:

$$\lambda x^{apartment}\ \lambda f^{apartment \rightarrow dialogue\_act}\ (f\ x[x.color = red])$$

Here, x[x.color=red] is a constrained variable, i.e. x can only take values such that x.color=red is true. In this case, contextual interpretation amounts to first applying the above expression to the appropriate apartment (whose icon should be red), and then applying the resulting expression to a function f, expressing what to do with the red apartment. Supposing apt2 denotes the red apartment the user is referring to, then first applying the above expression to apt2 yields:

$$\lambda f^{apartment \rightarrow dialogue\_act}\ (f\ apt2)$$

We will discuss how to find functional antecedents in section 5.2; for now we will just stipulate that the correct antecedent is

$$\lambda y^{apartment}\ ?p^{money}\ (y.price = p)$$

since

$$(\lambda f^{apartment \rightarrow dialogue\_act}\ (f\ apt2)\ \ \lambda y^{apartment}\ ?p^{money}\ (y.price = p)) \rightarrow$$

$$(\lambda y^{apartment}\ ?p^{money}\ (y.price = p))\ \ apt2) \rightarrow$$

$$?p^{money}\ (apt2.price = p)$$

i.e. "How much does apt2 cost?".

5.2 Focus management

To keep track of which objects are potential targets for reference resolution, the ADAPT and NICE systems use several internal data structures.

The *visual context history* is a recency-ordered list of sets of objects, each set corresponding to a visual context. In ADAPT, each set consists of apartments whose icons are shown simultaneously on the map. Each time some icons are added or removed, a new visual context is created and added to the history. The visual context is used for resolving definite NPs like "the green apartment", and metonymies like "King's street".

As previously mentioned, the visual context is kept constant in the first scene of the fairy-tale system.

The *dialogue history* is a recency-ordered list of typed combinators, each combinator representing a (resolved) user utterance or a system utterance. The dialogue history is mostly used to resolve pronouns and ellipses, by searching backwards in the list for a (sub-)expression of compatible type. For instance, consider utterance N13, which is represented as two dialogue acts:

request(user, cloddy, goTo(cloddy, atMachine))

$$\lambda x^{thing}.request(user, cloddy, putDown(cloddy,x,magicSlot))$$

In order to find an argument of type thing, we have to go back to the representation of N11:

request(user, cloddy, pickUp(cloddy, hammer))

Here, the expression hammer is of type thing, and is indeed the expression needed to correctly resolve the reference in N13. Thus, the typing of expressions prevents unwanted resolutions (like resolving "it" by "the machine" in N13).

Resolution of certain kinds of ellipses involves finding a function of the appropriate type. To resolve utterance A7 (as discussed in section 5.1) the system must find a function of type apartment→dialogue_act. This is computed by a technique reminiscent of Dalrymple et al (1991). First abstraction (reverse functional application) from the resolved representation of the preceding user utterance A5 gives us[6]:

$$?p^{money}\ (apt1.price = p\ \&\ apt1.color = green) \rightarrow^{-1}$$

$$(\lambda y^{apartment}\ ?p^{money}\ (y.price = p\ \&\ apt1.color = green)\ \ apt1)$$

By removing the redundant constraint apt1.color = green, we can extract the combinator needed to resolve A7, namely

$$\lambda y^{apartment}\ ?p^{money}\ (y.price = p)$$

[6] In this particular case, there are several possible abstractions; any one or both occurrences of apt1 can be replaced by the variable y. The particular abstraction shown here is preferred by the system, since it does not create a constraint y.color = green. Such a constraint would be inconsistent with the representation of utterance A7; "What about the red one?".

Finally, an *event history* will be added to the NICE system to be able to resolve references like the one in utterance N10.

6 Discussion

This paper describes an approach to contextual reasoning for the interpretation of spoken multimodal dialogue which refrains from general-purpose reasoning and instead uses a much more restricted type system. By combining type information with a recency principle, we obtain a search strategy which is both highly efficient and accurate. By driving interpretation with respect to both the dialogue history and visual context by a process of β-reduction, we obtain a single, uniform machinery for contextual interpretation which is applicable to the resolution of many kinds of underspecified utterances, such as deictical expressions, anaphora and ellipses. Put differently, the search strategy amounts to finding correct arguments for the typed combinators representing user utterances.

Current work is mainly directed towards extending the NICE system to include the subsequent scenario taking place in the virtual fairy-tale world. To this end, the approach described here will have to be generalized. In particular, the visual context will require frequent updating as the user freely moves around in the large 3D world instead of being confined to a single room. There might also be a need for keeping track of the visual context at the time of previous utterances in order to correctly determine antecedents. We expect to report more on these aspects as part of future work.

References

Alshawi, H. (1992) *The Core Language Engine.* The MIT Press.

Bell, L., Boye, J. and Gustafson, J. (2001) Real-time handling of fragmented utterances. *Proc. NAACL workshop on adaptation in spoken dialogue systems.*

Boye, J. and Wirén, M. (2003a) Robust parsing of utterances in negotiative dialogue. *Proc. Eurospeech.*

Boye J. and Wirén, M. (2003b) Negotative spoken-dialogue interfaces to databases. *Proc. Diabruck, 7th workshop on the pragmatics and semantics of dialogue.*

Cheyer, A. and Julia, L. (1995) Multimodal Maps: An Agent-based Approach. *International Conference on Cooperative Multimodal Communication (CMC/95)*, 24–26 May 1995, Eindhoven, The Netherlands.

Dalrymple, M., Shieber, S. and Pereira, F. (1991) Ellipsis and higher-order unification. *Linguistics and Philosophy*, vol. 14, no. 4, pp. 399–452.

Edlund, J. and Nordstrand, M. (2002) Turn-taking Gestures and Hour-Glasses in a Multi-modal Dialogue System. *ISCA workshop on Multimodal Dialogue in Mobile Environments*, Kloster Irsee.

Grosz, B. (1977): The representation and use of focus in a system for understanding dialogs. *Proc. IJCAI*, pp. 67–76.

Gustafson, J., Bell, L., Boye, J., Lindström, A. and Wirén, M. (2004a) The NICE fairy-tale game system. *Proc. SIGDIAL*.

Gustafson, J., Boye, J., Bell, L., Wirén, M., Martin, J.-C., Buisine, S. and Abrilian, S. (2004 b). Collection and analysis of multimodal speech and gesture data in an edutainment application. NICE Deliverable D2.2b. http://www.niceproject.com.

Hendrix, G. (1975) Expanding the utility of semantic networks through partitioning. *Proc. IJCAI*, pp. 115–121.

Hindley, R. and Seldin, J. (1986) *Introduction to combinators and λ-calculus*. Cambridge University Press.

Hobbs, J. (1978) Resolving pronoun references, *Lingua*, vol. 44, pp. 311–338.

Jacobson, P. (1999) Towards a variable-free semantics. *Linguistics and Philosophy* 22, pp. 117–184.

Lemon, O., Bracy, A., Gruenstein A., and Peters, S. (2001) Information states in a multi-modal dialogue system for human-robot conversation. *Proc. Bi-Dialog, 5th workshop on formal semantics and pragmatics of dialogue* , pp. 57–67.

Mitkov, R. (1998) Robust pronoun resolution with limited knowledge. *Proc. COLING/ACL'98*, pp. 869–875.

Dynamic Optimisation of Information Enrichment in Dialogue

Stina Ericsson
Department of Linguistics
Göteborg University
Box 200
40530 Göteborg, Sweden
stinae@ling.gu.se

Abstract

Information enrichment is a process whereby explicitly realised information elements in a dialogue message make use of other information elements that are accessible through the context. I introduce information enriched constituents using four information structure primitives: *ground*, *focus*, *prominent element*, and *non-prominent material*. Five Optimality Theoretic (OT) constraints are then formulated for the generation of information enriched constituents in a dialogue system, and I show how dynamic constraint reranking is needed in a dialogue system. The usefulness of bidirectionality is also shown, and I end with a discussion of computational considerations.

1 Introduction

Information enrichment is the exploration of how some information elements in a dialogue message are explicitly realised as part of an utterance, and how others, the enriching elements, are accessible through the context.

Optimality Theory (OT), (Prince and Smolensky, 1993), has in recent years been applied to semantics and pragmatics, (e.g., (Hendriks and de Hoop, 2001; Buchwald et al., 2002; Zeevat, 2001; Beaver, 2004)). Below, OT is explored for information enrichment, and addresses the question: what light can OT shed on the generation of information enriched constituents in a dialogue system? The investigation leads to dynamic rerankings of a fixed set of constraints.

The paper focuses on the question of which information elements are to be realised and which ones not. The final realisation of an information enriched constituent includes considering morphosyntactic constraints, for instance, but these are not part of present considerations.

The following section elucidates the concept of information enrichment with the help of information structure. Section 3 presents the constraints, and section 4 the constraint rankings. 5 discusses the determination of which ranking to use, and 6 bidirectionality in the context of information enrichment. Finally, section 7 considers computational issues.

2 Information Structure and Information Enrichment

In what follows, utterances will be seen as connecting to the context along two dimensions. The first is illustrated by B' in the context of A in (1). '*Jones*' is the informative part that is meant to update the current information state, whereas '*my last name is*' is an anchor to what has already been established in the dialogue. I will call the former *focus* (F) and the latter *ground* (GR), following, e.g.,Vallduví (1992) and Ginzburg (1999).

(1) A: ok and what's your last name?[1]

[1] A and B are taken from the Amex Travel Agent Data, http://www.ai.sri.com/%7Ecommunic/amex/amex.html. B'

B: ah Jones

B′: my last name is Jones

The other dimension along which utterances connect to the context is illustrated by example (2).[2] Simplifying somewhat by leaving '*about*' out of the discussion, in $F3''$ the ground is something like '*I am ... from the left-hand-side of the page just now*', and the focus '*two inches*'. The element '*two*' in the focus is an alternative – in roughly the sense of alternative-evoking focus in (Rooth, 1996) – to '*four*' in $F1$, and will here be called a *prominent element* (P). Hence, prominence is here a semantic notion, used to mark constrastive or otherwise important material within the focus. The element '*inches*' in $F3''$ is non-prominent material (NON-P) within the focus.

For accounts that use the term focus in a way reminding of prominent element as introduced here, see (Pulman, 1997; Steedman, 2000).

Having introduced the relevant information elements, I define an *information enriched constituent*, or utterance, as one whose content in a shared context, *the contextual content*, is the result of embedding its *compositional content* in a larger semantic structure. For the purposes of the current study, the compositional content of an information enriched constituent consists of a single focus, or a single prominent element, or a prominent element together with ground, with the other information elements being supplied by the context. A non-information enriched constituent consists of a full ground and focus.

Examples of information enriched constituents in (1) and (2) – again ignoring '*about*' – are B (a focus), $F1$ (a focus), $G3$ (a focus consisting of a prominent element and some non-prominent material), $F3$ (a prominent element), and $F3'$ (a focus consisting of a prominent element and non-prominent material).

The information enrichment approach covers partly the same dialogue phenomena as do approaches to what is variously called ellipsis, fragments, non-sententials, etc. (e.g, (Ginzburg, 1999; Schlangen, 2003)).

3 The constraints

Various considerations govern the amount of information – in terms of which information elements – a given utterance is to contain. In an OT setting, some of these considerations are encoded as constraints, and some of them in the ranking of these constraints. For the information elements introduced above and concerning information enrichment, five constraints are involved (not given in rank order):

FOCUS: Generate focus
***NON-PROM**: Avoid non-prominent material
GROUND: Generate ground
***GROUND**: Avoid ground material
PROM ELEM: Produce the prominent element

GROUND is a faithfulness constraint conveying that what is part of the input should also be part of the output. The mirror constraint *GROUND is instead a markedness constraint prescribing economy and simplicity.[3]

All five constraints reflect that the optimisation of utterances in terms of their desired degree of reliance on information enrichment, is a matter of balancing between markedness and faithfulness constraints, between dialogue economy and explicitness. The approach to discourse anaphora by Buchwald et al. (2002) involves similar considerations for noun phrases and the salience of referents.

4 Dynamic reranking of constraints

The generation component of a dialogue system can take a number of issues into account for determining the level of reliance on information enrichment in an utterance to be generated:[4]

• High speech recognition scores, rely on information enrichment; low speech recognition scores, rely less on information enrichment, or not at all

is a constructed utterance. 'Jones' in utterance B has been substituted for the anonymised name 'C' in the Amex transcipt to make the example more readable here.

[2] $G1$-$F3$ are from the HCRC Map Task corpus, http://www.hcrc.ed.ac.uk/dialogue/maptask.html. $F3'$ and $F3''$ are constructed utterances.

[3] Note that both of these constraints are needed – neither is sufficient on its own for all the rankings.

[4] The first three issues are also considered by Jokinen and Wilcock (2001) for NLG, but not in terms of OT.

(2) G1: Where are you in relation to the top of the page just now?
F1: Uh, about four inches.
G2: Four inches?
F2: Yeah.
G3: Where are you from the left-hand side?
F3: About two.
F3′: (About) two inches.
F3″: I am (about) two inches from the left-hand side of the page just now.

• Polite/formal system, rely less on information enrichment; informal system, rely more on information enrichment
• Beginning of the dialogue, rely less on information enrichment; rest of the dialogue, rely more on information enrichment
• Naive users (or a system that is used seldom by the same person), rely less on information enrichment; expert users (or a system that is used often by the same person), rely more on information enrichment
• Adapt to the user's level of reliance on information enrichment, making the system appear more co-operative (cf. (Garrod, 1999))

Any one of these factors affects the ranking of the constraints introduced above, and their values will give rise to different rankings. For instance, a dialogue system that is designed to be very formal and correct, will make use of a constraint ranking where information enrichment is rare among the optimal candidates throughout the dialogue. The opposite is true for a more informal system.

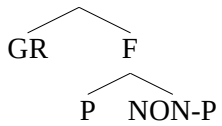
What's more, a dialogue system can be designed to *rerank the constraints* depending on conditions that change during the dialogue. For example, a system making use of recognition scores, will use one type of ranking if the score was high, and will need to rerank the constraints if the score was lower. Another example is the reranking of constraints to give an optimal candidate that relies on information enrichment to the same extent that a preceding user utterance does. In human-human dialogue the latter can be seen in the frequent occurrence of information enriched question-answer pairs (see, e.g., $G3 - F3$ in (2)).

Concretely, in a dialogue like (2) above, the question is whether to generate $F3$, $F3'$, or $F3''$, and how to rank the constraints to give precisely the desired optimal candidate.

I will now go through the different rankings that can be selected dynamically. First, a note on input and candidates. Input is here the contextual content of an utterance to be generated, where the content is marked up for information structure. For the tableaux below, a contextual content with all of ground, focus, prominent element and non-prominent material is used. For explanatory purposes, a prominent element may correspond to 'two', a focus (prominent element together with non-prominent material) to 'two inches', a ground to 'the distance is', and a ground-focus to 'The distance is two inches', all in the context of, say, the question 'What is the distance?'.

The candidates created from the input are the 'power set' of the information elements in the contextual content, with the reservation that the mark up is hierarchical (see the figure in example (3) below): the presence of a prominent element and non-prominent material implies the presence of a (full) focus. In the tables below this is indicated using the notation 'P_NON-P/F'.

(3)



As is usual, a dotted line between two constraints indicates that the ranking of these two constraints in relation to each other is unknown, that is, the outcome is independent of the order of these two particular constraints in relation to each other. In addition, I will use a double line to indicate indeterminacy between several constraints, in the

sense that it demarcates partial rankings.

Note that my use of partial rankings is not to be confused with the partial orderings of Anttila (1997). He uses partial rankings between constraints to explain examples of variation in Finnish morphology; different total orders (different tableaux) created from the partial one (the grammar) give different winners. In my approach, constraints are partially ordered because there is no conflict between them; the various total rankings that can be created from the partial ranking all give *the same* winner.[5]

4.1 Maximal reliance on information enrichment

Maximal reliance on information enrichment is the generation of just the prominent element when such can be determined. It involves the following partial rankings: *NON-PR, PROM ELEM >> FOCUS, and *GROUND >> GROUND. The tableau is given in figure 1.

In this tableau, the left-most column lists all the output candidates, as described above. The ranking between PROM ELEM and *NON-PR is not known (or, equivalently, their ranking in relation to each other does not affect the outcome), as indicated by the dotted line, but both of them are ranked higher than FOCUS. *GROUND is ranked higher than GROUND, but the ranking of these two constraints in relation to the other three does not change the result, which is the meaning of the double line.

Each star indicates a violation of a constraint by a candidate, and the optimal candidate is determined in the usual way, which can be described as: the optimal candidate is the one with the fewest violations of the highest constraint on which the two candidates differ.

Thus, in figure 1, *P* ('Two') is the optimal candidate. Informally, and intuitively, what this tableau says, is that for maximal reliance on information enrichment, avoiding ground is more important than producing ground, and producing the prominent element and avoiding non-prominent material are both more important than producing a full focus.

4.2 Minimal reliance on information enrichment

Minimal reliance on information enrichment means producing a full ground-focus utterance. The partial rankings are FOCUS >> *NON-PR, and GROUND >> *GROUND, and the tableau is given in figure 2. The optimal candidate is *GR P_NON-P/F* ('The distance is two inches').

4.3 Intermediate reliance on information enrichment

Intermediate reliance on information enrichment occurs in two cases. In one case, the optimal candidate is *P_NON-P/F* ('Two inches'), and the partial rankings involved are FOCUS >> *NON-PR, and *GROUND >> GROUND. This is depicted in figure 3.

In the other case, figure 4, the partial rankings are *NON-PR >> FOCUS, and GROUND >> *GROUND, to make *GR P* ('The distance is two') the optimal candidate. For this second case, the presence of PROM ELEM is required to separate *GR P* from the candidate involving just *GR*, and it can be ranked anywhere among the constraints.[6]

4.4 Focus-ground and all-focus utterances

The discussion and tableaux above assumed an utterance whose contextual content could be partitioned for all of ground, focus, prominent element, and non-prominent material. Now, many utterances have a contextual content consisting of only a focus and a ground, or just a focus. These can also be handled by the rankings and constraints given so far.

For focus-ground contents, there will be four candidates: *GR F*, *GR*, *F*, and $\emptyset$. The constraints that play a role in determining the optimal candidate are FOCUS, GROUND, and *GROUND, the other two (PROM ELEM and *NON-PR) being violated or vacuously satisfied by all candidates.

[5] The assumption in OT is that theoretically there *is* a complete ranking. My partial rankings are then to be interpreted as that given the current constraints there is no way of finding this ranking.

[6] The effect of PROM ELEM in the rankings for maximal and intermediate reliance on information enrichment is to ensure that the null candidate, the empty utterance, does not end up as the optimal candidate. The same effect could be achieved, in perhaps a more transparent way, through a constraint stating that a candidate should have semantic content.

	PR. ELEM	:	*NON-PR	FOCUS	*GROUND	GROUND
a. GR P_NON-P/F		:	*		*	
b. GR P		:		*	*	
c. GR NON-P	*	:	*	*	*	
d. GR	*	:		*	*	
e. P_NON-P/F		:	*			*
▷ f. P		:		*		*
g. NON-P	*	:	*	*		*
h. ∅	*	:		*		*

Figure 1: Tableau for maximal reliance on information enrichment

	PR. ELEM	FOCUS	*NON-PR	GROUND	*GROUND
▷ a. GR P_NON-P/F			*		*
b. GR P		*			*
c. GR NON-P	*	*	*		*
d. GR	*	*			*
e. P_NON-P/F			*	*	
f. P		*		*	
g. NON-P	*	*	*	*	
h. ∅	*	*		*	

Figure 2: Tableau for minimal reliance on information enrichment

	PR. ELEM	FOCUS	*NON-PR	*GROUND	GROUND
a. GR P_NON-P/F			*	*	
b. GR P		*		*	
c. GR NON-P	*	*	*	*	
d. GR	*	*		*	
▷ e. P_NON-P/F			*		*
f. P		*			*
g. NON-P	*	*	*		*
h. ∅	*	*			*

Figure 3: Tableau for intermediate reliance on information enrichment (focus)

	PR. ELEM	*NON-PR	FOCUS	GROUND	*GROUND
a. GR P_NON-P/F		*			*
▷ b. GR P			*		*
c. GR NON-P	*	*	*		*
d. GR	*		*		*
e. P_NON-P/F		*		*	
f. P			*	*	
g. NON-P	*	*	*	*	
h. ∅	*		*	*	

Figure 4: Tableau for intermediate reliance on information enrichment (prom. element and ground)

For all-focus contents there are two candidates, F and $\emptyset$, and one constraint, FOCUS, will determine the winner.

5 Determining which ranking to use

Just how is one to determine which ranking is to be used by a particular system or for a given utterance in some context? It is important to note that this will be something *outside of* the ranking. As with the design of any dialogue system, the solution lies in the answers to questions asked about various features of a dialogue system; when, where, and how is the system to be used, who is going to use it, what kind of system behaviour is desired, etc.

For information enrichment, as for many other aspects of dialogue system development, one needs studies of human-human dialogue, or experimental setups such as Wizard-of-Oz, or evaluations involving real users.

Looking in more detail at the factors identified at the beginning of section 4, for speech recognition scores, various levels need to be tried when the system is being developed, and possibly also during evaluation with real users.

Regarding the choice between a polite/formal system and an informal one, questions such as the following may need to be answered: What is best suited to the system? What do users think? Maybe what is needed is a system that comes with a choice regarding degree of formality?

When it comes to the distinction between the beginning and the rest of a dialogue, it needs to be determined, as for all factors, whether it is a useful distinction in the system, and a measure is needed for what counts as the beginning of a dialogue.

For distinguishing between naive and expert users, some form of user modelling is needed.

An example of a system used often by the same user may be a personalised system in the home used several times daily, and one used seldom a flight information system utilised by a number of different people and less frequently by each one. Determining which category the dialogue system belongs to, also determines information enriched behaviour.

Finally, adaptation to the user's level of reliance on information enrichment can be determined from linguistic studies and experiments, and through system evaluation.

All of these factors will interact with the various rankings in different ways. For instance, a situation involving intermediate level of reliance on information enrichment, say the production of a full focus although a prominent element has been determined, can be a high recognition score in conjunction with a system that is not completely informal.

6 Interference and bidirectionality

When producing utterances that rely on information enrichment, speakers need to take into account hearers' ability to construct an appropriate embedding structure, hence an appropriate contextual content, given a compositional content and the context. This becomes evident in examples like (4). Suppose that Edith's extension number is 1439, and this is what B is going to tell A. To what extent can B, in $B2$, rely on information enrichment (assuming B wants to exploit information enrichment maximally)?

(4) A1: What is your extension number?
B1: One eight three nine
A2: And what is Edith's extension number?

The contextual content of $B2 - to - be - generated$ can be paraphrased as: '*Edith's extension number is _*' as the ground, '*1_39*' the non-prominent part of the focus, and '*4*' the prominent element. Now, although it is possible for the speaker to utter only 'Four', that is, just the prominent element, this gives the hearer no chance of unambiguously recovering the contextual content. This is an example of what I call *interference*. The term is borrowed from Givón (1983), and adapted to information enrichment it involves the presence of semantically compatible contextual material that can give rise to ambiguity.

The solution to interference that can be constructed within OT is one that lends itself naturally to a dialogue context: bidirectionality (e.g., (Blutner, 2000; Jäger, 2002; Buchwald et al., 2002)). Blutner and Jäger formalise two communicative principles, one minimising hearer effort, and the

other minimising speaker effort, and show the interaction of these principles using bidirectionality.

These two principles are clearly in play in dialogue involving information enrichment. For example (4) above, the optimisation of the (speaker's) output needs to be followed by an optimisation of the (hearer's) interpretation. That is, given the input form *4*, what is its optimal interpretation in the given context? Clearly, no such interpretation can be found. The optimal candidate using bidirectionality will instead be one that includes the full number, *1439*. That is, in the given context, the maximal reliance on information enrichment is not using a prominent element, but a full focus.[7]

I omit full details here, but the analysis involved will need to create all the candidate interpretations for the input *4*. Next, these candidates are evaluated with regard to the constraints, and candidates involving *4839*, *1439*, *1849*, and *1834* will be equally optimal – there is no constraint separating them.[8]

Once this has been determined, the system needs to 'back off' to a lesser degree of reliance on information enrichment. If this is intermediate reliance using a full focus – 'One four three nine' – candidate contents will be created for this. The winner from the interpretation perspective will be *1439* as a focus relying on information enrichment for its ground.

In comparison to (Blutner, 2000), my approach needs to handle the existence of several different tableaux, for instance through the 'backing off' to a lesser degree of reliance on information enrichment as just described. An alternative is to do bidirectional optimisation for all the tableaux, which gives that both 'One four three nine' and 'Edith's extension number is one four three nine' give optimal candidates. Then, other factors are used to determine which of these two candidates is to be selected.

[7] An alternative is possibly using the prominent element together with only part of the non-prominent material, as in the utterance 'One FOUR', where capitals indicate nuclear stress.

[8] These candidates are based on the assumption that in this particular context, an extension number always consists of four digits. Without this piece of information, the number of candidates will of course be even larger.

Note that the step optimising interpretation involves the utterance's being marked up for information structure. The generation step in section 4 similarly assumes that this has already been done. I presume that it is possible to determine OT constraints and rankings also for this.

7 A computational note

The OT analysis presented here is intended for a dialogue system, that is, it is intended to be implemented, so a few remarks on OT in a computational setting are in order.

Several approaches to the implementation of OT constraint checking make use of finite-state techniques. Karttunen (1998) uses an example from phonology and shows how the generation of candidates and constraint application can be composed into a single transducer, a single network. Jäger (2002) reformulates Blutner (2000) and also discusses some formal properties of bidirectional OT as outlined by the latter. Notably, Jäger discusses bidirectionality in a finite-state setting.

Now, both Karttunen and Jäger acknowledge the limitations of the OT models that can be formulated as finite-state transducers. Jäger mentions that finite-state techniques are in general too simple to handle syntax, semantics and pragmatics, so the implementation of such analyses in OT seems to be an open research question.

The OT analysis that I presented in section 4 involves a small and finite set of candidates, the constraints all involve checking whether a particular informational element is part of the candidate or not, and constraints only have at most one violation. This may mean that a finite-state implementation is possible. However, the step presupposed in section 4, the assignment of information structure (a step that is also involved in the bidirectional analysis), involves reasoning using a fairly complex information state, which is probably less likely to lend itself to a finite-state analysis.

Instead, I think that in a context such as this – the computation of information structure and the generation of information enrichment in a practical dialogue system – work could usefully be spent on making the system avoid having to create all the candidates. One possible solution is to incorporate the effect of the constraints and their ranking in the

GEN component, making GEN different for the different degrees of reliance on information enrichment, and only producing the optimal candidate in each case. Similarly, when bidirectionality is considered, the system should not produce all of, say, *4839*, *1439*, *1849*, and *1834*, but be able to determine, *a priori*, that these would be equally (un)optimal.

8 Conclusion

Five OT constraints have been introduced to handle the kinds of information enrichment discussed above. Various rankings of the constraints are needed to give different optimal candidates, and the notion of dynamic constraint reranking in the generation component of a dialogue system was introduced to model the flexibility of information enrichment. The different rankings show that the degree of reliance on information enrichment arises from, on the one hand, a conflict between generating a full focus and avoiding non-prominent material, and, on the other, a conflict in whether to generate ground or not. The need for bidirectionality in a dialogue system generating information enriched constituents was also discussed, and some computational considerations were presented. Given the theory, a subsequent step is the precise formulation of bidirectionality for information enrichment, and the implementation of the constraints and the dynamic rerankings in a dialogue system.

Acknowledgements. I would like to thank Benjamin Lyngfelt and the two anonymous reviewers for helpful comments and suggestions.

References

Arto Anttila. 1997. Deriving variation from grammar: A study of finnish genitives. In Benjamins, editor, *Variation, Change and Phonological Theory*, pages 35–68. F. Hinskens and R. van Hout and L. Wetzels.

David Beaver. 2004. The optimization of discourse anaphora. *Linguistics and Philosophy*, 27(1):3–56.

Reinhard Blutner. 2000. Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17(3):189–216.

Adam Buchwald, Oren Schwartz, Amanda Seidl, and Paul Smolensky. 2002. Recoverability optimality theory: Discourse anaphora in a bidirectional framework. In *Proceedings of Edilog 2002*.

Simon Garrod. 1999. The challenge of dialogue for theories of language processing. In Simon Garrod and Martin Pickering, editors, *Language Processing*, pages 389–415. Psychology Press.

Jonathan Ginzburg. 1999. Semantically-based ellipsis resolution with syntactic presuppositions. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning*, volume 1, pages 255–279. Kluwer.

Talmy Givón. 1983. Topic continuity in discourse: An introduction. In Talmy Givón, editor, *Topic Continuity in Discourse: A quantitative cross-language study*. John Benjamins.

Petra Hendriks and Helen de Hoop. 2001. Optimality theoretic semantics. *Linguistics and Philosophy*, 24:1–32.

Gerhard Jäger. 2002. Some notes on the formal properties of bidirectional optimality theory. *Journal of Logic, Language and Information*, 11:427–451.

Kristiina Jokinen and Graham Wilcock. 2001. Confidence-based adaptivity in response generation for a spoken dialogue system. In *Proceedings of the 2nd SIGdial workshop*.

Lauri Karttunen. 1998. The proper treatment of optimality in computational phonology. In *Proocedings of FSMNLP'98*, pages 1–12. International workshop on Finite-State Methods in Natural Language Processing, Bilkent University, Ankara, Turkey.

Alan Prince and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University. RuCCS Technical Report 2.

Stephen Pulman. 1997. Higher order unification and the interpretation of focus. *Linguistics and Philosophy*, 20:73–115.

Mats Rooth. 1996. Focus. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*. Blackwell.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press/Bradford Books.

Enric Vallduví. 1992. *The Informational Component*. Garland.

Henk Zeevat. 2001. The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics*, 17(3):243–262.

Information State Update: Semantics or Pragmatics?

Raquel Fernández, Matthew Purver
Department of Computer Science
King's College London, Strand, London WC2R 2LS, UK
{raquel,purver}@dcs.kcl.ac.uk

Abstract

We argue for an approach which treats the compositional semantic content of an utterance as including its basic dialogue update effects – those which can be derived entirely from its semantic and syntactic properties. This allows us to capture the distinction between these integral *semantic* contextual effects and those *pragmatic* effects which can only be determined from the interaction between features of the utterance and the context itself.

1 Introduction

This paper presents an approach to dialogue update processes that captures the distinction between that part of an utterance's contextual import that can be derived entirely from its semantic and syntactic properties, and that which results from interaction between features of the utterance and the context – by treating the former as part of the utterance's compositional *semantic* content and only the latter as having to be specified separately as *pragmatic* processes. We then show that this does not prevent an utterance's representation from articulating constraints on context, and illustrate this for context-dependent phenomena such as givenness and ellipsis.

1.1 Background

We adopt the approach to utterance representation introduced in (Purver and Fernández, 2003), which views utterances and their sub-constituents as instructions for contextual update: programs in a dynamic logic defined with respect to the dialogue gameboard (DGB) of (Ginzburg, 1996). The DGB provides a structured view of context in dialogue by keeping track of the following components: a set of commonly accepted FACTS; a partially ordered set QUD of questions under discussion (QUDs); and the LATEST-MOVE (LM) made in the dialogue.

In (Fernández, 2003), the DGB is formalised using first-order Dynamic Logic (DL) as it is introduced in (Harel et al., 2000). In short, DL is a multi-modal logic with a possible worlds semantics, which distinguishes between *formulae* and *programs*. Programs are interpreted as relations between states that change the values assigned to particular variables. They can be combined to form complex programs by means of a repertoire of program constructs, such as *sequence* **;** , *choice* $\cup$, *iteration* * and *test* ?. The different DGB components are modelled either as individual variables ranging over terms (e,g. LM, for the latest move), or as *stack* variables ranging over strings of terms (e,g. QUD, a stack of questions). Update operations are brought about by program executions that involve changes in variable assignments. The atomic programs are simple assignments ($x := t$), where x is an individual variable and t is a term; and X.**push**(x) and X.**pop** programs, where X is a stack variable and x stands for the element to be pushed onto X. Such programs can then be assigned to utterances and their sub-components by a HPSG grammar which relates programs to

grammatical types. The approach allows us to reflect the basic insights of dynamic semantics (e.g. indefinite NPs can be assigned programs which introduce new referents) and define a process of grounding and clarification, as well as specify update effects of utterances familiar from Information State (IS)-based theories of dialogue.

2 Update Programs

The basic assumption underlying the IS approach to dialogue modelling is that the main aspects of dialogue management are best captured by (i) keeping track of the relevant information available to each dialogue participant at each state of the conversation, and (ii) providing a full account of the possible update mechanisms that change this information. The notion of IS *update* is key, usually being governed by a set of *update rules* triggered by the observation and performance of dialogue moves.

Our starting point is, in fact, a fairly straightforward extension of this view: as long as dialogue move types can be incorporated into the grammatical representation of utterances, their update effects can also be seen as part of the utterance's linguistically conveyed information. The integration of direct illocutionary force into the grammar has been argued for in (Ginzburg et al., 2001b). The authors present an HPSG grammar where each illocutionary type introduces a constraint on the type of its message argument (*ask-rel* types are associated with *questions*, *assert-rel* types with *propositions*, and so on), with these message types being determined by syntactic form. SDRT (Asher and Lascarides, 2003) also assumes a *uniform semantics* of declaratives, interrogatives and imperatives, where each clause type is linked to its illocutionary force by means of compositional semantics. Our approach goes one step further in that it views the immediate contextual effects of these various illocutionary types (which are usually seen as brought about by independent IS update rules or pragmatic inference) as compositionally linked to syntactic and semantic properties of utterances. In (Purver and Fernández, 2003) this is achieved by associating appropriate DL programs with particular clause types, as shown inAVM (1) and AVM (2) for interrogatives and declaratives:

(1)
$$\begin{bmatrix} \textit{interrogative} & \\ \textsc{cont} & \boxed{1}\left[\textit{question}\right] \\ \textsc{c-prog} & A; \mathsf{QUD}.\mathbf{push}(\boxed{1}) \\ \textsc{head-dtr} \mid \textsc{c-prog} & A \end{bmatrix}$$

(2)
$$\begin{bmatrix} \textit{declarative} & \\ \textsc{cont} & \boxed{1}\left[\textit{proposition}\right] \\ \textsc{c-prog} & A; \mathsf{QUD}.\mathbf{push}(whether(\boxed{1})) \\ \textsc{head-dtr} \mid \textsc{c-prog} & A \end{bmatrix}$$

Introducing DL programs into the grammatical representation of utterance types allows us to reflect the part of their contextual import which is compositionally derivable. Just as an indefinite NP intrinsically introduces a new entity into the context, *ask* moves, and therefore questions, intrinsically introduce new QUDs (in our formalisation, push their content onto QUD). Similarly, moves which assert a proposition p push the question $whether(p)$ onto QUD. Note that this is not to deny that some questions and assertions might have further contextual effects, or even that QUD introduction might also be achievable by other, less obvious means. The point here is that an important part of the context change potential of *ask* and *assert* moves (namely the fact that they introduce particular QUDs) can be fully derived from their grammatical properties. As far as dialogue goes, it is therefore possible and, we think, desirable to consider such updates as part of the semantic contribution of interrogative and declarative utterances (just as much as the introduction of new referents is part of the semantic contribution of indefinites). This is precisely what our programs achieve.

The main issue to consider now is: can this approach be extended to all dialogue move types? In other words, is it possible to encode the main contextual updates brought about by dialogue moves into the grammatical representation of utterances, thus removing the need for independent update rules?

3 Semantic vs. Pragmatic Updates

Here we must distinguish two different kinds of updates: updates whose assignment can be determined purely by properties of the utterance itself, and those which should only be assigned to ut-

terances given certain information in the current state.

3.1 Semantic (Direct) Updates

The immediate update effects of direct moves such as *ask* and *assert* (as given above) can be determined by simple examination of the linguistic properties of utterances – they don't have to be inferred using pragmatic information. The same can be said for many other move types included in dialogue act taxonomies such as *greetings*, *closings* and *acknowledgements*. In fact, one could argue that the meaning of an acknowledgement can *only* be represented as a contextual update – in our approach, acknowledgements are associated with a program that pushes onto FACTS whatever proposition was previously under discussion:

(3)
$$\begin{bmatrix} \textit{ack-cl} & \\ \text{CONT} & \textit{acknowledge-rel} \\ \text{C-PROG} & \textbf{head}(\texttt{QUD}) = whether(p)?; \\ & \texttt{FACTS}.\textbf{push}(p); \texttt{QUD}.\textbf{pop} \end{bmatrix}$$

The complex program shown in AVM (3) requires for its success the existence of some question $whether(p)$ under discussion. If there is no such question, the program will not succeed, the utterance cannot be understood or grounded (and on our account, will cause a clarification question). This seems correct: if there is nothing to be integrated into the common ground, or if the current QUD is a wh-question, an acknowledgement will seem quite odd. Acknowledgements require suitable QUDs in order to be understood (just as understanding an *ask* move seems to require recognition of its intention to raise a new QUD). It is important to note that although an acknowledgement therefore imposes a restriction on the type of state to which it can apply (expressed as a *test* subprogram), there is no need for pragmatic information to determine what its update effects should be (what program to associate with it).

3.2 Pragmatic (Indirect) Updates

Most dialogue act taxonomies and implemented dialogue systems include other move types which are less directly associated with the linguistic or grammatical form of the utterance. Indirect speech acts such as requests or commands can take the form of questions (*"Can you close the door please?"*); questions can be rhetorical (*"Do I look like an idiot?"*).

Answers in the Grammar? A common example in dialogue systems is an *answer* move. Answers differ from assertions and questions in many respects: if we were to specify the contextual update effect of an answer by a program, it might be of the form QUD.**pop** – i.e. a program that downdates QUD by popping the maximal question under discussion, rather than one which adds a new question to the stack. The notion of answerhood employed by many dialogue systems involves assertion of a proposition that unifies with the propositional content of a QUD question (see e.g. (Traum, 2003)). This could be easily defined within the grammar as in (4):

(4)
$$\begin{bmatrix} \textit{answer} & \\ \text{CONTENT} & \boxed{1}\textit{proposition} \\ \text{C-PROG} & (\textbf{head}.(\texttt{QUD}) = \lambda\{\ldots\}.\boxed{1})?; \texttt{QUD}.\textbf{pop} \end{bmatrix}$$

The problem is of course that there will be no way of associating this program with an utterance based on its internal grammatical properties alone: to determine which update effects to associate with a declarative (those of an *assert* program as in (2) or an *answer* program as in (4)), we must take into account its relation to some relevant contextual information (precisely the maximal QUD). Given our formalism, this could be phrased as a single program using the *choice* operator:[1]

(5)
$$\begin{bmatrix} \textit{declarative} & \\ \text{CONTENT} & \boxed{1}\textit{proposition} \\ \text{C-PROG} & ((\textbf{head}.(\texttt{QUD}) = \lambda\{\ldots\}.\boxed{1})?; \texttt{QUD}.\textbf{pop}) \\ & \cup\, \texttt{QUD}.\textbf{push}(whether(\boxed{1})) \end{bmatrix}$$

However, we see several problems with such an approach. The first is that downdating QUD, which must be one of the update effects of an answer, does not need to be performed in order to *understand* it: one can understand an answer without accepting it, and indeed can discuss whether it is true – so making this downdate part of the semantic content seems inappropriate. A second is that the contextual effects expressed as the se-

[1] Another equally unattractive solution would be to see *all* declarative sentences as ambiguous between being answers and assertions, with two alternative analyses assigned by the grammar and with the decision between the two made later.

mantic content have now become *determined by* context, rather than just expressing *restrictions on* context as before. Even worse, the third is that this approach seems very difficult to scale up to more complex notions of answerhood: in particular, indirect answers could not be detected by the test of unification with the head QUD as above, but would require some further inference – thus the C-PROG program would have to involve such inference and presumably access to further contextual information. Semantic content, then, would not only be context-dependent but include (possibly unrestricted) access to pragmatic components.

Note that this is not the case for acknowledgements: the program in AVM (3) shows no contextually determined variation in its possible effects – the program simply imposes a restriction on the current state that has to be met for the update program to be executable: if the restriction is not met, the program will just fail.

Answers outside the Grammar A more reasonable approach therefore seems to be to take answers as having update effects at two levels: at the *direct* level, expressible as part of the grammatically assigned semantic content, the effect of an assertion as in AVM (2) (introducing $whether(p)$ to QUD); and then at the *indirect* level the further answering effect (popping the QUD stack). This indirect effect must be outside the realm of grammar, as its applicability will depend on the current IS – reasoning or update rules must decide whether p answers the current maximal QUD, and if so whether it is to be accepted.

The semantic content then no longer varies with context (although it can still express a restriction on context as before), and can be grammatically assigned as long as this basic program is not inconsistent with the possible later indirect updates. For answers, the basic effect is an assertion which is then used to license QUD downdate; for rhetorical questions, the basic effect would be to introduce a new QUD which is seen to be already answered (by domain/world knowledge or context) and thus immediately downdated; for indirect requests, again the basic effect would be to introduce a new QUD, which further inference would then presumably determine to be influenced by the indirectly requested task (see (Ludwig, 2001) for a similar approach to inferring requests from basic declaratives).

This distinction, between direct updates which stem from the utterance's internal properties on the one hand, and indirect updates which stem from its relation to context on the other, now allows us to draw a line between the kind of updates that can be thought of as part of an utterance's *semantic content*, and those that should be specified separately by means of *pragmatic* operations (e.g. update rules or inference). Note that this distinction does not correspond to the one drawn between forward and backward looking acts (Allen and Core, 1997) – acknowledgements and answers are both usually classified as backward-looking. In a typical system such as GoDiS (Larsson et al., 2000) the only move type which requires separate pragmatic processes (and which we would therefore classify as indirect) is *answer*.

3.3 Discourse & Turn-Taking Effects

So far we have assumed that the direct update effect of questions and assertions is to introduce a question which becomes topmost in QUD (q, in the case of asking a question q, and $whether(p)$ in the case of asserting a proposition p). In Ginzburg's account, this topmost position is taken to explicate why the last question posed takes conversational precedence (has to be addressed first) and why elliptical forms are licensed as responses to it. Several authors (Asher, 1998; Ginzburg, forthcoming), however, have pointed out that when multiple moves are performed by the same speaker within a single turn, the evolution of QUD seems to be somewhat different.

(6)
A : Where were you? Did you talk to anyone?
B : I was at home. I didn't talk to anyone.
B': I didn't talk to anyone, I was at home.

(7)
A : Who did you invite? Did you invite Jill?
B : Yes. Also Merle and Pat.

(8)
A : Who did you invite? And why?
B : Merle and Pat, because they are very undemanding folks.

Examples like the ones above have motivated a view according to which the way several queries

asked in sequence by the same speaker are integrated into QUD depends on the discourse relation that links them. Thus, the questions in example (6) (adapted from (Asher, 1998)) are taken to be in what has been called *coordinate structure*, with none of them taken precedence over the other one. The questions in (7), on the other hand, would be related by *query-elaboration*, which would account for the fact that apparently the second one takes precedence over the first one. Contrastingly, the questions in example (8) would be related by *query-extension*, which would explain why in this case the first question tends to be answered first.

At a first glance, one may think that three different QUD updating operations are needed to account for these examples: one that pushes the second question next to the maximal QUD, the standard push on top operation, and a "push-under" operation (or *QUD-FLIP*, as it is called by Ginzburg (forthcoming)) that would push the second question under the topmost element in QUD. If we were to specify these distinctions in our account, we would presumably have to do so by a program that first tests the kind of rhetorical relation that holds between the head of QUD and the current question, and then applies the right QUD.**push** program, as in AVM (9):

(9)
$$\begin{bmatrix} \textit{interrogative} & \\ \text{CONT} & \boxed{1}\left[\textit{question}\right] \\ \text{C-PROG} & \textit{q-elab}(\textbf{head}(\text{QUD}),\boxed{1})?;\text{QUD}.\textbf{push}(\boxed{1})\cup \\ & \textit{coor}(\textbf{head}(\text{QUD}),\boxed{1})?;\text{QUD}.\textbf{push-next}(\boxed{1})\cup \\ & \textit{q-ext}(\textbf{head}(\text{QUD}),\boxed{1})?;\text{QUD}.\textbf{push-under}(\boxed{1}) \end{bmatrix}$$

However, as with answers in the previous section, the test subprograms in AVM (9) not only express restrictions on the kind of state the program can be applied to (like the program for acknowledgements in AVM (3) above), but crucially they both require further pragmatic information, and use it to determine the program's effects. To decide on the kind of push program that has to be applied, we must first compute the rhetorical relation that holds between the current question and the maximal QUD, and this will involve using pragmatic reasoning. Thus, to use a grammar to assign the complex program in AVM (9) to interrogative clauses seems both problematic and rather pointless, given that its update effects are actually ambiguous between three different QUD updating operations, and such ambiguity is only resolved by pragmatic knowledge about rhetorical relations.

Instead, we think that the *semantic* update effects of questions and assertions are still best characterised by the simple programs proposed in AVMs (1) and (2). In fact, a closer look at the examples above reveals that the discourse relations that link different questions in the same turn do not play such a significant role in determining availability and licensing of elliptical forms:

(10) A : Who did you invite? Did you invite Jill?
B': (I invited) Merle, Pat, and Jill, yes.

(11) A : Who did you invite? And why?
B': I thought we'd need a guitar, so Merle.

As (10) and (11) show, regardless of the rhetorical relation that holds between the questions, both questions are still available: they can be answered by a fragment and it is up to the addressee to choose which one to answer first. We can therefore assume that the basic QUD update mechanism (and therefore our basic programs) do not require, and need not be affected by, computation of the rhetorical relation that links moves within a single turn. This is not to claim that discourse relations are not needed at any level: they may be required to establish the coherence of the dialogue at the pragmatic level, or indeed to decide which member of QUD to answer first. We do claim however that one can still specify some basic *semantic* contextual update potential brought about by questions and assertions as monotonically introducing QUDs.

QUDs introduced in the same turn must then have equal satus in the QUD stack. We regard this coordinate status as a consequence of the dynamics governing turn management. It is implicitly assumed that information about turn taking and turn change is part of the resources commonly shared by dialogue participants. To encode this information explicitly in the dialogue context, here we assume that QUD not only includes the questions under discussion themselves, but also information on turn change that acts as an additional structuring mechanism of the QUD order. Assuming that

turn change is recorded in QUD, the maximal elements of QUD are then those questions between the top and the turn change indicator.[2]

4 The Contextual Interface

As we have seen, *semantic* update programs such as acknowledgements can specify restrictions on the current state, without requiring state information to determine their form. How does this distinction apply for other contextually-dependent phenomena such as ellipsis?

4.1 Conditions on State

Some interaction between the utterance representation and the context is required not only by moves like acknowledgements, but by the treatment of givenness: given referents such as those associated with definite NPs and proper names contribute sub-programs which express restrictions on the type of state to which the utterance program can be successfully applied – to whit, that the state contain a suitable antecedent (seeAVM (12) and AVM (13)). The same is true for other givenness effects such as the focus/ground distinction: following (Engdahl et al., 1999; Ginzburg, forthcoming) a particular focus/-ground partition introduces a sub-program which must find a particular maximal QUD in the current state. This type of program, then, expresses a condition on the kind of state to which it can apply: in other words, the kind of context in which an utterance is licensed.

(12)
$$\begin{bmatrix} \textit{definite} & \\ \textsc{cont} & \boxed{1}[\textit{parameter}] \\ \textsc{c-prog} & (\boxed{1} \in \mathtt{BG/FACTS})? \end{bmatrix}$$

(13)
$$\begin{bmatrix} \textit{root-clause} & \\ \textsc{info-struct} & \begin{bmatrix} \textsc{focus} & \boxed{2} \\ \textsc{ground} & \boxed{3} \end{bmatrix} \\ \textsc{c-prog} & (\mathbf{head}(\mathtt{QUD}) = \lambda\boxed{2},\boxed{3})? \end{bmatrix}$$

4.2 Fragments

Elliptical fragments can also be seen in this way: as being licensed only in certain types of context, and therefore as expressing conditions on the kind of state to which their programs can apply. Fragments, of course, specify their content only partly, requiring the presence of some information in context in order to resolve their fully specified sentential content. Ginzburg et al. (2001a) analyse this by use of two contextual features in their HPSG grammar, MAX-QUD and SAL-UTT: the content of a fragment is specified in terms of constraints on these, by identifying the propositional content of the elliptical utterance with that of MAX-QUD and the referential index of the fragment itself with that of SAL-UTT. Until resolution in context, this information is essentially underspecified. Schlangen (2003), on the other hand, regards the content of such an elliptical utterance as containing an unknown anaphoric propositional relation, which must be enriched using contextual inference.

Instead, we regard elliptical fragments as introducing sub-programs which must ensure that the required contextual information is present in the current state, and by finding it, fully instantiate the content. The grammatical approach can directly follow that of Ginzburg et al. (2001a): the content of a (declarative) elliptical fragment utterance is taken to be a proposition which must be associated with the current MAX-QUD question; the referential index of its head daughter must be identified with that of a SAL-UTT utterance which is also constrained to be syntactically parallel to it. This is expressed in the grammar via the type *decl-frag-cl* (see AVM (14)).[3]

(14)
$$\begin{bmatrix} \textit{decl-frag-cl} & \\ \textsc{content} & \boxed{1} \\ \textsc{head-dtr} & \begin{bmatrix} \textsc{cat} & \boxed{2} \\ \textsc{cont}\,|\,\textsc{index} & \boxed{3} \end{bmatrix} \\ \textsc{context} & \begin{bmatrix} \textsc{max-qud} & [\textsc{prop} \ \boxed{1}] \\ \textsc{sal-utt} & \begin{bmatrix} \textsc{cat} & \boxed{2} \\ \textsc{cont}\,|\,\textsc{index} & \boxed{3} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Now, the only change that must be made is that top-level sentences (in our grammar, signs of type *root-cl*) must add sub-programs which require the specified contextual information to be found, as shown in AVM (15).[4] Note that the order of the

[2] A way of implementing this idea is to think of QUD as a stack of sets. See (Fernández and Endriss, ms) for a formalisation of this in the context of dialogue protocols.

[3] Similar specifications can be given for short interrogatives, sluices, bare adjuncts and so on following (Fernández et al., 2004) directly.

[4] This *root-cl* specification also includes a sub-program to

program is important: the contextual information must be identified in the initial state, before it is changed by the utterance program (which may of course update QUD), and of course before the LM state variable can be set to the fully specified move (the overall utterance content).

(15)
$$\begin{bmatrix} \textit{root-clause} & \\ \text{CONTENT} & \boxed{1}\left[\textit{illoc-rel}\right] \\ \text{CONTEXT} & \begin{bmatrix} \text{MAX-QUD} & \boxed{2} \\ \text{SAL-UTT} & \boxed{3} \end{bmatrix} \\ \text{C-PROG} & (\textbf{head}(\texttt{QUD}) = \boxed{2})?; \\ & (\textbf{head}(\texttt{UTT}) = \boxed{3})?; A; \texttt{LM} := \boxed{1} \\ \text{HEAD-DTR} \mid \text{C-PROG} & A \end{bmatrix}$$

This seems to make the status of this contextual information clearer than in either Ginzburg et al. (2001a) or Schlangen (2003)'s approach. In the former, the utterance is left underspecified by the grammar, and we must assume separately specified pragmatic routines (update rules?) to fill it in; in the latter, this underspecification is replaced by anaphora essentially unaccompanied by information about possible antecedents, which must be identified by pragmatic inference. In our approach, not only is the method of content specification fully defined by the grammar as a program, the source of the antecedents (particular state variables) is also made clear.

Note the similarity between this program and that introduced by information structure in-AVM (12). Both programs express a constraint on the current maximal QUD, and therefore restrict their utterances' use to suitable contexts.[5]

4.3 Setting Up State Conditions

Note that not only does the program for the fragment specify the state variables where antecedents must be found, the program for the previous utterance will have specified how the values of these state variables were updated. As already shown in (1) above, interrogatives introduce questions to QUD – this will automatically provide a suitable head value of QUD for an elliptical answer which follows it. In fact, the program for wh-interrogatives also pushes a salient utterance (the wh-word corresponding to the question's abstracted parameter) onto the UTT stack, thus providing a state which will fulfill both the requirements of an elliptical fragment:

(16)
$$\begin{bmatrix} \textit{interrogative} & \\ \text{CONT} & \boxed{1}\left[\lambda\boxed{3}.p\right] \\ \text{C-PROG} & A; \texttt{QUD}.\textbf{push}(\boxed{1}); \texttt{UTT}.\textbf{push}(\boxed{2}) \\ \text{HEAD-DTR} \mid \text{C-PROG} & A \\ \text{CONSTITS} & \left\{\ldots\boxed{2}\left[\text{CONT} \quad \boxed{3}\right]\ldots\right\} \end{bmatrix}$$

Similarly, declarative utterances (as we have already seen in (2) above) introduce $whether(p)$ QUDs; indefinites also introduce programs to push themselves onto UTT for later resolution of sluices.

4.4 Ordering Sub-Programs

The specification of AVM (15) is designed to ensure that sub-programs are executed in a certain order: firstly, checks on state variables, before any utterance programs have had any effect; secondly, the sub-programs projected by individual phrases (and inherited by the sentence from its daughters); and thirdly the top-level effects of the utterance – updating QUD, UTT and LM. The ordering of the daughter sub-programs themselves will also be important to account for e.g. intrasentential anaphora and presupposition projection. Anaphoric definites and pronouns must be able to identify variables introduced by preceding indefinites as their referents, so we must ensure that the indefinite programs which introduce them are executed before the definite programs which attempt to find them. In English at least, this requires sub-programs to be put together in linear order, and this is simply expressed:

(17)
$$\begin{bmatrix} \text{C-PROG} & A; \ldots; B \\ \text{DTRS} & \left\langle\left[\text{C-PROG} \quad A\right], \ldots, \left[\text{C-PROG} \quad B\right]\right\rangle \end{bmatrix}$$

5 Summary

Update effects which are specified entirely by, and are inseparable from, utterances themselves can be represented as part of their grammatically assigned content, even when this content is contextually dependent. It is only when the context determines the form of these effects (the type of pro-

set the latest-move LM variable to the value of the utterance's content, a move – see (Purver and Fernández, 2003) for details.

[5] Of course, the same *could* be said for answers, if (as discussed and rejected above) they were to be represented as testing for a suitable QUD and popping it from the QUD stack.

gram which represents them), as with answers, that we need these effects to be determined by pragmatic processes. This is, of course, not to deny that these pragmatic processes govern dialogue to a large extent: merely to say that the dividing line between semantics and pragmatics can be drawn in a different place. This approach is currently being implemented in a HPSG grammar and a prototype IS-based dialogue system.

6 Acknowledgements

We would like to thank two anonymous Catalog reviewers for several helpful comments that have significantly contributed to the final version of this paper. The authors are supported by ESRC grants RES-000-23-0065 and RES-000-22-0355 respectively.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers.

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

Nicholas Asher. 1998. Varieties of discourse structure in dialogue. In J. Hulstijn and A. Nijholt, editors, *Proceedings of the 2nd Workshop on Formal Semantics and Pragmatics of Dialogue (Twendial)*, Enschede, May.

Elisabet Engdahl, Staffan Larsson, and Stina Ericsson. 1999. Focus-ground articulation and parallelism in a dynamic model of dialogue. In *Task Oriented Instructional Dialogue (TRINDI): Deliverable 4.1*. University of Gothenburg.

Raquel Fernández and Ulle Endriss. ms. Abstract models for dialogue protocols. Under review.

Raquel Fernández, Jonathan Ginzburg, Howard Gregory, and Shalom Lappin. 2004. SHARDS: Fragment resolution in dialogue. In H. Bunt and R. Muskens, editors, *Computing Meaning*, volume 3. Kluwer Academic Publishers. To appear.

Raquel Fernández. 2003. A dynamic logic formalisation of the dialogue gameboard. In *Proceedings of the Student Research Workshop, EACL 2003*, pages 17–24, Budapest. Association for Computational Linguistics.

Jonathan Ginzburg, Howard Gregory, and Shalom Lappin. 2001a. SHARDS: Fragment resolution in dialogue. In H. Bunt, I. van der Sluis, and E. Thijsse, editors, *Proceedings of the 4th International Workshop on Computational Semantics (IWCS-4)*, pages 156–172. ITK, Tilburg University, Tilburg.

Jonathan Ginzburg, Ivan Sag, and Matthew Purver. 2001b. Integrating conversational move types in the grammar of conversation. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue (BI-DIALOG)*, pages 45–56.

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

Jonathan Ginzburg. forthcoming. *A Semantics for Interaction in Dialogue*. CSLI Publications. Draft chapters available from: `http://www.dcs.kcl.ac.uk/staff/ginzburg`.

David Harel, Dexter Kozen, and Jerzy Tiuryn. 2000. *Dynamic Logic*. Foundations of Computing Series. The MIT Press.

Staffan Larsson, Peter Ljunglöf, Robin Cooper, Elisabet Engdahl, and Stina Ericsson. 2000. GoDiS - an accommodating dialogue system. In *Proceedings of ANLP/NAACL-2000 Workshop on Conversational Systems*.

Bernd Ludwig. 2001. Dialogue understanding in dynamic domains. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pages 287–297. BI-DIALOG.

Matthew Purver and Raquel Fernández. 2003. Utterances as update instructions. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck)*, pages 115–122, Saarbrücken, September.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh.

David Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *Proceedings of the International Workshop on Computational Semantics*, pages 380–394, January.

Reference Resolution Mechanisms in Dialogue Management

Petra Gieselmann
Interactive System Labs
Universität Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
petra@ira.uka.de

Abstract

Humanoid robots which are able to walk and behave human-like became very popular in the last few years. Now it is high time that they are able to use more natural communication means so that the human-robot interaction resembles more and more to human-human communication. Therefore, in this paper, we evaluate different reference resolution mechanisms within a dialogue management system for human-robot communication in a household environment. User studies showed that most of the pronouns can be resolved by a pragmatic, simplified approach.

1 Introduction

Dialogue management systems as well as mechanisms for reference resolution are well known research areas. Nevertheless, they have mostly been analyzed from different points of view until now. In this paper, we want to combine both by using well known pronoun resolution mechanisms within a dialogue management system for human robot communication in a household environment. In this context which is specifically tailored for unexperienced users, it is important that the user can talk to the robot in the same way as to a human servant for example. Therefore, the communication has to be as natural as possible which also includes pronoun resolution and multimodal communication mechanisms.

This paper deals with reference resolution of personal and deictic pronouns. Natural human robot interaction in a household environment is also explored. Section two gives an overview of related work on anaphora resolution in general and on special reference resolution mechanisms used in dialogue management systems in particular. In section three, our dialogue manager is explained. Section four deals with context management and our mechanisms for reference resolution. Section five gives a conclusion and outlook.

2 Related Work

Pronoun resolution is a well examined field in computational linguistics. Different theoretical articles have been written on this topic and methods from the field of Artificial Intelligence, such as inference mechanisms and world knowledge, have been explored in detail. Here, we want to have a look at the problem from a more pragmatic point of view. Therefore, we want to concentrate on deictic pronouns which can be resolved by means of gesture recognition and personal pronouns which are resolved by our pronoun resolution mechanisms. Other resolution mechanisms are the topic of future research.

2.1 Reference Resolution in General

Since there are so many researchers dealing with reference resolution from different point of views, such as philosophy, psychology, linguistics, computer science, etc., we want to take into account here only a small part of them which is relevant for our research.

One of the oldest algorithms for resolving pronouns is Hobb's naive algorithm (Hobbs, 1977). It simply traverses the surface parse trees of the sentences in a text looking for noun phrases of the correct number and gender as antecedents for pronouns. Although this algorithm is quite simple, it works fine and about 90% of the pronouns can be resolved (Hobbs, 1977).

The theory of discourse structure and centering invented and further developed by Grosz et al. (Grosz and Sidner, 1986; Brennan et al., 1987; Grosz et al., 1995; Walker, 1998) serves for tracking discourse context and binding pronouns. First, a set of all the cospecification relationships is created. Then it is filtered, classified and finally ranked by some rules. These rules rely on the relationship between antecedent and pronoun, such as parallelism of grammatical function, recency, etc. Furthermore, continuing with the same entity in the discourse center is preferred over retaining it which is preferred over shifting the discourse entities completely. Although the algorithm is much more complicated than Hobb's naive one, the results are similar (Tetreault and Allen, 2003).

As an extension of the centering model, Strube uses a list of salient discourse entities which is called S-list (Strube, 1998). This list is ranked based on information status. Therefore, it uses the distinction between new and old information in the discourse and incorporates also preferences for inter- and intrasentential anaphora which is not included in the original centering model.

CogNIAC (Baldwin, 1995) is a pronoun resolution engine which defines a set of rules for finding the correct antecedent in a list. These rules are somewhat simple, such as "If there is only one possible antecedent in the preceding input sentence, use this"; world knowledge is not used for pronoun resolution. Nevertheless, these rules seem to be quite efficient given the fact that he reported about 92% precision.

All of these mechanisms have been developed by means of written texts. They can be also used for spoken communication to a certain extent, but have to be adapted to its special needs, especially covering spontaneous effects. Therefore, the next chapter deals with reference resolution mechanisms used in spoken natural language dialogues.

2.2 Reference Resolution in Multimodal Dialogue Management

Until now, there are only very few dialogue systems which use a reference resolution module because most of them have been specifically tailored for communication via phone, such as flight and train timetable information systems (McTear, 2002; Allen et al., 2000; Stallard, 2000), call-routing systems (Gorin et al., 2002), weather information systems, (Zue et al., 2000) etc. and do therefore not need reference resolution. But now since the number of systems for direct human machine communication from face to face, such as human robot interaction, increases, we need to take into account the situated and context-dependent communication, the changing environment, the multimodal interaction, etc. Therefore, we want to have a look at the resolution mechanisms necessary in situated and context-dependent communication.

For example, Kumar et al. uses an approach based on cognitive grammar which assumes conceptual semantics (Kumar et al., 2003). Reference domains identify representations for subsets of contextual entities to which can be referred, such as individual objects and also collection of objects. The important feature of a reference domain are its partitions which define in conjunction with focus and salience the criteria for reference resolution. Underspecified reference domains are composed with the existent context structure by means of grouping and assimilation. In this way, references can be resolved by finding the corresponding node within a context structure. Since the same mechanism is used for linguistic expressions and for gestures, different kinds of references, such as deixis and pronouns, can be resolved.

Other researchers (Landragin and Romary, 2003) propose a classification of referring modes which describes referring actions, and disambiguation principles to define the correct referent. References can be resolved by means of unification with the information available in context so that the one with the best unification result is kept. In this way, also deictic pronouns and pointing gestures can be resolved.

For the galaxy system, a whole context resolu-

tion server has been developed (Filisko and Seneff, 2003) which includes repairing mechanisms, anaphora and ellipsis resolution, history functions, etc. Pronouns are resolved by means of a discourse entity list which is searched for a possible antecedent.

Out of these approaches, we created a reference resolution model which uses similar methods, such as a list of possible antecedents and rules for the agreement between the antecedent and the pronoun. It also works for personal and deictic pronouns and is specifically tailored for human robot communication by including for example some knowledge about the actual situation of the robot. Therefore, it is not as theoretically complex as some of the mentioned approaches, but works efficiently in our scenario.

3 Dialogue Management

Our dialogue manager is based on the approaches of the language and domain independent dialogue manager ARIADNE (Denecke, 2002) which is specifically tailored for rapid prototyping because general concepts are already available and can be reused. Only the domain and language dependent components have to be implemented for new applications, such as: An ontology, a specification of the dialogue goals, a data base, a context-free grammar and generation templates.

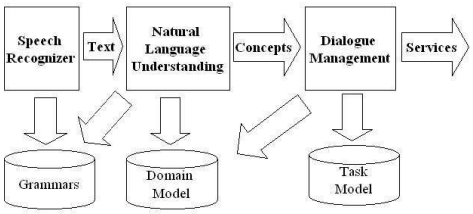


Figure 1: The Dialogue Management Workflow with Its Resources

The dialogue manager uses typed feature structures to represent semantic discourse information (Carpenter, 1992). In figure 1, the whole dialogue management workflow can be seen: First of all, the user utterance is parsed by means of a context-free grammar which is enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk. In figure 2, you can see a part of the ontology we defined for our robot dialogue system. It consists of different objects available in the kitchen, actions the robot can accomplish for the user and properties of the objects resp. the actions.

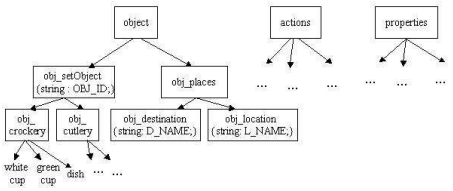


Figure 2: Part of the Ontology

An example of the semantic representation which is created during parsing can be found in figure 3. This semantic representation is compared against the dialogue goals. If all the necessary information to accomplish a goal is available, the dialogue system calls the corresponding service. But if some information is still missing, the dialogue manager uses clarification questions to get this information from the user. The spoken output is created by means of generation templates.

```
[ act_put  OBJ
  [ obj_puttable
    [ generic:NAME [ "it" ] ]
  ]
  [ DESTINATION
    [ DEST [ "table" ] ]
  ]
```

Figure 3: Semantic Representation of the Sentence ”put it on the table”

The database serves as a context model which includes different world knowledge sources and is used for the resolution of references as described below. Therefore, you can find there information on the position of the objects in the world as well as information on possible antecedents.

Also the ontology plays an important role in reference resolution because it is used to define the semantic agreement between the reference and its antecedent: If both of them belong to the same category or to a subcategory in the ontology, then there is a semantic agreement between them. In the example in figure 3, you can see that ”it” refers to an object which is puttable because of the verb

”put” which expects a puttable object. This means that other possible antecedents are semantically excluded. If the user said in the previous sentence for example ”get the cup from the board”, the ”board” cannot be an antecedent for the pronoun ”it” because it belongs to another category in the ontology. In this way, we assure that only semantically useful antecedents are taken into account by our algorithm.

4 Context Modeling for Reference Resolution

As you can also see in the example below (see figure 4), the two different types of references we want to resolve are personal pronouns and deictic pronouns. The reference resolution for both of them takes place during the creation of the semantic representation. Therefore, the input is the parsed user utterance transformed into a semantic representation as you can see in figure 3 and the output is the semantic representation enhanced with reference resolution information.

4.1 Our Context Model

The context model contains information on the environment: For example, all the available objects are stored there with their three-dimensional position in the room. This information can also be updated during the actual dialogue processing, if an object is moved by the user or by the robot itself.

In addition, possible antecedents are stored in the context model in a list similar to Strube’s S-list. Since we only found nominal antecedents for the pronouns in our user studies, we decided to resolve only these pronouns in a first step. In addition, some expletive pronouns are already covered by the grammar by means of expressions such as ”it is too dark in here”; others cannot be resolved at the moment.

We implemented our context model in such a way that it works similar to the human brain and therefore ”forget” old antecedents after a certain period of time (Clark, 1978). Whenever a new user utterance comes in, the context model is updated with the corresponding possible antecedents.

4.2 Mechanisms for Pronoun Resolution

For reference resolution, the context model is used and linguistic expressions, such as personal pronouns, as well as pointing gestures and deictic pronouns are both resolved - in multimodal parsing or in pronoun resolution.

4.2.1 Deictic Pronouns

We made a user study with our household robot where the users interacted with the robot via speech and gestures. They were told that they can use pointing gestures and we found in about 10% of the sentences pointing gestures coupled with deictic pronouns (see table 1).

Total Number of Turns	1151
Turns with Deictic Pronouns	125
Deictic Pronoun Rate (in %)	10.86

Table 1: Number of turns with deictic pronouns in an experiment with our household robot

For resolving deictic pronouns, we assume that a referring pointing gesture is available at the same time, as you can see in the second example of figure 4. We use a gesture recognizer and multimodal parsing of speech and gestures so that the information from both input modalities is merged on a semantic base by means of time stamps (Gieselmann and Denecke, 2003).

Therefore, gesture input is resolved by means of the context model which consists of different objects in the kitchen, such as cups, dishes, forks, knifes, spoons and lamps. An n-best list with all the pointing gestures matching a possible target object from the context model is created. The disambiguation is then performed by merging speech and gesture in a multimodal parsing process (Stiefelhagen et al., 2004). Deictic pronouns without a referring gesture cannot be resolved at the moment.

4.2.2 Personal Pronouns

In another small user study, where the users had to make the robot set the table, we found in about 6% of the sentences personal pronouns (see table 2).

By means of the context model, personal pronouns can be resolved, as you can see in the first

```
User: Robbi, get the blue cup from the board.
Robbi: Going to take the blue cup from the board.
User: Bring it to me.
Robbi: Going to bring you the blue cup.

User: Switch on that light. + pointing gesture to the big lamp
Robbi: Switching on the big lamp.
```

Figure 4: Example Dialogue taken from our user studies with a household robot

Total Number of Turns	572
Turns with Personal Pronouns	37
Personal Pronoun Rate (in %)	6.47

Table 2: Number of turns with personal pronouns in an experiment with our household robot

example in figure 4 in two different ways:

- out of the dialogue context taking into account information from the previous sentences
- out of the situation. This means that some kind of simple world knowledge is used. For example, if the robot has a cup in its possession, and the user tells it "Put *it* there", then it can be assumed that "it" refers to this cup.

Therefore, there are two different ways how pronouns can be resolved. On one hand, the information on what can be found in the robot's possession is in the context model and can therefore be used for the resolution. In this way, pronouns can be simply resolved by replacing the pronoun by the object in the robot's possession.

On the other hand, we use our list of possible antecedents in the context model and look there whether there is a possible antecedent. Similar to the pronoun resolution mechanisms mentioned above, we also use some rules, such as that the pronoun and the antecedent have to agree in their syntactic and semantic features. This means that they have to have the same number and gender as far as syntax is concerned and both of them have to belong to the same category or a subcategory in the ontology, as far as semantic is concerned. Since the antecedents are ranked by their appearance and also deleted, if they are too old, we can use the first possible antecedent which is found, and put its semantic representation in the discourse.

Both methods are not very complex, but work efficiently in our scenario so that about 90% of the pronouns can be resolved. In our user study even all the pronouns can be resolved just out of the situation by means of the world knowledge in the context model. Therefore, we do not even need the more complex mechanism with all the possible antecedents in the context model. But since this might also be due to the fact that the scenario is quite simple at the moment, we will test this with an enhanced version in a more complex scenario.

Also a combination of both methods sounds promising. Namely, there are situations where the method based only on the previous sentences will fail because the previously mentioned correct antecedent is too many sentences away and cannot be found therefore. On the other hand, also the method of just using the information what is in the robot's possession can fail easily, if the robot has something else than the user is referring to. Therefore, we want to do further experiments with a combination of both methods to see whether we can resolve even more pronouns by this combination.

5 Conclusion and Outlook

In this paper, we developed some methods for reference resolution in human robot communication. We focused our attention on the pragmatic aspects of the resolution and started with personal and deictic pronouns. Both of them are resolved by means of the context model.

In our user studies, we found out that it was possible to resolve the personal pronouns just by taking into account the current situation without us-

ing any knowledge of the previous sentences. For the future, we want to evaluate whether this is also feasible in more complex situations which would facilitate reference resolution a lot.

Furthermore, we also want to evaluate whether a combination of the two mentioned methods leads to better results and how these methods can be efficiently combined to take advantage from both of them while avoiding their disadvantages.

Acknowledgments

This work was supported in part by the German Research Foundation (DFG) as part of the SFB 588 and within the FAME project by the European Union as IST-2000-28323.

References

James Allen, George Ferguson, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski Zollo. 2000. Dialogue systems: From theory to practice in trains-96. *Robert Dale, Hermann Moisl, and Harold Somers, eds.: Handbook of Natural Language Processing*, pages 347–376.

Breck Baldwin. 1995. *CogNIAC: A High Precision Pronoun Resolution Engine*. University of Pennsylvania Department of Computer and Information Sciences Ph.D. Thesis, Pennsylvania, US.

E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. *Proceedings of the 25th Annual Meeting of the Association of Computational Linguistics*, pages 155–162.

Bob Carpenter. 1992. The logic of typed feature structures.

H. H. Clark. 1978. On inferring what is meant. *W. J. M. Levelt and G. B. Flores d'Arcais (Eds.). Studies in the Perception of Language*, pages 295–322.

Matthias Denecke. 2002. Rapid prototyping for spoken dialogue systems. *Proceedings of the 19th International Conference on Computational Linguistics*.

Edward Filisko and Stephanie Seneff. 2003. A context resolution server for the galaxy conversational systems. *Proceedings of the Eurospeech*.

Petra Gieselmann and Matthias Denecke. 2003. Towards multimodal interaction with an intelligent room. *Proceedings of the Eurospeech*.

A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright. 2002. Automated natural spoken dialog. *IEEE Computer Magazine*, 35(4):51–56.

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

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Jerry R. Hobbs. 1977. Resolving pronoun references. *Lingua*, 44:311–338.

Ashwani Kumar, Susanne Salmon-Alt, and Laurent Romary. 2003. Reference resolution as a facilitating process towards robust multimodal dialogue management: A cognitive grammar approach. *International Symposium on Reference Resolution and Its Application to Question Answering and Summarization*.

F. Landragin and L. Romary. 2003. Referring to objects through sub-contexts in multimodal human-computer interaction. *Seventh Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck'03)*, pages 67–74.

Michael F. McTear. 2002. Spoken dialogue technology: Enabling the conversational interface. *ACM Computing Surveys*, 34(1):90–169.

David Stallard. 2000. Talk'n'travel: A conversational system for air travel planning. *Proceedings of the Association for Computational Linguistics 6th Applied Natural Language Processing Conference (ANLP 2000)*, pages 68–75.

R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 2004. Natural human-robot interaction using speech, gaze and gestures. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*.

Michael Strube. 1998. Never look back: An alternative to centering. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1251–1257.

Joel Tetreault and James Allen. 2003. An empirical evaluation of pronoun resolution and clausal structure. *Proceedings of the 2003 International Symposium on Reference Resolution and its Applications to Question Answering and Summarization*, pages 1–8.

Marilyn A. Walker. 1998. Centering, anaphora resolution, and discourse structure. *Marilyn A. Walker, Aravind K. Joshi and Ellen F. Prince: Centering in Discourse*.

Victor Zue, Stephanie Seneff, James Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, and Lee Hetherington. 2000. Jupiter: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):100–112.

Alignment in Dialogue: Effects of Visual versus Verbal-feedback

Kerstin Hadelich
Department of Psycholinguistics
Saarland University
Germany
hadelich@coli.uni-sb.de

Holly P. Branigan
Department of Psychology
Edinburgh University
U.K.
holly.branigan@ed.ac.uk

Martin J. Pickering
Department of Psychology
Edinburgh University
U.K.
martin.pickering@ed.ac.uk

Matthew W. Crocker
Department of Psycholinguistics
Saarland University
Germany
crocker@coli.uni-sb.de

Abstract

It has been shown that restrictions on feedback in communicative tasks have an important impact on how speakers ground their communicative acts and the effectiveness of their communication. Generally speaking, the more interlocutors are allowed to interact, the quicker they solve communicative tasks, and the quicker they converge at a linguistic level on referring expressions for objects under discussion. Whereas the effects of verbal feedback have so far been mainly investigated with respect to linguistic measures, the effects of non-verbal feedback have been thought of as mainly influencing a more affective component or the outcome (efficiency) of communication. However, recent research has shown that visual-feedback (in terms of a shared work space) also has an effect on the smoothness and effectiveness of linguistic communication.

In our study we investigated the different effects of visual and verbal-feedback on alignment in a communicative task. In addition to commonly used measurements like the number of words of referring expressions, we also computed the lexical overlap of subsequent descriptions. We found that visual feedback also has effects on linguistic measures, and that differences in communication related to visual and verbal feedback do not necessarily show up in relatively superficial measurements such as number of words per turn.

1 Introduction

In investigating human communication, mostly task-oriented dialogues have been used. These offer the advantage that, on the one hand, participants are free to talk, but on the other hand, topic and goals of the communication are constrained by the specific task at hand. A wide variety of experiments on task-oriented dialogues have been carried out. One important issue that has been addressed is the difference that modalities used by either the speaker or the listener make on communication. In order to tackle these differences, these tasks have been conducted using different feedback conditions as variables.

One line of investigation focuses on the effects of different (verbal) task conditions on linguistic parameters. For example, the issue of coordination in the making of mutually agreeable references was addressed in a number of studies (e.g. Anderson et al., 1991; Boyle, Anderson and Newland, 1994; Clark and Wilkes-Gibbs, 1986; Clark and Krych, 2004; Horton and Keysar, 1996; Krauss and Weinheimer, 1964; Schober and Clark, 1989; ...). Clark and Wilkes-Gibbs (1986), for instance, conducted an experiment in which two participants had to arrange a set of abstract shapes (i.e., tangrams) in a linear order. One of the two participants was asked to give instructions in form of descriptions whereas the second participant was the listener who sorted the tangrams. The shapes were abstract in order to induce negotiations of names for the figure under discussion. The degree of feedback was manipulated reaching from full verbal-feedback to no-feedback. Clark and Wilkes-Gibbs measured the effects of the different feedback conditions, for example, in terms of the number of words used per referring expression.

Another line of investigation deals with the effects of visual-feedback on communication. Visual-feedback is thereby addressed either in terms of the effects of visual contact, i.e. mutual gaze or a shared visual scene, or the effect of the transmitting channel (e.g., Boyle, Anderson and Newland, 1994; Anderson, 2004; Clark and Krych, 2004; De Ruiter et al. 2003; Drolet and Morris, 2000; ...). De Ruiter et al. (2003), for example, had subjects perform a communication task in the *spatial logistics task* (SLOT), a psycholinguistic version of the so-called *social dilemma scenario*. In SLOT two participants have to negotiate a route through a map that meets certain optimisation criteria. In their experiment, the visual information of the scene was shared across all conditions. De Ruiter et al. looked at the effects of the presence and absence of eye contact on the outcome of the task. In one condition they used a one-way mirror that only allowed asymmetric visual contact. De Ruiter et al. found that in this condition negotiation times increased significantly but the successful outcome of the task was not affected. This result is consistent with findings in earlier work (e.g., Drolet and Morris, 2000). Remarkably in this respect, Anderson (2004) reports that in one map task experiment (Anderson et al., 1991) only 30 % of words were actually uttered in the time span of mutual gaze.

Most of these problem-solving tasks are asymmetric by virtue of the way in which roles are assigned to (the two) interlocutors, such as instruction giver versus instruction receiver. Even though this design reveals obvious disadvantages when it comes to the generalisation of results, it nonetheless appears to be an approach that satisfies many of the relevant constraints.

Taken together, the evidence suggests that visual-feedback affects the way in which participants solve a task in dialogue. But apart from more general measures like efficiency and affective components (e.g. rapport), it remains unclear what influence different feedback modalities have on the linguistic dimensions of dialogue. In our study, we compared the effects of visual-feedback (shared visual information about the scene but no eye contact) versus verbal-feedback on linguistic measures of communicative success. In measuring the linguistic effects, we use the concept of alignment (Pickering and Garrod, in press) and analyse, for example, the lexical overlap in subsequent utterances.

2 Experiment

We tested 32 Edinburgh University students who received £5 each for taking part in the (30 - 60 minute) experiment. Participants were paired randomly and randomly assigned to one of the four experimental conditions.

2.1 General set-up

Participants were separated by a head-high divider. Each participant was seated in front of a monitor and given a separate mouse. Their task was to move a set of tangrams from an initial set of positions into their final positions as indicated on a given target configuration. The two boards were identical and showed all eight tangrams. However, each participant had their own individual target card with four of the eight tangrams displayed on it. Both the board to play on and the target card were displayed on the monitor (see Figure 1). Participants were asked to take turns instructing each

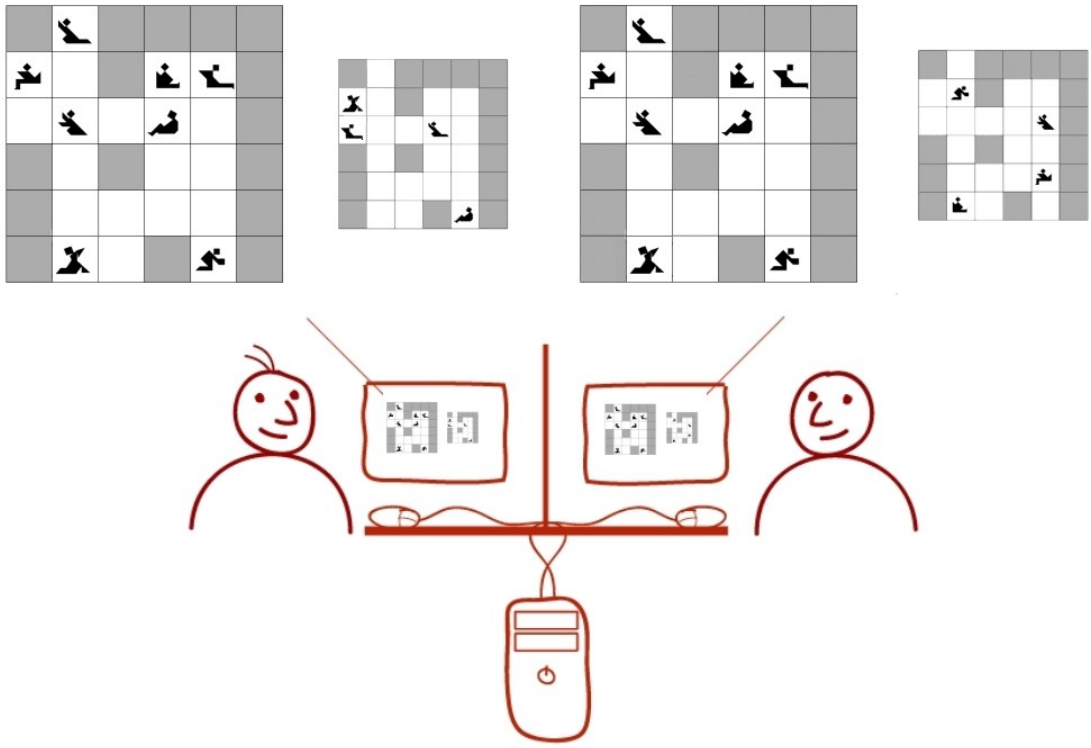


Figure 1: Schematic illustration of the experimental set-up. The board with all eight tangrams and, next to it, the target card showing the final positions of four of the tangrams as the two participants saw them on their screens.

other. This means that they alternately selected a tangram on their target card and gave instructions to the instruction receiver until this particular tangram had reached its final position. In doing so, other tangrams had potentially to be moved out of the way first. After the selected tangram had reached the target position, participants swapped roles. This sequence was repeated until all eight tangrams had reached their final positions and the target configuration was accomplished. Our aim was to approach the symmetric character of natural conversation by introducing a more dynamic role assignment. Also, the turns taken in giving instructions can be seen as an equivalent to subgoals in conversation.

Prior to running the experiment, participants completed a practice session illustrating the rules and technical features of the set-up. In this practice session we used geometric shapes instead of tangrams, to avoid giving the participants practice in the specific task.

2.2 Conditions

We varied the type of feedback that participants could give in a between-participants design. Each pair of subjects was randomly assigned to one of four conditions: *full-feedback*, *verbal-feedback*, *visual-feedback*, and *no-feedback*.

In the full-feedback condition, we allowed participants to talk freely; additionally the two monitors were connected, so that the instruction giver also could see on their screen which of the items the instruction receiver was moving, and to which position. In the only-verbal-feedback condition, participants were also allowed to talk freely. But this time their monitors were not connected, so they did not get any information about which item their partner was moving. In the only-visual-feedback condition, the instruction receiver could not give any verbal feedback, but participants

could again see what their partner was doing on their screen. Finally, in the no-feedback condition, the instruction receiver could not give any verbal feedback and the participants' monitors were also not connected.

2.3 Hypotheses

Generally, alignment takes place most effectively by the use of same channels in interaction and shows up in same representations used by interlocutors (Pickering and Garrod, in press). Participants are thus expected to prime each other in the use of, e.g., lexical items. This effect should be stronger when interlocutors are allowed to interact verbally as opposed to a more passive participation in the communication when listening to instructions. We thus expected a greater reduction of number of words and greater lexical overlap in subsequent descriptions in verbal-feedback conditions. This advantage should result in fewer disfluencies in the verbal-feedback conditions.

2.4 Analysis

We identified the first phrase of each referring expression that was delimited by intonational phrase boundaries. We analysed the number of words in a phrase in order to measure the process of convergence and additionally looked at the number of disfluencies (e.g., filled pauses, such as *uh* and *uhm*). In the conditions without verbal-feedback (i.e., visual-only and no-feedback) the descriptions could not be interrupted by the listener and thus tended to be much longer than the verbally more interactive conditions. In cutting down the descriptions into smaller, phrasal units, we increased comparability of the utterances across conditions. We also computed lexical overlap of subsequent descriptions. The *relative lexical overlap* for a description *k* was calculated by relating the number of lemmas in description *k* shared with description *k-1* to the total number of lemmas in descriptions *k* + *k-1*. As in the first description of an item in conversation the preceding description is missing, we only included descriptions two, three, and four in the analyses to compute lexical overlap.

2.5 Results

We conducted univariate ANOVAs and paired comparisons (Scheffé Posthoc Test) with participants as random factors and disfluencies, number of words used in the first phrase, and lexical overlap as dependent measures. Overall, the visual-feedback conditions differed from the verbal-feedback conditions with respect to disfluencies and relative lexical overlap, but not with respect to length of the description.

The results showed a significant main effect of condition on number of disfluencies in the referring expressions ($F(3, 1509) = 73.022$; $p < .005$). The paired comparisons revealed an effect of feedback modality: Conditions without verbal feedback (visual and no-feedback) showed significantly more disfluencies than the two conditions with verbal-feedback (full and verbal-feedback). Additionally, the absolute number of words used for the first phrase in a description also showed significant effects of condition type ($F(3, 650) = 44.996$; $p = .005$). Here, the posthoc tests revealed that in the no-feedback condition significantly fewer words per phrase were used than the other three feedback conditions (see Figure 2). The third dependent measure, relative lexical overlap, also showed significant effects of condition type ($F(3, 647) = 14.388$; $p < .05$). As in the analysis of disfluencies, there was again a significant effect of feedback modality. But this time the effect was inverted: In the two verbal-feedback conditions, referring expressions shared significantly fewer lemmas with their preceding utterance than the two conditions without verbal feedback.

3 Conclusion

The data provide only partial support for the hypothesis that verbal-feedback is more effective for alignment than visual-feedback. With respect to fluency, verbal feedback turned out to have the expected effects, i.e. in conditions with verbal feedback, utterances were more fluent than those in conditions without verbal feedback. The second commonly used measurement in the analyses of dialogue, length of referring expressions, did not reveal differences of verbal versus visual feedback. Only in the no-feedback condition, in

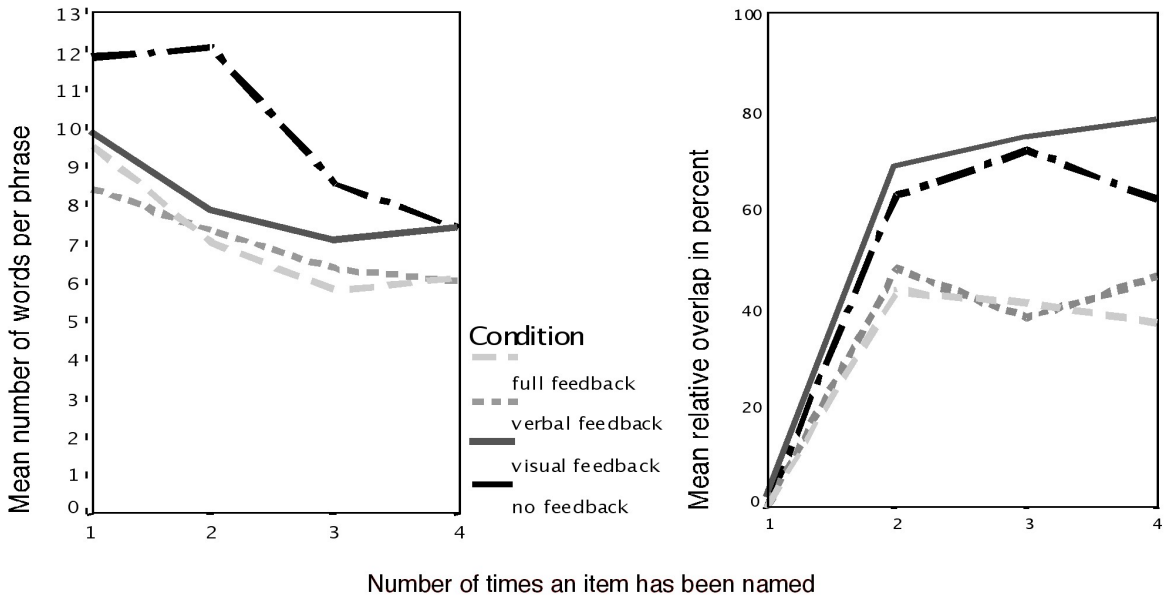


Figure 2: Mean number of words per phrase and mean relative overlap per phrase relative to the number of times an item has been named.

which participants could neither see what their partner was doing nor negotiate names for an item to be moved, did participants produce significantly longer referring expressions than in the three other feedback conditions. Noteworthy at this point is that, obviously, in task-oriented dialogues as the one described above, the type of feedback does not seem to make a difference with respect to the length of the first phrase. The more important factor appears to be the actual possibility of having feedback in communication, be it visual or verbal. Moreover, the fact that the conditions with verbal-only and visual-only feedback are not significantly different from the full feedback condition suggests that the communicative benefit on the first phrase is not larger with an increase of feedback. However, the predicted difference between visual and verbal-feedback did show up in the measure of relative lexical overlap that we computed for subsequent descriptions. Here we found that there was less overlap in the two verbal-feedback conditions than in the visual or the no-feedback condition. To some extent, this is surprising as the assumptions drawn on the basis of the alignment model pointed into the opposite direction. One way to interpret these results is to consider the overlap showing up in the verbal-feedback conditions as the automatic portion of overlap and the additional overlap in the visual-feedback conditions as stemming from other origins, such as pragmatic or situational influences or an aspect of audience design. In conditions without verbal feedback, participants have to make sure that their descriptions are understandable. This is even more the case in the no-feedback condition, because misunderstandings are much more difficult to resolve. In order to make sure that referring expressions are understandable, an appropriate strategy can be to reuse successful lemmas, which leads to a relatively big overlap. The interactive character of the verbal feedback conditions, however, offered instruction receivers the possibility to actively take part in the process of finding a name for items under discussion. Thus, subsequent descriptions in the verbal feedback conditions don't necessarily have to be driven by an automatic tendency to

align on a name, but could also show effects of this collaboration.

Taken together, we have shown that visual feedback obviously has effects not only on more general and affective components of communication but also on linguistic measures such as the number of words used in a referring expression and lexical overlap. This further highlights the fact that differences between visual and verbal-feedback are not revealed in relatively superficial measures such as the number of words and require more fine-grained measures such as degree of lexical overlap.

Acknowledgements

We would like to thank Alissa Melinger and Andrea Weber for helpful comments on the experimental design and an earlier draft of this paper. The presentational software for the study described was programmed by Alan Marshall (Edinburgh University).

References

A.H. Anderson, M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, and H.S. Thompson. 1991. The HCRC Map Task Corpus *Language & Speech*, 34:351–366.

A.H. Anderson. 2004. Feedback in problem solving dialogue. Talk given on the Multi-Modal Interaction Projects Feedback in Interaction Workshop, MPI Nijmegen.

E.A. Boyle, A.H. Anderson, and A. Newlands. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language & Speech*, 37(1):1–10.

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

H.H. Clark and M.A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

J.P. De Ruiter, S. Rossigno, L. Vuurpijl, D.W. Cunningham, and W.J.M. Levelt. 2003. SLOT: A research platform for investigating multimodal communication. *Behavior Research Methods, Instruments, and Computers*, 35(3):408–419.

M.J. Pickering and S. Garrod. in press. Toward a mechanistic Psychology of Dialogue. *Behavioral and Brain Sciences*

G. Doherty-Sneedon, A.H. Anderson, C. O'Malley, S. Langton, S. Garrod, and V. Bruce. 1997. Face-to-Face and video mediated communication: A comparison of dialog structure and task performance. *Journal of Experimental Psychology*, 3, 2:105–123.

A.L. Drolet and M.W. Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed motive conflicts. *Journal of Experimental Social Psychology*, 36:26–50.

W.S. Horton and B. Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59(1):91–117.

R.M. Krauss and S. Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114.

M.F. Schober and H.H. Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21:211–232.

Dialogue History Modelling for Multimodal Human-Computer Interaction

Frédéric Landragin and **Laurent Romary**
LORIA – UMR 7503
Campus scientifique – B.P. 239
F-54506 Vandœuvre-lès-Nancy Cedex – France
{landragi,romary}@loria.fr

Abstract

The design of multimodal dialogue systems requires a particular attention on the way of managing dynamic heterogeneous information. We present a theoretical model of a multimodal dialogue history, that includes a global history and local histories linked to the various modalities. We describe the nature of these components and the relations they entertain. To save the information and to exploit and confront them during the interpretation of an utterance from the user, we need an unified representation format. We developed the MMIL (MultiModal Interface Language) model that we present here with two main aspects: the representation of a simple utterance and the representation of a dialogue structure. Then, we draw some conclusions concerning the exploitation of this framework in the OZONE system, with an interest on the management of attentional scores inside the dialogue history.

1 Introduction

A dialogue history is always needed in oral understanding systems, for example to resolve ellipses and anaphors. The interpretation of an utterance can exploit the previous utterances that are saved in this history. When the task includes objects that can be referred to, this history must keep the objects identity along with the expressed utterances. Problems can arise in multimodal systems, where gesture and speech are combined and linked to the visual perception of the displayed scene. What type of information has to be saved to resolve references to objects? Are there separate histories for visual perception, gesture, speech, referents, reference domains? How can a history take forgetting phenomena into account? In this paper, we propose some theoretical leads to these infrequently debated topics. The particular aspects we want to address are the following ones:

- Components of the dialogue history: what is a dialogue history and how is it shared considering the modalities in natural language and multimodal systems (visual perception history, linguistic history, etc.).
- Nature and roles of the global history: temporal marks and pointers to local histories, and, eventually, a model of attention and forgetting (i.e., the former information are forgotten step by step). First main problem: limits of histories. We cannot save all information that can help the reference resolution and the utterance interpretation (information structure, task marks, interaction marks, etc.). We need to save only the information that have been the object of a computation during a previous phase of the dialogue. Second main problem: a model of the user's attention may be useful. The idea is to identify a focused part of the histories, that corre-

sponds to the part which is at the moment in the mind of the user. Third main problem: we need a model of forgetting. Idea: when the user talks about a blue triangle, the concepts that are linked to "triangle" and to "blue" are activated. Their activation rates will then decrease step by step, as the time goes by (except if they are activated again, i.e., mentioned in the dialogue).

- Unified representation of information in the various histories: first through the systematic notion of reference domain (see Landragin & Romary, 2003), second by using a unified representation format (MMIL: MultiModal Interface Language) which we describe in this paper.
- Illustration and applications of the model we propose in the framework of the OZONE European project.

2 Components of the dialogue history

The main purpose of the linguistic history is to keep the trace of referring actions. The information to be saved are referring expressions, referents, and reference domains. Referring expressions have to be saved for a further exploitation of the referent accessibility. Referents have to be saved because of evolutive referents and objects deletions (see the interpretation of "it" in Example 1). Concerning reference domains, Example (2) shows the linguistic construction (using the coordinating conjunction "and") of a reference domain including two tables. This domain is kept in the history and constrains the resolution of a further referring expression. (2a) is then authorized, but not (2b). Another interest of reference domains is to propose a default set of objects for the interpretation of "other one" expressions. Reference domains are linked to each others, in order to model referential and anaphoric chains (Salmon-Alt, 2001).

(1) Remove the big desk. Replace it with a round table.
(2) Add a wooden table and a black plastic table.
(2a) Put the wooden table on the left and remove the other one.
(2b) *Put it on the left.

The visual history keeps the successive states of the scene. The information to be saved are the objects and their properties (including the coordinates), the perceptual groups, and their structuring (one group can include several groups). Visual salience and focus spaces (Beun & Cremers, 1998) are modelled into visual reference domains (Landragin, 2001). The visual history is necessary to face to situations such as (3). In this example, the interpretation of "underneath" needs a return to a previous state of the scene. As the pointing gesture was linked to the visual context, it was also kept in this history. Determining the position of the shelf uses a combination of the coordinates of the painting and of the gesture trajectory.

On the importance of the gestural part in the visuo-gestural history: when the user produces several times the same type of trajectory, it is more and more easy to interpret (even if they are more and more ambiguous, more and more imprecise, more and more quick, or less close to the target objects).

(3) Remove this painting (+ *gesture*). Add a shelf underneath.

The task history groups the performed actions, the referents to which they applied at the time, and links between these objects and the task's purpose and sub-purposes. Following (Grosz & Sidner, 1986), we take intentional structures into account, and we model them as task-linked reference domains. We show in a complex multimodal example (extracted from Ozkan corpus and described in details in the paper) how keeping these domains can be useful for the dialogue understanding.

3 Nature of the global history

One important point is that linguistic, visuo-gestural and task-linked reference domains are all structured in the same way. Consequently, domains can be confronted and integrated. That is a strong point of this model compared to heteroge-

neous modality-dependent theories. However, the global dialogue history does not correspond to the integration of all local histories, but includes the following elements: pointers to parts of them, results of the referring actions, *a posteriori* evaluations, system's reactions, and time stamps. Local histories can then be seen as specialized agents, and the global history as the coordinating agent.

The last problem deals with the storage capacity of the global history. With a cognitive concern, we can consider the history as a model of short-term memory, and limit its capacity to the seven most recent items, whatever the nature of these items (Miller, 1956). More recent works in psychology tend to limit this capacity to five or only four items (see Rousselet & Fabre-Thorpe, 2003). But since numerous items can be accessible, the working of the history appears to be much closer to long-term memory. The problem that arises is that all information has to be saved and is at the same level of accessibility. We prefer to consider the global history as a model of forgetting. The longer an entity has been the focus of user's attention, the more important it is to emphasize it in the history. We thus propose a methodology for tracking attentional scores, that apply to every object, category, event, or property:

- the more the user refers to an object, the greater the object's attentional score;
- the more he evokes a property (like a red colour), the greater the score of its related concept ("red");
- the more he performs an action, the greater the score of the corresponding software primitives (for example, a function or a class).

The score of an entity involves the scores of all its properties. These scores are managed so that the entities with the best scores are favoured during the interpretation process. Scores also (like human attention) decrease as the dialogue progresses. This approach has some common aspects with linguistic work like that of (Ariel, 1988) or (Lappin & Leass, 1994), but is more inspired by psycholinguistics. More precisely, we want to extend the principle of the Logogen Model (Morton, 1982). Our propositions concerning local and global histories seem to fit well such an aim. Though this is on-going research (no concrete algorithm has yet been implemented), the nature of information to be saved is already sufficiently precise to deduce the main characteristics of an algorithm and of a representation format.

4 Representing semantic content: MMIL

To represent the various histories, we need a unified representation format. The purpose is to confront in the global history information from different histories. This confrontation is only possible if the heterogeneous information are represented into similar structures. So we need a model to represent visual, gestural, linguistic and task-linked information in a same manner. For that, we use the MMIL model (MultiModal Interface Language) that was designed for the MIAMM European project and that was updated for the OZONE European project (see references).

The MMIL meta-model abstracts different levels of dialogue information (phone, word, phrase, utterance) by means of a flat ontology, which identifies shared concepts and constraints. The definition layer of the ontology includes two kinds of entities: events and participants. Events are objects associated to the temporal level, while participants are static entities acting upon or being affected by the events. Dependencies between entities are represented as typed relations linking structural nodes. Contrary to other semantic information models, the MMIL meta-model does not include relations, which are perceived as qualifying descriptors defining anchors among entities. As the other information units of the MMIL model (e.g., morpho-syntactic, domain, annotation descriptors), relations act in the information architecture as a set of descriptors (data categories) that formally describe the specification constraints. The data categories, expressed in an RDF format compatible with ISO 11179-3, give the necessary openness to the design of the semantic structures, so to cope with the potential flexibility of the model.

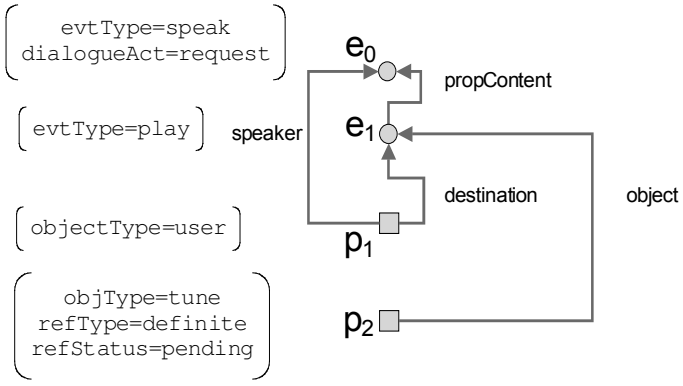
4.1 Simple utterance representation

One of the central purposes of MMIL is of course to be able to represent the actual semantic content of the various utterances processed by the linguistic modules in the MIAMM and Ozone architectures. To do so, we framed a core organization for MMIL structures at the output of linguistic modules that articulates:

- A specific event e0 that systematically represents the speech event associated with the utterance. Being categorized as such (/event type/=/speech event/), the event is anchored on temporal node that informs its beginning and ending date, and it may be further refined by various data categories corresponding to the /speaker/, /addressee/ and /speech act/;
- The actual (possibly underspecified) meaning of the utterance represented by an event e1 corresponding to the main predicate expressed by the utterance and which is related to e0 by a /propositional content/ relation. The event e1 is in turn associated with all the necessary descriptive elements such as its actual arguments, which are represented by one or more participants associated to it by the basic semantic roles (/agent/, /patient/, /location/, etc.) identified by the linguistic parser.

The fact that a full representation is provided for the speech event instead of just providing the corresponding propositional content offers several advantages that by far compensate the little extra complexity that it brings to the representation. First, it is an essential aspect in dialogue management (see below), to be able to relate an utterance to an other, and it may not be possible to make this boil down to the sole organization of contents. Second, it provides a clear and coherent background for personal and temporal deixis interpretation, which can be directly related to the information available at speech event level, rather taking up information maintained specifically to this purpose. Finally, it is an essential basis unifying references to the application domain and to the discourse proper. An utterance such as "Please repeat" can only be processed if a homogeneous treatment is made of speech event within the space of the various events expressed along the course of a dialogue.

As an example, Figure 1 shows a graphics summarizing the main component of the utterance "Play me the song", together with the corresponding full XML representation.




```
<mmilComponent>
  <event id="e0">
    <evtType>speak</evtType>
    <dialogueAct>request</dialogueAct>
    <speaker target="p1"/>
  </event>
  <event id="e1">
    <evtType>play</evtType>
  </event>
  <participant id= "p1">
    <objType>user</objType>
  </participant>
  <participant id= "p2">
    <objType>tune</objType>
    <refType>definite</refType>
    <refStatus>pending</refStatus>
  </participant>
  <relation source="e1" target="e0"
            type="propContent"/>
  <relation source=" p1" target="e1"
            type="destination"/>
  <relation source= "p2" target="e1"
            type="object"/>
</mmilComponent>
```

Figure 1. Graphical representation of the MMIL structure associated to the utterance "Play me the song!".

Furthermore, the inherent hierarchical structure of events and participants in the MMIL metamodel, together with the actual reification of the speech event in any language related MMIL representation, allows the representation of more complex phenomena related to the actual segmentation of the spoken input. Indeed, it may occur that what is considered as one single input at speech acquisition level may well be further segmented at speech recognition or parsing level

as several sub-utterance bearing, for instance, specific dialogue acts or propositional content. In those cases, the speech event is further subdivided into the necessary components, as exemplified in the following simplified representation for "No, to Paris" (Figure 2).

```
<mmilComponent>
  <event id="e0">
    <evtType>speak</evtType>
    <speaker target="p1"/>
    <event id="e0-1">
      <dialogueAct>reject</dialogueAct>
    </event>
    <event id="e0-2">
      <dialogueAct>inform</dialogueAct>
    </event>
  </event>
  ...
</mmilComponent>
```

Figure 2. Representation for "No, to Paris."

As can be seen, both the information related to the event type and speaker are factored out at the main event level, whereas the dialogue act information is, in this case, specifically attached to the sub-components of the utterance.

4.2 Dialogue structure in MMIL

As in any classical man-machine dialogue architecture, the dialogue manager is in charge of both contextualizing each utterance coming from lower level linguistic modules (e.g. by interpreting referring expressions) and maintaining the overall logics of the dialogue, through, among other thing, a proper management of the dialogue structure. As a matter of fact, dialogue structure is the result of combining sentence level information with higher principles of discourse organization, which also closely interacts with the actual semantic content of utterances, among which focusing information take some specific importance. Besides, dialogue management involves being able to put in relation the user's utterances with the actual decisions or actions (spoken feedback, information presentation, actions at application levels) taken by the dialogue manager itself (or its action planning component). This is again where the homogeneous background provided by MMIL offers a flexible way of dealing with those various phenomena, under the condition that some clear principles are stated to maintain the coherence across the various information sources, that is at user's input level, dialogue internal processing level, and discourse management level.

At user's input level, we have already mentioned that each utterance is initially represented by a speech event to which is attached the actual semantic content as understood at the parsing stage. The speech event is qualified by at most one dialogue act taken from a basic typology of six values, i.e. /opening/, /closing/, /inform/, /query/, /accept/, /reject/. Those values are considered as surface dialogue acts, as they result from the sole analysis of the utterance inner structure prior to any specific contextualization of the semantic content. They correspond to the core (and consensual…) values that can be observed from various systems or annotation schemes (e.g. DAMSL, HCRC, etc.) that have been around in the last decades. They also correspond (except for the obvious /opening/ and /closing/) to the basic dialogues acts identified in the work by FIPA on inter-agent communication, which, as we shall see, contributes to a more homogeneous treatment of events in our dialogue architecture. Appendix A provides the ISO 11179 conformant description of those 6 dialogue acts in the perspective of stabilizing those values within the man-machine dialogue community.

As a consequence, we can also represent each event occurring within the dialogue architecture proper as a MMIL event, which bears similar characteristics in common with users' speech events. Dialogue internal event are thus represented in MMIL by means of several core characteristics:

- They are typed according to a basic ontology of dialogue management action combining general purpose actions (e.g. spoken feedback to the user, graphical presentation of information) and application specific primitives (e.g. queries to an underlying database);
- Like any other MMIL event, they can contain a temporal anchor indicating either when the action has taken place or when it is to be taking place (in the case it is still in a pending state within the dialogue architecture);

- They are qualified by one of the following four dialogue acts: /inform/, /query/, /reject/, /accept/, the last two are acts being used to validate or invalidate (e.g. when the information is not available or a service is down) an initial inform or query from one module to another. It should be noticed here that in both the MIAMM and OZONE architecture, a clear difference is made between the technical management of communications flows between modules, as can be typically handled by SOAP mechanisms in the context of a web service based technological deployment, with the management of the exchanges between modules from a semantic point of view. For instance, the same kind of behavior is not to be expected when a module within the architecture is physically down and when it has the knowledge cannot deliver a certain service that has been asked to it;

- Such dialogue internal event can be generalized to be used to communicate to external processes that may provide services in relation to the dialogue underlying task.

The example of Figure 3 shows the simple representation of the master event associated to a query issued by the action planner to the MIAMM database.

```
<event id="e4">
  <evtType>database query</evtType>
  <dialogueAct>query</dialogueAct>
  <evtStatus>actuated</evtStatus>
  <tempSpan startPoint="2004-04
     -05T17:00:00" endPoint="2004-04
     -05T17:00:01"/>
</event>
```

Figure 3. Representation of a master event.

Finally, MMIL structures can be used, on the basis of what has thus been presented, to uniformly represent dialogue structures that have been construed at dialogue management level. Without entering into the details of the supporting arguments for doing this, it seems by far more appropriate to base dialogue structure representation on inter-event relations (discourse relations, when dealing with users' input) then try to infer deep dialogue acts from the user's utterances. In this context, each time the dialogue manager infers a connection between any two events in the course of dialogue, it can report about it (for instance to the action planner module) by means of simple MMIL structures combining those events. As an example, a basic acknowledgement by the user ("Fine for me") to a proposal by the system ("I have this song from the Beatles") will be reported by a MMIL structure such as follows (Figure 4).

```
<mmilComponent>
  <event id="e1">
    <evtType>speak</evtType>
    <speaker target="p2"/><!-- system-->
    <dialogueAct>inform</dialogueAct>
  </event>
  <event id="e2">
    <evtType>speak</evtType>
    <speaker target="p2"/><!-- system-->
    <dialogueAct>accept</dialogueAct>
  </event>
    ...
  <relation source="e2" target="e1"
     type="confirm"/>
</mmilComponent>
```

Figure 4. Representation of a link between two events.

4.3 Standardization in the domain of semantic content representation

The work we have conducted on the definition of the MMIL language can be seen as a kind of experiment to identify precise requirements on what a general framework for multimodal content representation. Those requirements should obviously go beyond what has been described in (Bunt & Romary, 2002), in order to identify classes of applications which bear enough features to be covered by one single model. Indeed, it may not be likely that the kind of representations needed for such applications as information extraction, named entity recognition, reference annotation, or the annotation of temporal structure will be based on exactly the same underlying structures. Still, it seems necessary that those various types of models do share a common semantics for any sub-structure they would share and even more for any elementary descriptor they would use (e.g., a certain dialogue act /inform/, or discourse relation /elaboration/, a temporal relation /overlap/, or an elementary role in relation to an event /agent/). Such a goal obviously requires that there is some kind of consensus on providing some shared definition of such concepts, as well as an international infrastructure to submit, select and

disseminate those descriptors. The first aspect is one of the topics which has been considered as underlying the activity of the ACL/SIGSEM working group on multimodal semantic content representation and is being pursued through a series of meetings that have taken place since November 2002.

The second aspect is the core of a standardizing effort in ISO committee TC 37 to deploy an online data category registry intended to cover a wide variety of descriptors (also known as *data categories* in the TC 37 terminology) identified in existing representation or annotation practices. In this context, we would like to see MMIL as one instance of such a descriptive and modelling activity which would nicely fit the needs of multimodal dialogue system when conveying meaning from one component to another, and when managing meaning inside a component (and particularly inside the dialogue history). If it is the case, we could also contemplate using MMIL—or a dialect thereof—for such tasks as the evaluation of dialogue systems.

5 Applications of the model

A first application of the attentional scores that we propose is a help during the resolution of verbal ellipses. For instance, in a man-machine dialogue system that consists in the interrogation of a music database (queries about authors and songs, and commands like "play my favourites", see MIAMM European project for the implementation of such a system), when the user asks several times to "play this tune", the attentional scores of "play" and of "tune" are maximal. Then, if the user produces the incomplete utterance "this tune", the system can easily infer that the requested action is "play". This method for exploiting attentional scores can also be used for the resolution of anaphoric expressions, like "play this one" after "play this tune".

A second application is to allow more spontaneous reactions of the dialogue system, by using recalls when some information may have been forgotten. For instance, in reservation tasks like the reservation of a train ticket or a room in a hotel (see OZONE European project), one of the purpose of the dialogue is to specify a number a parameters that allow to launch a reservation request. For a train ticket like in our OZONE's demonstrator, the parameters are the departure station, the destination, the way (including changes of train), and the time (of departure or arrival). For a room reservation, the parameters are the number of persons, the date of arrival, the date of departure, and some options (breakfast, etc.). In OZONE, the user can begin the dialogue with "I want to go to Paris", and then can ask for some information about the possible ways, their duration, the changes, etc. The resulting dialogue can be quite long, and the destination (Paris) that has only been mentioned once at the beginning can have a very low attentional score. Thus, the system may be aware of this low score and may produce a sentence like "Do you still want to go to Paris?" or "You confirm a train ticket to Paris, don't you?". This is important, not only because it reactivates the salience of the destination, but also because it adds a collaborative aspect to the dialogue. Even if the system has not forgotten the destination, it's important for it to show a human-like cognitive behaviour. Of course, experimentations have still to be done to determine if subjects who interact with such a system feel (or not) a kind of strangeness in the reactions of the system.

6 Conclusion and future work

For now, our participation to the IST-OZONE European project has consisted in the realization of a dialogue system demonstrator for a transport information service task. The research work we have described in this paper will allow us to improve this demonstrator and to provide a second system. For this improvement, we will focus (among other points) on the design of the dialogue history using attentional scores. The OZONE's application appears to be an efficient framework for that. As other future works, we want to test our model in generation systems, and for other modalities like written texts. To conclude, we want to show with this experience that designing multimodal systems with spontaneous communication abilities has to make the most of linguistic and psychological results, and that a crucial point in this design is the representation of information that is managed by the system during the dialogue.

References

Mira Ariel. 1988. Referring and Accessibility. *Journal of Linguistics*, 24:65-87.

Robbert-Jan Beun and Anita Cremers. 1998. Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition*, 6(1/2):121-152.

Harry Bunt and Laurent Romary. 2002. Towards Multimodal Content Representation. In K. Lee, K. and K. Choi (Eds.) *Proceedings of LREC 2002 Workshop on International Standards of Terminology and Linguistic Resources Management*, pp. 54-60.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175-204.

Frédéric Landragin. 2001. Visual Salience and Perceptual Grouping in Multimodal Interactivity. In: *First CLASS Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, pp. 151-155.

Frédéric Landragin and Laurent Romary. 2003. Referring to Objects Through Sub-Contexts in Multimodal Human-Computer Interaction. In: *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck'03)*, Wallerfangen, pp. 67-74.

Shalom Lappin and Herbert J. Leass. 1994. A Syntactically Based Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535-561.

MIAMM, Multidimensional Information Access using Multiple Modalities. IST-2000-29487 European Project. Website: http://www.miamm.org/.

John Morton. 1982. Disintegrating the Lexicon: An Information Processing Approach. In: J. Mehler, E. C. T. Walker and M. F. Garrett (Eds.) *On Mental Representation*. Erlbaum, Hillsdale, NJ, pp. 89-109.

OZONE (O_3), Offering an Open and Optimal roadmap towards consumer oriented ambient intelligence, IST-2000-30026 European Project. Website: http://www.extra.research.philips.com/euprojects/ozone/.

Guillaume A. Rousselet and Michèle Fabre-Thorpe. 2003. Les mécanismes de l'attention visuelle. *Psychologie Française*, 48(1):29-44.

Susanne Salmon-Alt. 2001. Reference Resolution within the Framework of Cognitive Grammar. In: *Proceedings of the Seventh International Colloquium on Cognitive Science* (ICCS'01), San Sebastián, Spain.

Appendix A: MMIL core dialogue acts

This annex describes the possible values of the **/dialogue Act/** data category. It is a selection of the dialogue act listed in the literature, which appears to us as sufficient for interpreting users' utterances in MIAMM (as opposed to annotation tasks, which would have probably required a more elaborate scheme).

Dialogue acts are usually described within an extensive hierarchy providing comprehensive groupings for them. It is not the intention of this simple typology to describe such a hierarchy, even if we have tried to organize thee list on the basis of some general dialogical categories. Further work within ISO/TC37/SC4 should incorporate this aspect more neatly.

Discourse management/Conventional acts:

/Opening/ **Def:** An utterance or segment establishing the communicative contact between a speaker and an addressee. **Note:** also know as 'Greet.'

/Closing/ **Def:** An utterance or segment finishing the communicative contact between a speaker and an addressee.

Initiative:

/inform/ **Def:** The speaker provides information to the user. **Note:** known under various names depending on the encoding scheme; Update (LINLIN), Explain (HCRC) Statement (DAMSL), Inform (TRAINS).

/request/ **Def:** The speaker aims to get the addressee to perform some action. **Note:** know under various names depending on the encoding scheme; Instruct (HCRC), Influencing Addressee Future Action (DAMSL), Action-directive (DAMSL), Request (TRAINS), Open-option (DAMSL), Suggest (TRAINS). No distinction is made here between Request and Suggest. Queries are represented as requests.

Response/Backward Looking Function:

/accept/ **Def:** The speaker agrees to all of the antecedent. **Note:** Corresponds to 'confirm_positive'. No distinction is made here between 'accept' and 'accept-part' (as in DAMSL).

/reject/ **Def:** The speaker disagrees with all of the antecedent. **Note:** Corresponds to 'confirm_negative'. No distinction is made here between 'reject' and 'reject-part' (as in DAMSL).

Context-sensitive speech recognition in ISU dialogue systems: results for the grammar switching approach

Oliver Lemon
School of Informatics
Edinburgh University
Edinburgh EH8 9LW
olemon@inf.ed.ac.uk

Abstract

This research explores how to employ context-sensitive speech recognition in a general way, in the Information-State Update (ISU) approach to dialogue management. The central idea is that different contexts, or dialogue "Information States", can be associated with different language models for speech recognition. In this paper a "grammar switching" approach is presented, based on "active" dialogue move types. It is then shown that this technique leads to more robust speech recognition. An evaluation of a dialogue system using this technique found that 87.9% of recognised utterances were recognised using a context-specific language model, resulting in an 11.5% reduction in the overall utterance recognition error rate, and a 13.4% reduction in concept error rate.

1 Context-sensitive speech recognition in dialogue systems

The basic idea of context-sensitive speech recognition is not new. Finite-state dialogue managers typically define a recognition language model (LM) at each state, and form-based managers often define a LM for each slot, as is commonly done in Voice XML for example. However, this is a laborious and unsystematic process since a designer must anticipate the likely range of user utterances at each point in the dialogue. It also often curtails the freedom of the speaker to say anything at any time in the conversation. In addition, these approaches are domain- and task-specific, and thus are not reusable. The approach presented here is to implement a similar idea more generally and systematically, within a richer (non-finite-state or form-based) model of dialogue context: the Information State Update (ISU) approach (Traum et al., 1999). The general method presented here could be used for a variety of applications, since it only depends on representing the dialogue move types of the user and system, and their dependencies, and not on any application-specific information.

The central idea is to use an "active move list" from dialogue Information States to define a changing search space of language models for speech recognition, to be used whenever the user speaks. We assume that at any point in the dialogue there is a "most active move" of some dialogue move type (for a full description of the system see Lemon and Gruenstein (2004)). In ISU systems generally this is typically the last uttered dialogue move, although there are cases, for example where a clarification subdialogue has just successfully closed, where another dialogue move should be chosen. We then define, for each move type, the name of a language model to be used for speech recognition if that is the type of the most active move. These LMs are defined by dialogue move type, rather than domain-specific slot-value types (e.g. *wh-answer* rather than, say, *city-name*). For instance, if the most active move is a *yes-no-question* then the appropriate language model is

defined by a small context-free grammar covering phrases such as "yes", "that's right", "okay", "negative", "maybe", and so on. We call this language model `[yn-answer]`.

In the experimental system, evaluated in the next section, the following LMs were implemented:

- `[full]`: generated by the whole grammar for the application.
- `[wh-answer]`: generated by a subgrammar consisting only of "wh-answer" forms such as "the office", "to the school".
- `[yn-answer]`: generated by a subgrammar consisting only of "yn-answer" forms such as "yeah", "that's right", and so on.
- `[alt-answer]`: generated by a subgrammar consisting only of "alt-answer" forms such as "now", "later", "do it later", and so on.
- `[no-answers]`: generated by the whole grammar minus all the "answer" forms.
- `[no-corrections-no-wh-answers]`: generated by the whole grammar minus "answer" forms and "correction" forms such as "I meant the office", "not the office the lab", and so on.

The dialogue move types were associated with different LMs as shown in the table of Figure 1[1]. This technique is a variant of "conversational games", also known as "dialogue games" (Carlson, 1983), and in the context of task-oriented dialogues, "discourse segments" (Grosz and Sidner, 1986). Such accounts rely on the observation that answers generally follow questions, commands are generally acknowledged, and so on, so that dialogues can be partially described as consisting of "adjacency pairs" of such dialogue moves. A statistical analysis of which dialogue move types typically follow each other could also be used (see e.g. Gabsdil and Lemon (2004)[2].

But what should happen in cases where the user produces an utterance which is not in the coverage of the currently active language model? For example, the currently active LM could be `[yn-answers]` but the user could produce a command. In cases where recognition fails with the currently active LM, there are several options:

- Reprocess the utterance using the LM related to the next most active move[3].
- Back-off to a "full" LM consisting of all the sentences recognizable for the application, and reprocess the utterance.

Due to the amount of time taken to perform another recognition pass (roughly proportional to the size of the LM) the second strategy is preferable, except for cases where the next most active node has a small associated LM. This is the technique used in the evaluation system. With current processor speeds, both these techniques are feasible. In fact, it is perfectly feasible to run the recognition processes in parallel, as was done in Hockey et al., (2003).

Figure 2 is an excerpt from a Nuance recognizer logfile, showing these dynamic language models in action. Here, the recognizer is in a context where it is using the LM `[no-answers]` (for instance after just uttering a report) but cannot recognize the user input ("maybe") which is an answer to an earlier system question (e.g. "is this the right car?"). So the system backs-off to the LM `[full]`, and then succeeds in recognizing the answer[4]. Then another user utterance arrives for recognition. In this context there are no active commands that could be corrected, and no open questions, so the recognizer uses `[no-corrections-no-answers]` and successfully recognizes the user command "zoom in on the car".

[1] "Root" is s special dialogue move type, used at the start of a dialogue, and in other contexts where there are no open questions or active commands.

[2] Here a feature "DMBigramFrequency", calculated from a corpus of dialogues with the system, is used to predict recognition performance, in combination with other features.

[3] This can be iterated, or performed to a certain depth of the active move list.

[4] The recognizer uses a new utterance label (65552) because it treats the back-off recognition pass as a new utterance.

DM Type	Language Model
command	[no-answers]
confirmation	[no-answers]
report	[no-answers]
wh-question	[wh-answer]
yn-question	[yn-answer]
alt-question	[alt-answer]
correction	n/a
yn-answer	n/a
wh-answer	n/a
root	[no-corrections-no-answers]

Figure 1: Language Models associated with Dialogue Move Types

```
started utterance 65550 with grammar .UTTERANCE-no-answers
Result #0:     <rejected> (conf: 37, NL conf:  0)

started utterance 65552 with grammar .UTTERANCE-full
Result #0:     maybe (conf: 81, NL conf: 81)

started utterance 65554 with grammar .UTTERANCE-no-corrections-no-answers
Result #0:     zoom in on the car (conf: 50, NL conf: 47)
```

Figure 2: Excerpt from a Nuance Logfile, showing Context-sensitive Speech Recognition

1.1 Defining suitable Language Models

It might be thought that the process of constructing the required multiple language-models is laborious and time-consuming. However, Gemini, SRI's system for developing bi-directional unification grammars (Dowding et al., 1993), makes this process quite simple. Gemini can be used for parsing and generation, and grammars can be compiled to language models for the Nuance speech recognition system. Similar systems (Bos, 2002; Rayner et al., 2003) are also in development.

Every Gemini grammar rule can be given a feature which is the name of the subgrammar (if any) that it belongs to. When the unification grammar is compiled to its context-free version (Dowding et al., 2001), these subgrammars are preserved, and the Nuance language model compilation process also preserves these named language models. This means that all that is required is to define the subgrammars in the top-level unification grammar formalism. A more laborious alternative is to partition the context-free grammar (Nuance GSL in this case) by hand before compilation. Since the partitioning is to be done by dialogue move type, this is still more general and less labour and maintenance-intensive than finite-state or form-based approaches, which mix together task and dialogue representations.

2 Evaluation

The technique described above was implemented in the WITAS dialogue system (Lemon et al., 2002). Seven members of the University community volunteered to use the system to complete a total of 35 tasks. There were both male and female subjects, all in their twenties or thirties. The subjects were given minimal written instruction on how to use the system before the interaction began. They were then asked to use the system to complete five tasks, in which they directed a simulated robot helicopter to move within a city environment. An example task is "There are reports of a fire at the tower. Check it out and fight the fire if you find one. Then fly the helicopter to the warehouse". Each task was given immediately prior to the start of the interaction, in language the system could not process to prevent users from sim-

ply reading the tasks aloud to the system. A given task ended when the user indicated to the system that he or she had finished, or they indicated that they had given up on the task. The system was run in open-microphone mode.

With the context-sensitive recognition system in use, subjects' speech was recorded and the system behaviour logged for each of the five tasks. Data was collected regarding task completion time, steps to completion, and speech recognition error rates. All dialogues were recorded, and the Information States logged as HTML files. The data thus consists of 35 tasks, resulting in 362 user turns, and 731 recognised words (as counted using Nuance batch-recognition). Of utterances which were recognised (at all), 87.9% were recognised using a context-specific language model, with the remainder being handled by backing-off to the full language model when recognition with the context-specific language model had failed to produce any result.

Each subject's speech data was then batch-recognized, without access to dialogue context information, using the full language model for the domain (call this the "normal case"), and the resulting statistics and recognition logs were compared to those from the context-sensitive recognition case. The Nuance batch recognition process effectively simulates (for the purposes of determining speech recognition performance) the performance of the system without the context-sensitive recognizer. We used the same recognition parameters in both cases (i.e. beam width, pruning, etc.).

The performance of the context-sensitive recognition system was evaluated in two ways: overall percentage of utterances recognized and concept accuracy of the recognized utterances (see section 2.2).

2.1 Overall recognition performance

The percentage of utterances recognized in the context-sensitive recognition case was 82.4%, while it was 80.2% in the normal case. Using a paired samples t-test this 2.2% difference between the overall utterance recognition rates in the two samples (number of utterances recognized per subject in the context-sensitive case compared with the normal case) was found to be significant $(t = 2.75, df = 6, p < 0.05)$. The reduction in overall recognition error rate was 11.5%.

Note that the context-sensitive system as implemented here cannot actually perform more poorly than the normal case in terms of number of recognized utterances, due to the fact that it backs-off to the full grammar should its first recognition attempt fail. In such cases the context-sensitive system will be slower than the normal system, but it is faster in the cases where the first recognition attempt succeeds[5] (since a smaller, faster LM is used), so a further study is needed to determine the speed/accuracy trade-offs here.

Note that the context-sensitive case can perform more poorly in the sense of "jumping to conclusions" based on a limited language model (see examples below), so we also need to determine the accuracy of the recognized utterances in each case. For this reason we also evaluate the concept accuracy of the system.

2.2 Concept accuracy

Rather than simply knowing that more utterances are recognized using context-sensitive recognition, we wish to know whether they are recognised correctly, and whether they lead to the correct system actions. It might be the case that context-sensitive recognition indeed recognises more utterances, but recognises them incorrectly, possibly harming overall system performance.

There are several important cases here:

- the user's utterance is recognized correctly by the context-sensitive system, but is not recognized at all by the normal system (e.g. Figure 3, rows 1-4),

- the user's utterance is recognized differently in the normal and context-sensitive cases. The recognition hypothesis in the context-sensitive case is correct (or partially correct) but incorrect for the normal case. Furthermore, the recognition hypothesis for the normal case does not give rise to the user's intended effect[6] (e.g. Figure 3, rows 5-7),

[5] As reported above, this was 87.9 % of the recognized utterances.

[6] There are cases where the normal recognizer output is

- the user's utterance is recognized differently in the normal and context-sensitive cases, and the recognized utterance in the context-sensitive case is incorrect and does not give rise to the user's intended effect, whereas the normal recognition hypothesis is correct or partially correct, or
- both recognition hypotheses are only partially correct, or
- both recognition hypotheses are completely incorrect, or
- there are no recognition hypotheses in either case.

An example of the first case is where the context-sensitive system recognized "fly to the tower" using the LM `[no-corrections-no-wh-answers]`, but the normal system rejected the utterance. An example of the second case is where the dialogue system has asked "Shall I fly to the building now or later?" (and so is subsequently using the LM `[alt-answer]` for recognition) and the user replies with "now" – which is correctly recognized using the context-sensitive system, but is recognized as "no" using the normal system, which would lead to an unintended action. An example of the third case is where the user said "forget about the house" and this was incorrectly recognised as "to the pond" by the context-sensitive system (using the LM `[wh-answer]`), but was correctly recognised using the full LM under batch recognition.

We used the concept accuracy measure of Boros et al., (1996) to compare the performance of the context-sensitive system with the normal system, in respect of each system's ability to correctly recognize user utterances. Concept accuracy is closely correlated with word accuracy, but allows that some word errors do not have a semantic effect (see Chotimongkol and Rudnicky (2001) for examples). Concept accuracy for each utterance is given by the following formula:

not absolutely correct in terms of word errors, but still leads to the user's intended effect (e.g. recognizing "yep" when the user said "yes" leads to the same action in this domain). Such word errors do not count against concept accuracy.

$$CA = 100\left(1 - \frac{SU_s + SU_i + SU_d}{SU}\right)\%$$

– where SU is the total number of semantic units in the reference answer (i.e. the logical form of the utterance were it recognised correctly), and SU_s, SU_i, and SU_d are the number of semantic units that must be substituted, inserted, or deleted respectively, to correct the actual parser output for the recognised utterance.

Examples of cases where the normal system (in the first column) suffers a concept error that the the context-sensitive system (second column) avoids are shown in Figure 3.

The average concept accuracy in the context-sensitive recognition case was 68.9%, while it was 64.1% in the normal case. Using a paired samples t-test this 4.8% difference in concept accuracy between the two samples was found to be significant ($t = 2.58, df = 6, p < 0.05$). The reduction in concept error rate was 13.4%.

2.3 Related work

SRI's CommandTalk system (Stent et al., 1999) used a related technique which:

> "used a main grammar (for full sentences), and a second grammar that had full sentences plus isolated NPs. If the system asked a question that could be answered with an isolated NP, then the larger grammar would be activated. The idea was that users were not forced to answer the question, since they had the complete sentence grammar available too." (John Dowding, personal communication).

Note that this technique was adopted to handle isolated NPs occurring as wh-answers, because they were not covered in the full CommandTalk grammar. In the WITAS dialogue system, such isolated NPs are legal utterances in the full grammar, so the problem is not how to include them in the context of a wh-question, but how to exclude them when there is no active wh-question.

Recognition with full LM	Context-sensitive recognition	Context-sensitive Language Model	Concept Acc % for full LM
< rejected >	that's it	[yn-answer]	0
< rejected >	and follow a truck	[no-answers]	0
< rejected >	fly to the power station	[no-corrections-no-answers]	0
< rejected >	where are you	[no-corrections-no-answers]	0
no	now	[alt-answer]	0
and the tower	and stop	[no-answers]	0
to the tower	go to the tower	[no-corrections-no-answers]	33.3

Figure 3: Examples of Recognition Hypotheses occurring in the Evaluation Study

3 Conclusion

Speech recognition performance in ISU dialogue systems can be improved by the use of context-sensitive recognition, using a grammar-switching approach based on dialogue move types. Both overall recognition error rates and concept error rates are significantly improved (11.5% and 13.4% reductions respectively) using a general technique which is less labour-intensive and easier to maintain than finite-state or form-based approaches, which mix together domain-specific and dialogue-general representations. A key idea is to define grammars and language models at the more abstract level of dialogue move type (e.g. *wh-answer*) rather than using application-specific slot-filler types (e.g. *destination-city*).

Future work will explore more advanced techniques for determining the correct LM to use in a particular dialogue context – for example the use of machine learning methods (Gabsdil and Lemon, 2004). Further investigation of such techniques is planned in the TALK project[7], see e.g. Lemon and Henderson (2004).

Acknowledgements

With thanks to Alex Gruenstein, Laura Hiatt, and Mark Core. This work was partially funded by Scottish Enterprise under the Edinburgh-Stanford Link programme, and partially by the EC IST FP6 project 507802 "TALK". Thanks also to Nuance Communications Inc. for the use of their speech recognition software.

[7] http:///www.talk-project.org

References

M. Boros, W. Eckert, F. Gallwitz, G. Görz, G. Hanrieder, and H. Niemann. 1996. Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Proceedings ICSLP '96*, volume 2, pages 1009–1012, Philadelphia, PA.

Johan Bos. 2002. Compilation of unification grammars with compositional semantics to speech recognition packages. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 106–112.

Lauri Carlson. 1983. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel.

A. Chotimongkol and A. I. Rudnicky. 2001. N-best speech hypotheses reordering using linear regression. In *Proceedings of European Conference on Speech Communication and Technology (EuroSpeech 2001)*, pages 1829–1832.

John Dowding, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. 1993. GEMINI: a natural language system for spoken-language understanding. In *Proc. 31st Annual Meeting of the ACL*.

J. Dowding, B.A. Hockey, M. J. Gawron, and C. Culy. 2001. Practical issues in compiling typed unification grammars for speech recognition. In *Proceedings of the Thirty-Ninth Annual Meeting of the Association for Compuational Linguistics*.

Malte Gabsdil and Oliver Lemon. 2004. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of ACL-04*, page (to appear).

Barbara Grosz and Candace Sidner. 1986. Attentions, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Beth-Ann Hockey, Oliver Lemon, Ellen Campana, Laura Hiatt, Gregory Aist, Jim Hieronymus,

Alexander Gruenstein, and John Dowding. 2003. Targeted help for spoken dialogue systems: intelligent feedback improves naive users' performance. In *Proceedings of European Association for Computational Linguistics (EACL 03)*, pages 147 – 154.

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 (ACM TOCHI)*. (to appear).

Oliver Lemon and James Henderson. 2004. Machine learning and the Information State Update approach to dialogue management. In *Proceedings of the Joint AMI-PASCAL-IM2-M4 workshop*, page (to appear).

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.

Manny Rayner, Beth-Ann Hockey, and John Dowding. 2003. An open source environment for compiling typed unification grammars into speech recognisers. In *Proceedings of EACL 2003 (demonstrations)*, pages 223–226.

Amanda Stent, John Dowding, Jean Mark Gawron, Elizabeth Owen Bratt, and Robert Moore. 1999. The CommandTalk spoken dialogue system. In *Proceedings of the Thirty-Seventh Annual Meeting of the ACL*, pages 183–190, University of Maryland, College Park, MD. Association for Computational Linguistics.

David Traum, Johan Bos, Robin Cooper, Staffan Larsson, Ian Lewin, Colin Matheson, and Massimo Poesio. 1999. A Model of Dialogue Moves and Information State Revision. Technical Report D2.1, Trindi Project.

Statistical Support for the Study of Structures in Multi-Modal Dialogue: *Inter*-Rater Agreement and Synchronization

Andy Lücking **Hannes Rieser** **Jens Stegmann**
SFB 360 "Situated Artificial Communicators", B3
Bielefeld University
{andy.luecking|hannes.rieser|jens.stegmann}@uni-bielefeld.de

Abstract

We present a statistical approach to assess relations that hold among speech and pointing gestures in and between turns in task-oriented dialogue. The units quantified over are the time-stamps of the XML-based annotation of the digital video data. It was found that, on average, gesture strokes do not exceed, but are freely distributed over the time span of their linguistic affiliates. Further, the onset of the affiliate was observed to occur earlier than gesture initiation. Moreover, we found that gestures do obey certain appropriateness conditions and contribute semantic content ("gestures save words") as well. Gestures also seem to play a functional role wrt dialogue structure: There is evidence that gestures can contribute to the bundle of features making up a turn-taking signal. Some statistical results support a partitioning of the domain, which is also reflected in certain rating difficulties. However, our evaluation of the applied annotation scheme generally resulted in very good agreement.

1 Introduction

In ordinary face-to-face communication, people make use of both speech and non-verbal gesticulation. No reductive relationship holds between these modes of communication in either direction. This assumption is in accordance with empirical work, e. g. in psycholinguistics (McNeill, 1992, e. g.), as well as with philosophical considerations, mainly about reference and demonstration (Wittgenstein, 1958; Peirce, 1965). Hence, we take it as a truism that accounts of dialogue must be extended to include a treatment of gesture.

Empirical investigations of multi-modal dialogue comprising gesture and speech can pursue at least two interests: First, one wants to know how speech and pointing gestures are related to each other, especially whether the information from the auditory and from the visual channel synchronizes. Here the focus is on relations within individual dialogue moves. We call this '*intra*-move synchronization'. Secondly, a similar interest exists concerning pointing gestures and exchanges of turns, where the question is how speech and gesture of one speaker are related to the gestures and the speech of his addressee and *vice versa* ('*inter*-move synchronization'). Here the focus is on relations between different dialogue moves within one dialogue game.

The distinction between *intra*- and *inter*-move synchronization reflects different research lines that have been pursued in recent years. Psycholinguistics serves as an illustrative example here. One point of reference is the body of work in gesture studies that builds on McNeill (1992), whose main empirical focus is on the relationships holding among gestures and speech within utterance units. On the other hand, much current work in dialogue theory centers on issues that are intimately con-

nected with coordination among language users, e. g. building upon the *joint actions* framework of Clark (1996); but see also the notion of alignment in (Pickering and Garrod, in press).

Our investigation is based on original empirical studies. The task we set for our subjects involved the choice of referents from a restricted domain, see figure 1 and figure 2. They had to negotiate or to align reference using dialogue games of a certain type. In order to get results showing relations obtaining between gesture and speech in dialogue, we applied descriptive and analytical statistical methods to the time-based annotation stamps of suitable dialogue data. Such statistical analysis is pointless, of course, unless the employed annotation scheme isn't evaluated and confirmed to be reliable.

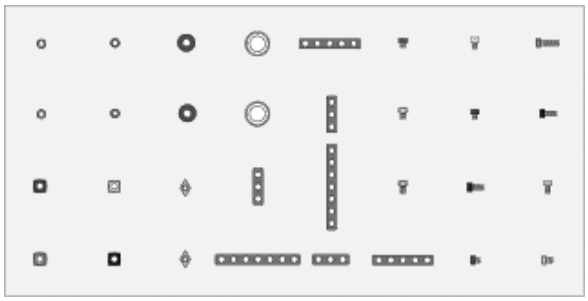


Figure 1: The pointing domain (form cluster), taken from (Kühnlein and Stegmann, 2003).

Accordingly, we present our study as follows: First, we set the stage with a description of the annotation of the empirical data (section 2). We then report on assessing of *inter*-rater agreement on our annotation scheme (section 3). In section 4 we present the results of further empirical investigation, mainly concerned with synchrony. We conclude the paper with a summary of our findings and a discussion of those topics that might be explored in further studies (section 5).

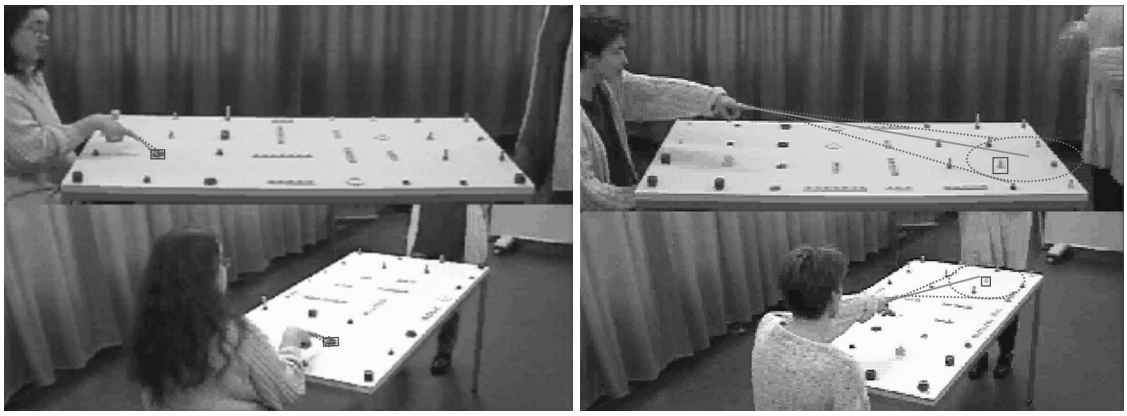
One last word of caveat: note, that our empirical studies are preleminary in the sense that only some variables have been controlled. This is due to the fact that the studies had not been conducted with issues of precise statistical hypothesis testing in mind. However, the results reported here are reasonably robust and will be reproducible in more carefully controlled experiments (see section 5).

2 Annotation of simple reference games

The analysis of our corpus of digital video data is based on an annotation with the TASX-ANNOTATOR software package[1] (Milde and Gut, 2001) which allows for the pursuit of an XML-based bottom up approach. Since the annotation data are stored in XML format, the extraction of the relevant information for purposes of statistical analysis could be realized *via* XSLT script processing straightforwardly. Details connected with the empirical setting and principles of annotation are laid out in (Kühnlein and Stegmann, 2003).

Figure 3 is a screenshot from a TASX annotation session that exemplifies the annotation scheme applied in score format. The set of annotation tiers includes a transcription of the agent's speech at word level (*speech.transcription*) and a classification of the dialogue move pursued (*move.type*). The annotation of deictic gestures follows the framework established by McNeill (1992). A gesture token has three phases: wrt pointing gestures the maximally extended and meaningful part of the gesture is called *stroke*, and *grasping* if an agent grasps an object. Stroke or grasping are preceded by the *preparation* phase, that is the movement of the arm and (typically) the index finger out of the rest position into the stroke position. Finally, in the *retraction* phase the pointer's arm is moved back into the rest position. We presume that pointing gestures serve one of two semantic functions: they uniquely pick out an object (*object pointing*) or merely narrow down the region in which the intended object lies (*region pointing*). In order to clarify this distinction, in figure 2 an occurrence of each gesture function is shown. The extension of pointing gestures is modelled with a pointing cone. Subfigure 2(b) depicts a case of region pointing, where several objects are located in the conic section of the pointing cone and the table top. In addition, the extension of the index finger does not meet the object in question. Against this, in object pointing the object is unequivocally singled out, i. e. it is the only object within the conic section, see subfigure 2(a). Seeing the "fuzziness" of pointing gestures as antic-

[1] It can be obtained at `http://tasxforce.lili.uni-bielefeld.de/`.



(a) Object pointing (b) Region pointing

Figure 2: Pointing cones. The extension of the index finger is indicated with a line, the pointing cone is indicated by dotted lines, and the box frames the intended object.

ipated by Quine's (1960) thesis of the indeterminacy of reference, the philosophical stance taken here can be labelled as *neo*-Peirce-Wittgenstein-Quinean (Rieser, 2004). The distinction between object and region pointing is captured on the *gesture.function* tier.

All tiers are specified for instructor and constructor, i. e. the respective tier names have an *inst.* or *const.* prefix, cf. figure 3.

To get a better grip on the kind of data we are concerned with, the speech portions of the sample dialogue from figure 3 were extracted and are reproduced below.

(1)	Inst:	The wooden bar [pointing to object1]
(2a)	Const:	Which one?
(2b)		This one? [pointing to object2]
(3a)	Inst:	No.
(3b)		This one. [pointing to object1]
(4)	Const:	This one? [pointing to object1 and grasping it]
(5)	Inst:	O.K.

We have the dialogue move of a *complex demonstration* of Inst's in (1) here, followed by a *clarification* move involving a pointing of Const's (2a, 2b). Inst produces a *repair* (3a), followed by a new *complex demonstration* move (3b) to the object she had introduced. Then we have a new *checkback* from Const (4) coming with a pointing and a grasping gesture as well as an acceptance move by Inst (5). The whole game is classified as an *object identification game.* The following events from different agents' turns overlap: (2b) and ((3a) and (3b)); (3b) and (4).

3 Reliability of the Annotation Scheme

Annotation-based projects must decide on the appropriateness of the annotation scheme. The standard way to handle this is using a bag of statistical methods that goes under the heading of *inter-rater agreement* or *inter-rater reliability.* Basically, the underlying idea is that of conducting a test on the results of raters who have annotated the same set of data. Different aspects of reliability (stability, reproducibility, and accuracy) go with different test designs (test *vs* retest, test *vs* test, and test *vs* "gold standard") and different *foci* of measured error (*intra*-observer, *inter*-observer, and deviation from norm) (Krippendorff, 1980). We are concerned with the second aspect of reliability (reproducibility, test *vs* test, *inter*-observer) here, since we have evaluated our annotation scheme comparing two raters' codings of the same video data.

In dialogue research the most widely known proposal concerning measures of *inter*-rater agreement is (Carletta, 1996) who argues in favor of the *kappa* statistics. However, there are serious problems associated with its interpretation, cf. (Feinstein and Cicchetti, 1990) on kappa paradoxes.

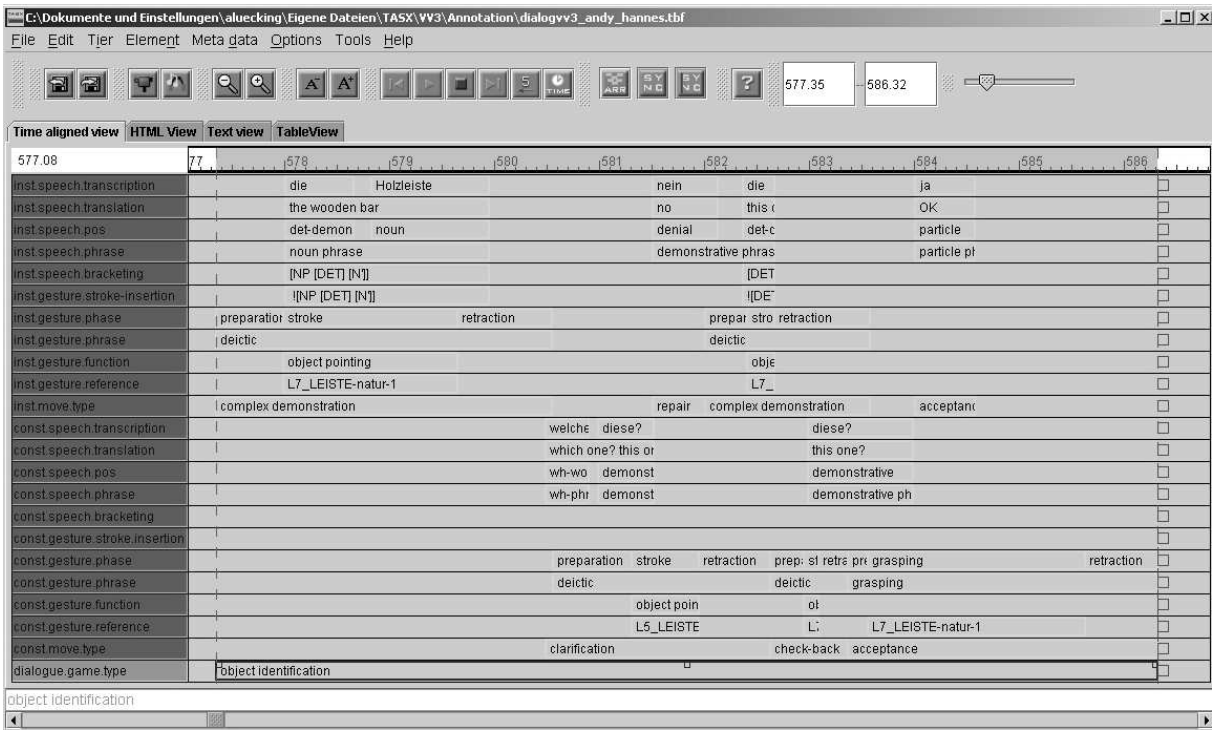


Figure 3: Annotation of a more complex Dialogue Game.

The point is that in the calculation of kappa the term representing the proportion of agreement by chance is systematically overestimated. Therefore, where appropriate with respect to the type of data involved, we pursue an alternative proposal based on the methodological framework of Gwet (2001), i. e. his *AC1* statistics. The latter—more adequately chance-corrected—coefficient is appropriate with respect to data resulting from a *type-ii* measurement on nominal-scale niveau.[2] Concerning judgments on magnitude scale niveau, which are usually classifiable as being of *type-i*, we use well-established conventional techniques, mainly correlation analysis. All calculations were implemented making use of the statistical programming environment R (R Development Core Team, 2003)[3].

Our *type-i* annotation data on a magnitude scale are the time-stamps for words and gestures, i. e. the points in time when words begin or end, and the start or end times of the gesture phases. In the TASX-ANNOTATOR a time bar is incorporated and synchronized with the video, so that a mark on the *speech.transcription* tier, say, at 201.4 seconds, means that the word in question starts at second 201.4 of the respective entire videotaped session. Since performing a gesture is a continuous action, the coding of gesture phases splits it into three parts where the end time of the preceding phase is identical with the start time of the following one. For example, the end of the preparation simultaneously marks the start of the stroke. The correlation of those time-based annotations was calculated over 108 words and 25 gesture occurrences using the Pearson product-moment correlation coefficient r. The outcomes are given in table 1. Despite almost perfect values of nearly 1, there is need for a closer look, since this result is influenced by the strict linearity of the underlying time scale. We transformed linear measurement data into nominally scaled data because of the category of stroke insertion, which is derived from allocating the stroke element's time in-

[2] *Type-ii* measurements are those, where the process leading to the measured datum is not well understood. Comparably well-understood measurements go by the name of *type i*. We will overload the term to refer to respective data, where appropriate.

[3] `http://www.r-project.org`.

	preparation	stroke		retraction	word boundaries	
	start	start	end	end	start	end
r	0.9999999	0.9999999	0.9999998	0.9999976	0.9999999	0.9999999

Table 1: Results for the correlation of gesture and word boundaries.

terval relative to the part of speech portions. This means, basically, a projection from temporally extended entities onto a sequence of symbols on, say, a modality-neutral representation at roughly word level, which could be fed into a parser. Essentially, we abstract away from exact timing—only the relative order remains, cf. example sentences 1 and 2 below, where $\searrow$ symbolizes gestural stroke.

(1) $\searrow$ the wooden bar (2) the $\searrow$ wooden bar

Resulting in nominally scaled data, the agreement regarding stroke insertion could be calculated using AC1, leaving us with a value of "merely" 0.73, which still can be regarded as good agreement. However, this result reveals that minor deviations in determining the boundaries of parts of speech and gesture phases can make a difference for the exact placement of the stroke.

One main concern was to assess whether the distinction between object pointing and region pointing is a concept reproducible by different raters. Being a nominal response category resulting from a *type-ii* measurement, the degree of correlation in classifying gesture functions was calculated using AC1. With a value of 0.4842 that is based on the judgment of 56 gesture occurrences, the agreement has to be classified as being fair at best. This shows that there are different habits in judging gestures as being related to object or region, which, in turn, indicates that either a clear-cut empirical category is lacking, or that the two-dimensional video data are not good enough to admit of this categorization.

Nevertheless, there was close agreement among raters concerning certain regions of the pointing domain. The domain of the reference games can be partitioned into three regions, according to the distance measured from the instructor, cf. figure 1. The two leftmost columns form the proximal region, the two rightmost columns the distal region, and the remaining four columns are called the mid-range region. Observe now that there is nearly perfect correlation with respect to the categorization of pointing to objects located in the proximal or distal regions. Dissent arises wrt pointing into the mid-range area. This shows that reliability of assignment of gesture functions is conditioned by the relative position of the objects that are referred to by the instructor.

Being interested in the dialogue structure of reference games, we also checked the reliability of our dialogue move annotation scheme. This was carried out computing the AC1 separately for instructor and constructor moves. The highly schematic instructor moves form a recurrent pattern that could be judged fairly consistently in the annotations of both observers ($N = 92, \text{AC1} = 0.9$). Agreement diminished when concerned with the more variable constructor moves ($N = 65, \text{AC1} = 0.795$).

4 Empirical findings

Gestures contribute to the content of communicative acts rather than being mere emphasis markers. This hypothesis can be substantiated by findings related to the semantic, the pragmatic, and the discourse level. On the semantic level, gestures contribute content that otherwise would have to be cast into clumsy verbal descriptions, thus making communicative acts more efficient. We found strong evidence for this in comparing the number of words used in referential NPs without pointing gesture occurrences (hereafter DDs, for *definite descriptions*) with NPs that come with pointing gestures (CDs, short for *complex descriptions*). A *t*-test was carried out on the number of words used in 65 CDs *vs* that in 74 DDs, resulting in a (highly) significant difference ($t = 6.22, p = 5.696 \cdot 10^{-9}, \alpha = 0.05$), cf. figure 4. This result can be couched into the slogan "Gestures save words!".

A related hypothesis was that the time the con-

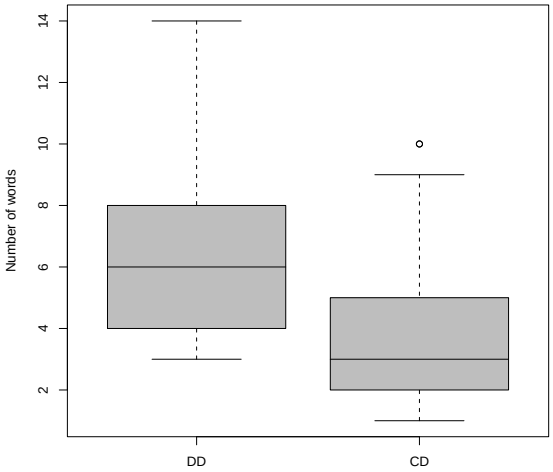


Figure 4: Boxplot displaying the number of words in CDs and in DDs.

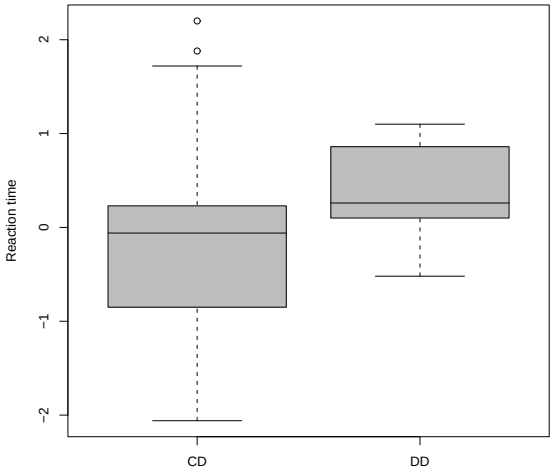


Figure 5: Boxplot for Const's reaction times (in seconds) following Inst's CDs and DDs.

structor needs to interpret the instructor's reference (reaction time) will be less after a CD than after a DD. The pointing gesture can be seen as guiding the constructor's eye towards the intended object—or at least towards a narrow region where the object is located—and thus as shortening the constructor's search effort. To assess this point, we calculated 48 (39 CDs and 9 DDs) differences between the start time of the constructor's move and the end time of the instructor's preceding referring act. A subsequent *t*-test did not result in a significant difference ($t = -1.4, p = 0.166, \alpha = 0.05$), but there seems to be a tendency for shorter reaction times after CDs, cf. figure 5.

What might have prevented a significant outcome was the fact that some objects are unique and therefore more salient, e. g. there is only one yellow cube (as opposed to several yellow bolts), so that the constructor could quickly spot such objects when directed with appropriate DDs only. In addition, the constructor may have used the instructor's gaze as a guiding device.

Moving from semantic to pragmatic issues, we also tried to find out whether there are contextual conditions constraining the use of gestures. This was defined in terms of frequencies of DDs *vs* CDs utilized to refer to objects in different columns of the pointing domain—that is, basically, wrt their distance as seen from the instructor. What is at stake here is whether the asymmetry that seems to be revealed in the bare data—comparethe plot depiction in figure 6—could be statistically validated; with DD's frequency peaks in the *periphery* (that are columns 1 plus 2 and 7 plus 8, or in terms introduced earlier, the union of the proximal and the distal region) and CD's frequency peaks in the *center* (the mid-range region, columns 3 to 6), there should be a bias to demonstrate objects in the middle of the domain using pointing gestures, whereas objects located in peripheral areas should be referred to only verbally.

There are two questions that have to be distinguished: First, is there a difference in the proportions of CDs *vs* DDs wrt the peripheral, resp. the center, region? Secondly, is there a difference in the proportions of CDs, resp. DDs, wrt the regions? To assess the second point the frequencies of peripheral and center CDs were compared using a χ^2-test, resulting in a significant outcome ($N_{\text{peripheral}} = 24, N_{\text{center}} = 41, \chi^2 = 7.8769, p = 0.005, \alpha = 0.05$). The comparison of the frequencies of DDs modelled through periphery and center yields an analogous result ($N_{\text{peripheral}} = 46, N_{\text{center}} = 28, \chi^2 = 8.7568, p = 0.003, \alpha = 0.05$). As regards the first issue, comparing the proportions of CDs *vs* that of DDs to

refer into the peripheral (and likewise the center area), we get significant outcomes, too (for periphery: $N_{\text{CD}} = 24, N_{\text{DD}} = 46, \chi^2 = 13.8286, p = 0.0002, \alpha = 0.05$; for center: $N_{\text{CD}} = 41, N_{\text{DD}} = 28, \chi^2 = 4.8986, p = 0.027, \alpha = 0.05$). Thus, the relative distance of the object in question to the instructor is a contextual factor for the choice of the mode of reference to that object. It is noteworthy that the partition of the reference domain imposed by the ratings of gesture function coincides with that of capturing the CD/DD-asymmetry.

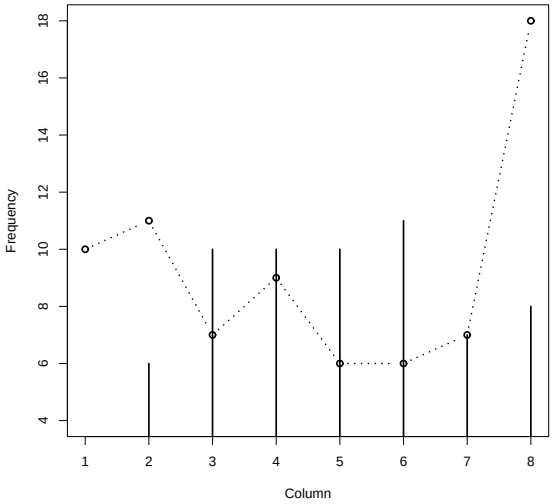


Figure 6: Plot for the modes of reference modelled by the eight columns of the reference domain; the bars depict the frequency distribution of CDs over the columns, the dashed line that of DDs.

At the beginning of this paper, a distinction was made between *intra*- and *inter*-move synchronization at the dialogue level. As regards *intra*-move synchronization we accounted for the temporal relations holding between gesture phases and escorting utterances. Above all, we focused on two synchronization effects, namely *anticipation* and *semantic synchrony* (McNeill, 1992, pp. 25-26, p. 131). The semantic synchrony rule states that gesture and speech present one and the same meaning at the same time (McNeill's "idea unit"). Anticipation refers to the temporal location of the preparation phase in relation to the onset of the stroke's co-expressive portion of the utterance. This rule states that the preparation phase precedes the linguistic affiliate of the stroke. Table 2 summarizes the descriptive statistics ($N = 25$).[4] Note, that we take the verbal affiliate to be the complete denoting linguistic expression, i. e. a possibly complex noun phrase. Row P gives the values for the start of the preparation phase relative to the onset of the first word of the noun phrase.Contrary to McNeill (1992, p. 25, 131), we found that the utterance usually starts a little before the initiation of the gesture (compare the positive mean value in table 2). This seems to contradict anticipation, given the way we operationalised McNeill's concept of the verbal affiliate or the idea unit. Similarly (compare the mean value in row R), the stroke ends (or the retraction starts) normally around 0.5 seconds before the end of the affiliate. Together with an average beginning of the stroke around 1 second after the onset of the utterance (mean for row S) this shows, that the prototypical stroke does not cross utterance boundaries. This is as to be expected in the light of McNeill's semantic synchrony rule. Note, however, that some extreme tokens (compare respective min. and max. values in table 2) were observed that seem to contradict the McNeill regularities, cf. (Kühnlein and Stegmann, 2003).

	Min.	Mean	Max.	Std. Deviation
P	-0.8	0.3104	4.68	1.0692
R	-0.86	0.564	3.38	0.89
S	-0.02	1.033	5.54	1.128

Table 2: *Intra*-move synchronization of preparation (P), retraction (R), and stroke (S).

Concerning *inter*-move synchronization one point of interest was the alignment of the end of Inst's preparation phase with Const's retraction phase. The resulting values, given in table 3, show that there is gap of around 0.3 seconds at aver-

Min.	Mean	Max.	Std. Deviation
-2.06	0.29	3.46	1.27

Table 3: *Inter*-move synchronization of Const's retraction and Inst's preparation.

[4] The different rows were calculated as follows: (P) $\text{preparation}_{\text{start}} - \text{speech}_{\text{start}}$, (R) $\text{speech}_{\text{end}} - \text{retraction}_{\text{start}}$, and (S) $\text{stroke}_{\text{start}} - \text{speech}_{\text{start}}$.

age. But the comparatively large values for the range (the span between the maximum and minimum values observed) and the standard deviation suggest that simply averaging the results camouflages a great deal of dispersion. A look at the dialogue video data reveals roughly two different sources for the resulting large and small values. If the object referred to lies within Const's reach, his initiation overlaps with Inst's retraction, indicating that the retraction phase contributes to a turn-taking signal. If the object referred to lies at the opposite side of the table Const first has to move around the table which delays initiation of her gesture.

5 Prospectus

As pointed out in the course of this paper, there are some rough edges in the employed annotation scheme as well as findings that can't be accounted for properly as yet. Accordingly, the top of our agenda includes experiments suitably designed to determine (or at least approximate sufficiently) the topology of the pointing cone. Such findings, we hope, will improve the classification of gesture functions and shed some light on the role the partitioning of the domain plays in the manner of how reference is established.To streamline our coding of move types we will hook up to some already established annotation scheme. At the time being, the one that seems to be most appropriate for our kind of data is the HCRC coding scheme (Carletta et al., 1996), which has to be augmented to capture pointing gestures. A third topic that could be fruitfully invistigated concerns the interaction of speech, gesture and gaze, which opens the door to truly *multi*-modal dialogue. As remarked above, the constructors in our settings might have used instructors' eye movement as an information source to find out the location of the object in question. As regards *intra*-move synchronization, we found a variety of temporal relationships that exceeds by far what was to be expected in the light of the current literature. In addition, we found surprising variability with respect to *inter*-move synchronization. Especially the frameworks aiming at a phenomenological account of gestures (mainly based on *iconics*) do not capture the structural flexibility of deictic gestures. A more promising direction to approach pointing and grasping in dialogues should perhaps be based on rigid semantics and underspecification approaches, cf. (Rieser, 2004).

References

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1996. HCRC dialogue structure coding manual. Technical Report TR-82, University of Edinburgh.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 2(22):249–254.

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

A. R. Feinstein and D. V. Cicchetti. 1990. High agreement but low kappa: I. the problem of two paradoxes. *Journal of Clinical Epidemiology*, 43:543–549.

Kilem Gwet. 2001. *Handbook of Inter-rater Reliability*. STATAXIS Publishing Company.

Klaus Krippendorff. 1980. *Content Analysis*, volume 5. SAGE Publications, Beverly Hills / London.

Peter Kühnlein and Jens Stegmann. 2003. Empirical Issues in Deictic Gestures. Technical Report 2003/03, Bielefeld University.

David McNeill. 1992. *Hand and Mind—What Gestures Reveal about Thought*. Chicago University Press, Chicago.

Jan-Torsten Milde and Ulrike Gut. 2001. The TASX-environment. In *Proceedings of the IRCS Workshop on linguistic databases, Philadelphia*.

Charles Sanders Peirce. 1965. *Collected Papers*, volume II. Harvard University Press, Cambridge, MA.

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

Willard Van Orman Quine. 1960. *Word and Object*. M.I.T Press, Cambridge, MA.

R Development Core Team, 2003. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Hannes Rieser. 2004. Pointing in dialogue. In *CATALOG04 Conference Proceedings*. pp.

Ludwig Wittgenstein. 1958. *The Blue and Brown Books*. Harper & Row, New York.

Speech Acts and Recognition of Insincerity

William C. Mann
SIL
bill_mann@sil.org

Jörn Kreutel
SemanticEdge GmbH
and Universität Potsdam
joern.kreutel@semanticedge.com

"Oh what a tangled web we weave when first we practice to deceive" – Scott[1]

Abstract

From the earliest years of speech act theory, sincerity, or the absence of it, has been one of the defining aspects of speech acts and their uses. It remains prominent today, but models of communication often give it little function. How could a model of dialogue be designed so that the sincerity of speech acts could be defined and examined? How could natural language understanding and generation programs recognize or use insincerity? Is sincerity part of a collection of speech phenomena that could share implementation methods? The issues are complex, but approachable.

What are appropriate recognition criteria for sincerity? Are the sincerity-conditions described by Austin or Searle adequate guides to recognition of insincerity? No.

Other ways of using assertions have a formal resemblance to insincere assertions. Several of these ways involve statements by a speaker who does not believe those statements. Not all of these ways involve deception. Examination of a collection of similar ways to use language leads to a much more accurate, possibly adequate, guide to recognizing the absence of sincerity.

This paper examines relationships between (in)sincerity and other language phenomena. Focusing on irony, exaggeration and understatement, it also identifies several others that share characteristics with sincerity, and thus might benefit from joint work on definitions, formalization and computational model building.

[1]Walter Scott's Marmion Canto 6,Stanza 17

Overview

Imagine a future computer system that has a strong capacity to understand, and perhaps participate in, many different kinds of natural language interaction. We would expect that part of the understanding process would focus on speech acts and rely on speech act interpretation processes. To do this, a theoretical basis would be needed, including all of the common aspects of each distinct kind of speech act. In addition to act identification and propositional content identification, the system would have to judge whether the act was sincere. This judgment is necessary because the consequences, the grounded understandings from particular speech acts are very different for acts judged to be insincere than for sincere acts.

Speech acts are defined in a way that includes a *sincerity-condition*. An act is judged insincere or sincere according to its conformity to its sincerity-condition. Correct formulation of sincerity-conditions is essential to sincerity recognition.

This paper examines sincerity-conditions as they are identified or defined in foundational work of Austin and Searle, and finds those formulations inadequate. The inadequacy has to do with improperly labeling some sincere ways of using language as insincere.

Finding those definitions inadequate, the paper makes a number of observations that appear to provide a basis for creating more adequate fresh definitions.

Background of Sincerity in Pragmatics

To some, sincerity might seem to be just a topic in psychology or sociology, but it has a long history in linguistics as well. Since the inception of speech act theory in Ordinary Language Philosophy (see (Austin, 1975)), the sincerity aspect of speech acts has been recognized. Searle reformulated Austin's conceptual scheme, again making sincerity one of the prominent concepts (Searle, 1969). The continuing development and use of speech act theory is a major theme of linguistic pragmatics, and the topic is still being developed in philosophy as well. For example, in Habermas' *Theory of Communicative Action* (Habermas, 1984), it is one of the three aspects of validity of speech acts.

Speech Act Type	Sincerity-Condition
Request	S wants H to do A
Assert, state (that), affirm	S believes P
(ask a) Question	S wants this information
Thank (for)	S feels grateful or appreciative for A
Advise	S believes A will benefit H
Warn	S believes E is not in H's best interest
Greet (on encounter)	none
Congratulate	S is pleased at E

Figure 1: Eight Speech Acts and Sincerity-conditions by Searle

More than with any other variety of speech act, people think about sincerity associated with assertions. The range of potentially insincere acts is much broader than just assertions, certainly including commissive acts (promises), and congratulations. Requests, questions, acknowledgments and various other acts also raise sincerity issues.

Austin said that for certain acts (including assertional acts) to be performed sincerely, the speaker must have "the right thoughts and feelings." Searle said that for certain insincere speech acts the speaker "purports to have (beliefs, intentions...) that he does not have." Both of these statements presume that speech acts are based on certain mental states of the speaker, and if a particular utterance is to be sincere, it will correspond to the speaker's mental state in a certain way.[2]

Part of the interest in sincerity surely comes from its involvement with deceptions (a larger category) and with lying. Another part surely arises from an episodic effort in philosophy to relate language to certitude.

Austin and Searle's defining characteristics for sincerity seem to be appropriate, but closer examination indicates that there are systematic exceptions.

According to Austin and Searle, for a statement by S, "Today is Tuesday," to be made sincerely, S must have a certain thoughts. In this case, S must believe that the day of saying the statement is Tuesday. The sincerity-condition of an assertional speech act such as this requires that the speaker believes the asserted proposition. So one of the effects of performance of the act is to communicate that the speaker believes the asserted proposition. Similarly, the sincerity-condition for commissive acts is for the speaker to intend to do what has been promised.

To bring more of the range of sincerity-conditions into view, the table in figure 1 is an extract from (Searle, 1969), p. 66-67, a table by Searle in which he defines 8 types of speech acts, with their sincerity-conditions. Clearly this is only an open, representative list. Based on the same book, we could add this:

Promise	S intends to do A

All but one of these can, according to Searle, be performed insincerely. All of the sincerity-conditions are different, but to a degree they share predicates: *believes, wants, feels grateful or appreciative, is pleased, intends to do.*

So, to recognize (in)sincerity in an interaction or written text, they present eight closely related tasks, each of which involves an assessment of the mental state of S. Below we focus on asserting.

Belief and Deception

Beside insincerity, there are other ways of using language that also involve the speaker making statements without any associated belief that those statements are true in the speaker's world of daily life. We will examine three other ways of using language that together challenge the adequacy of Searle's sincerity-condition for assertions. The three are *irony, exaggeration and understatement.* We call them *ploys.* Along with a possibly deceptive assertion, here is an example of each:

1) "I will send you the money after I get my first paycheck." - possibly deceptive assertion

2) "All of Bill Gates' influence is due to his good looks." – irony

3) "Every time the Beatles had a concert, ten million fans showed up." – exaggeration

4) "The Beatles had a few fans." - understatement

Using irony, as in example 2), involves saying something which is completely opposite to the intended meaning. The speaker expects that hearers will quickly recognize that the statement is not believed by the speaker. No deception is involved.

Use of irony violates Austin and Searle's sincerity-conditions on speech acts. Those conditions label ironic speech as insincere. Yet ironic language is usually understood for what it is. It is not insincere. Rather it draws a certain kind of attention to what is meant, and it requires the hearer to construct what is meant.

[2]It is not always the speaker whose mental state is at issue. Rather, using existing participant framework notions, especially of Levinson (Levinson, 1988) and McCawley (McCawley, 1989), we can often identify another participant in the act whose mental state is the one actually at issue. Space limitations prevent discussing this further here.

One of the important conclusions that can be drawn from comparing irony to insincerity is this:

The traditional defining conditions for insincerity are inadequate as recognition criteria for insincerity.

Recognition of insincerity cannot depend only on judging speaker's belief.

Consider exaggeration, as in 3). It is not credible that the speaker believes an exaggeration. Exaggeration is another ploy for manipulating the attention of the hearer, in this case to the scale on which the assertion depends. So it is like using a superlative such as "huge," but often with more marked effect.

Like ironies, exaggerations would be labeled as insincere by Searle's sincerity-condition. Yet they are not understood as insincere. The ploy is such that the hearer is not led to believe that the speaker believes an absurd statement. There is no deception and no appearance of insincerity. So again in this case, recognition of insincerity cannot depend only on judging speaker's belief.

Now consider understatement, as in 4). The effects of exaggeration and understatement are very similar. Both draw attention to the scale on which a degree-related assertion has its force. Unlike exaggerations, for understatements generally the speaker does believe the understatement, and more.[3] Again, there is no deception and no appearance of insincerity. These two ploys are quite similar, but their belief conditions are opposites.

Each of these ploys succeeds only if the hearer can determine, easily and with confidence, that the speaker believes the opposite (for irony), or significantly more (for understatement), or less (for exaggeration), than what is said. Ease of recognition (of the incompatible character of the assertion) is essential. None of them involves the speaker hiding his or her beliefs. When what is said is compared with speaker's belief, the ploys differ, but they all in similar ways draw attention to what the speaker is saying.

These examples together suggest that for sincerity and insincerity, attempted deception is a vital unstated element of the sincerity-condition of assertions, thus part of constituting the speech act. Similar arguments are expected to apply to commissives (promises) and perhaps to other speech acts as well.

Illocutionary Force

The same point about the inadequacy of the classical definitions of sincerity can be made by considering illocutionary force. Reconsider examples 1) through 4) above. The illocutionary force commonly assigned to assertion 1) agrees with the statement itself. The illocutionary force commonly assigned to 2) through 4) is an altered force, representing the speaker's obvious intent. It is opposite to what was said for 2) less than what was said for 3) and more than what was said for 4). In that sense they behave like kinds of indirect speech acts, and it would probably be useful to classify them as such.

In all four examples, what is said is obviously not compatible with the speaker's thought. The list of possible bases for incompatibility is open and quite diverse. It may be logical, emotional, motivational, a cultural taboo or a host of others.

In 1) that incompatible character is not obvious; in the other three, it is. This again suggests that Austin, Searle and successors had in mind images of deception when they were describing insincerity. The major role given to sincerity by Habermas (Habermas, 1984) also seems implicitly to have this character.

For the future, it might be best to make this aspect explicit, and continue to work with sincerity of speech acts as involving attempted deception.

Other Communicative Techniques with a Family Resemblance to Insincerity

One of the aims of this paper is to facilitate work on sincerity. This includes work in Ordinary Language Philosophy, in formalization of phenomena for models of language function, and development of computer-based models and agents capable of using human languages.

If the prevailing definitions are inadequate, as we claim above, then redefinition and reconception are called for. That rework is not here. When that work is done, the work might benefit from having a broad view of the interaction of language use and speakers' beliefs. In that hope, we now examine a loose collection of such uses of language.

In addition to attempted deception, irony, exaggeration and understatement, we now consider assertions that arise in pretending, play acting, written fiction, quotation, teasing, mistaken speech, forced speech, "confidence games,"[4] impersonation, deliberate misunderstanding, covert deliberate obscurity, legal representation of a client, overconfidence, politeness, outward respectful manner, and feigned ignorance. Each of these has a literature and most have some theoretical development. Many of them do not involve deception.

Figure 2 presents this arbitrarily chosen list, together with an indication of whether deception is commonly involved, and also the manner in which the speaker departs from believing the statement. The right hand column, labeled "Requires Assessing PM?" is about whether judging the sincerity of the item requires some estimation of the thoughts of the speaker (private memory, PM).

People regularly participate in these language uses, or

[3] This can depend on details of the assertion. Consider "The Beatles had only 100 fans." It could be an understatement which the speaker does not believe.

[4] One of these was the subject of the movie *The Sting*.

Language Use (ploy)	Dimension of Speaker's not Believing	Typically Deceptive?	Requires Assessing PM?
Irony	opposite belief	No	Yes
exaggeration	degree scale	No	Some cases
understatement	degree scale	No	Some cases
pretending	imagined world	No	No
play acting	imagined world	No	No
written fiction	imagined world	No	No
quotation	representing another speaker	No	No
mistaken speech	speaker commitment	No	No
legal representation of a client	role fulfillment toward a set of beliefs	No	No
forced speech	speaker commitment	not by speaker	No
"confidence games"	intended deception	Yes	Yes
impersonation	identity of speaker	Yes	Yes
covert deliberate obscurity	intention to communicate	Yes	Yes
politeness	apparent beliefs from convention	No	No
overconfidence	degree of confidence	Yes	Yes

Figure 2: Uses of Language in which the Speaker does not Believe What is Said

read about them, mostly with no extreme difficulty in deciding what is going on. (Of course, where deception is intended, we may be deceived.) In formal and computational models of natural language, there is not yet much to say.[5]

Certainly, for understanding ordinary natural language communication, the traditional characterizations of sincerity of speech acts must be supplemented. The table above may help in considering how more effective definitions might be constructed.

Use of sincerity and insincerity form the basis for some of the more complex phenomena. Because sincerity is at the base of some, modeling of sincerity can be expected to facilitate modeling more complex varieties of language.

Statements, beliefs and networks of beliefs

As all of us know by experience, effective lying does not proceed only on a statement by statement basis. We must present a view of ourselves that appears to be appropriately consistent, motivated and based on our immediate factual world. It requires a system of beliefs, commitments, intentions, interpretations of events and more. We call this a *self-presentation* (SP). In the literature of sincerity that is not oriented to speech acts, the focus is often on personality and habitual aspects of personal life. Often particular acts are taken as sincere because they come from people who have been judged trustworthy. For an interesting philosophical discussion of these aspects, see (Williams, 2002). The idea of a self-presentation (SP) is not about these aspects of sincerity.

Approaching sincerity from an interest in speech acts, we are in effect committed to enabling accounts of single acts, generally performed one at a time. Yet communication does not proceed as a set of independent acts. Acts are linked to context, and to other acts by other interacts or self. When a statement is received, if its sincerity is to be examined, no variety of knowledge is excluded. Various kinds of knowledge about the speaker, the subject matter, the occasion and the reasons for speaking contribute to judging sincerity, as well as to a number of other processes that apply to each statement. For example, ambiguity resolution, based on the plausibility of various readings, interacts with sincerity.

All such processes require a holistic use of available knowledge, interrelated knowledge that forms a network supporting interpretation. This means that judging sincerity, as part of overall language interpretation in communication, requires comparisons using a diversity of kinds of knowledge and techniques.

In ordinary interaction, when a speech act is performed and its sincerity is at issue, there is a prior history of knowledge of the speaker's thoughts, immediate purposes, cultural assumptions and more. There is a prior context of the interaction as well, so that there are already commitments in place, intentions being pursued and ideas under discussion. There is always to some degree a stable reconciliation of the parts, so that a new communication is judged for compatibility with a somewhat consistent network. A speech act is judged for sincerity relative to this larger body of knowledge. The SP is constituted of these sorts of knowledge, and when a speech act occurs, it becomes part of the self-presentation of the speaker.

Given this interrelated character of assertions in practical use, it is important to choose a representation of speech act effects that allows multiple collections of related information, networks that are similar in content but with differences, each network having its own kind of consistency.

In order to meet these requirements, we propose a memory organization for the information which must be examined to judge the sincerity of speech acts that can be sincere, and to make it possible for a speaker to be insincere. A hearer's memory, for ex-

[5]The commonalities between these ploys may justify exploring sharing parts in formalizations and implementations. Also, irony shares features with metaphor, so the possibilities of sharing are extended. See (Ortony, 1979) p. 108-111

ample, has four partitions. Two of them represent the private thoughts (PM_H) and public self-presentation (SP_H) of the hearer. The other two represent the private thoughts ($PM_{S/H}$) and public self-presentation ($SP_{S/H}$) of the speaker, in the hearer's view. Each participant's active memory is organized to reflect the possibility that what a speaker presents is not always consistent with what he or she thinks.

For all four of the modes of speaking above, deception, irony, exaggeration and understatement, what is said must be compared with beliefs attributed to the speaker.

Conclusions

Sincerity has been an important aspect of speech act theory ever since Austin and Searle introduced the theory. It was foundational; for certain speech acts to be performed sincerely, the speaker had to have certain thoughts: e.g. the speaker must believe what is asserted. Sincerity is part of the definitions of such acts.

However, if we explore recognition of insincerity, the definition is too imprecise to use.

There are sincere acts that are labeled insincere by the classic view. The classic definitions do not correspond to sincerity as we know it. When we compare deception, irony, exaggeration and understatement, we find that insincerity is involved with attempting to deceive rather than simply holding certain thoughts.

In order to recognize sincerity (or any of the other three ways of speaking), comparisons between networks of beliefs are required. They all, along with similar ways to use language, might be facilitated by using the four-partition model of active memory described above, the elaboration of which will be subject to further research.

References

J. L. Austin. 1975. *How to do Things with Words.* Harvard University Press.

J. Habermas. 1984. *The Theory of Communicative Action*, volume 1. Beacon Press.

S.J. Levinson. 1988. Putting linguistics on a proper footing: Explorations in goffman's concept of participation. In P. Drew and A. Wooton, editors, *Erving Goffman: Exploring the Interaction Order*, pages 161–227. Northeastern University Press.

J.D. McCawley. 1989. Participant roles, frames and speech acts. *Linguistics and Philosophy*, 22:595–619.

A. Ortony. 1979. *Metaphor and Thought.* Cambridge University Press.

J. Searle. 1969. *Speech Acts.* Cambridge University Press.

B. Williams. 2002. *Truth and Truthfulness.* Princeton University Press.

Ontologies and the Structure of Dialogue

David Milward
Linguamatics Ltd
david.milward@linguamatics.com

Martin Beveridge
Advanced Computation Laboratory
London Research Institute
Cancer Research UK
martin.beveridge@cancer.org.uk

Abstract

In this paper we will consider relatively simple dialogues, but in domains which involve multiple tasks and services, and concepts of different granularity. We re-examine the notion of focus of attention, and show how ontological information combined with information states can shed new light on the distinctions between linguistic and intentional structure. Elements of this work have been implemented in spoken dialogue systems for home information and control, and for a system that advises on appropriate action for doctors examining patients with suspected breast cancer.

1 Introduction

In current commercial dialogue systems, domain knowledge tends to be incorporated into the dialogue scripts, or used within very tight bounds e.g. via specific database queries at particular points in the dialogue. This kind of approach becomes costly as the domain becomes more complex. Some more recent systems provide a cleaner separation between domain knowledge and generic dialogue interaction rules. The most prominent system of this kind is AT&T's HMIHY system (Abella and Gorin, 1999), where a task inheritance hierarchy is kept separate from generic dialogue "motivators" (such as "missing information" or "clarification"). The task hierarchy encodes information that e.g. a "billing method" is either a "collect call", a "calling card" or a "billing number".

The approach we describe here follows the same motivation as the HMIHY system, but emphasises general ontological knowledge, of which a task hierarchy is a part. By 'ontology' here we mean simply a network of concepts and instances which are related to each other by semantic links. Ontologies, in particular those based on description logics, have been argued to be "the solution of first resort for all problems related to terminology and semantics" because they "occupy the sweet-spot between maximal expressive power on the one hand and computer tractability on the other" (Ceusters and Smith, 2003).

The systems we describe are for two specific and rather different domains. The first is home information and control. In this domain, the amount of domain knowledge required is not huge, but it is highly dynamic. It cannot be fixed in advance by the dialogue designer, or even after the installation of the system in the home. For example, the user can register for new services and add or move devices. In the second domain of cancer, the domain knowledge is more fixed, but is also much larger. Encoding this into a dialogue script by hand was therefore not an option, so a pre-existing medical ontology was used instead.

In sections 2 and 3 we show how is-a relations and part-whole relations can be used to handle clarification questions, and to influence the sequencing of a dialogue. We then relate this work to Grosz and Sidner's (1986) work on intentions and the linguistic structure of dialogue, and to the notion of focus of attention, as used for text generation by McCoy and Cheng (1991).

2 Clarification Dialogues

During a dialogue, the user may provide a response to a prompt that does not match any of the expected range of responses (as defined by the application domain) but is a hypernym or hyponym of an expected response (Milward and Beveridge, 2003). In this case the system should be able to discover the relation between the expected terms and the user response in order to resolve the discrepancy. In the case where the user's reply was a hypernym of an expected term, then their reply can be considered under-specified and the system can issue clarification questions in order to obtain a more detailed reply. Consider the following example from the cancer domain:

(1) S: Do you have a family history of chronic disease?
[*system expects: lung cancer, leukemia, sarcoma,...*]
U: Yes, cancer
S: What type of cancer?
U: Lung cancer.

In this example the system expected a specific disease but the user replied with a more generic term. The system therefore formulated a more specific question in order to elicit an answer at the required level of the disease ontology. Hypernyms are similarly treated in a generic fashion in the AT&T system (Gorin et al., 2002) via the use of a disambiguation operation.

In the case where the user's reply was a hyponym of an expected term then their reply can be considered over-specified and the system can (a) find a more general related term which matches the expected responses in order to answer the current question and (b) avoid asking subsequent more specific questions that have already been answered. For example

(2) S: Have you had any chronic diseases?
[*system expects: cancer, hypertension, diabetes, ...*]
U: Yes, Leukemia
S: ??What kind of cancer?

In this example the system expected a more generic answer to the question (such as "cancer"), but the user replied with a more specific term. The system, however, can match the answer to the question on the basis that leukemia is a cancer, which is a chronic disease. The system therefore avoids asking what kind of cancer since this has already been answered and therefore would be pragmatically ill-formed (indicated by "??").

In the home information and control domain the system uses an is-a hierarchy for tasks to allow users to skip directly to a task instance or a more specific task class.

(3) S: Which service do you want?
[*system displays the following options: entertainment, control, settings*]
U: Cinema booking
S: ??What kind of entertainment?

During a dialogue, the user may also provide a response to a prompt that does not match, and is not a hypernym/hyponym of, any expected response, but is associated with one or more of the expected responses by a non-is-a link that imposes an ordering on concepts, e.g. mereological ("part-whole") relations, or topological relations such as "in". In this case the system needs to discover the associative relation between the expected terms and determine whether the supplied term is more or less specific than the expected terms according to the ordering imposed by the relation. Hence, the system can issue clarification questions in order to obtain a more specific reply, in the same way as before. For example:

(4) S: Where does it hurt?
[*system expects: elbow, wrist, shoulder,...*]
U: In my arm.
S: Where in your arm?
U: In my elbow.

In this example the system initially expected a more specific body-part than the user supplied, but recognized that the supplied term "arm" was related to the expected terms in a part-whole hierarchy and was more general then the expected terms

3 Sequential Structure

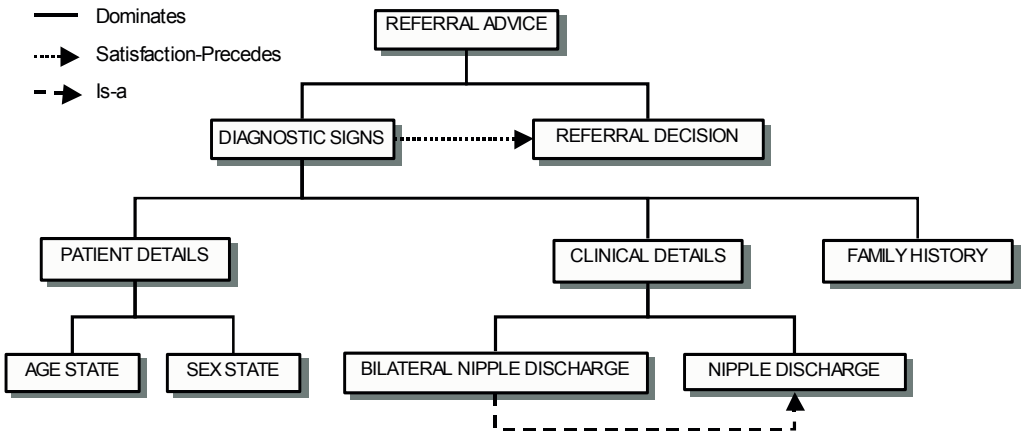


Figure 1. Intentional and informational relations for breast cancer referrals

In system-initiated dialogue, questions are grouped together at design time. Even in systems allowing some mixed initiative, this rarely affects the ordering of subsequent questions. Consider, the following dialogue:

(5) S: What is the patient's sex?
U: Female with some nipple discharge
S: What is the patient's age?
U: Fifty five
S: And is it a bilateral nipple discharge?
U: No

A standard form filling dialogue system might allow more than one answer to be given at once, but the order of the questions does not change. The Philips system (Aust et al., 1995) did deal with the case where the value for the new slot is underspecified, allowing an immediate clarification question, but did not deal with the more general case where two slots are related to each other. We can achieve much more natural dialogues if we cluster questions dynamically according to the user utterance. Consider the following:

(6) S: What is the patient's sex?
U: Female with some nipple discharge
S: And is it a bilateral nipple discharge?
U: No
S: What is the patient's age?
U: Fifty five

This appears to be a much more natural exchange, with the system immediately asking the follow-on question concerning nipple discharge. Questions which elaborate a previous question, either by asking about a particular attribute, or by asking for more specific information (determined by hyponymic, mereological or topological relations) will ideally appear straight afterwards. We achieve this behaviour by specifying what information the task requires, but not providing a strict ordering. The dialogue manager chooses the precise ordering according to the current state of the dialogue and its own domain knowledge.

For example, Figure 1 shows (a fragment of) the information available to the dialogue manager for the breast cancer referrals application (Beveridge and Milward, 2003). Boxes indicate tasks (implemented using the PRO*forma* process specification language (Fox et al., 2003)) with their associated topics. These are related by intentional dominance and satisfaction-precedence relations (Grosz and Sidner, 1986) as well as by ontological relations such as subsumption (Is-a). This is similar to recent approaches to discourse analysis. For example, Moser and Moore stress the need for representations of both intentional and informational relations between discourse segments, where the "informational structure [is] imposed by domain relations among the objects, states and events being discussed" (Moser and Moore, 1993, p. 416). The informational structure would therefore typically include "causal relations of various sorts, set relations, … the relation of identity between domain objects" and so on, creating "a complex network of domain relations that is defined independently of the intentional structure" (Moser and Moore,

1996, p. 417). In our case the informational structure is provided by a domain ontology.

In determining the sequential structure of a dialogue, satisfaction-precedence relations obviously provide the strongest constraint. For example, in Figure 1, the decision task should not be considered until all the enquiries have been completed. In the absence of satisfaction-precedence constraints, the default behaviour is to group siblings together following the dominance relations. Overlaid on this, however, is the effect of ontological relations between topics. For example, the presence of an ontological subsumption relation between bilateral discharge and discharge in Figure 1 causes the dialogue manager to infer a rhetorical *elaboration* relation (Mann and Thompson, 1988) between the related tasks and so overrides the default ordering in two ways. First, the question concerning the satellite task of the elaboration relation (e.g. bilateral discharge) will not be asked until the nucleus task (e.g. discharge) has been completed. This means that topics will "by preference be 'fitted' to prior ones – topics therefore often being withheld until such a 'natural' location for their mention turns up" (Levinson, 1983, p. 313), in this case when the topic of discharge has been introduced, either by the system or by the user. Secondly, once the nucleus task has been completed, the satellite task will have high priority for being considered next. This means that related topics are pursued as soon as possible, whilst they are still relevant, in order to avoid "unlinked topic 'jumps'" (Levinson, 1983, p. 313) later on. Indeed, "the relative frequency of marked topic shifts of this sort is a measure of a 'lousy' conversation" (Levinson, 1983, p. 313) in human-human dialogue.

4 Focus of Attention

In a very simple menu based dialogue system, the dialogue context can be simply the current node in the ontology. However, even if we just allow skipping (as in the case of hyponyms above) it becomes convenient to store the path to the node, not just the node itself, since we may not want to allow the user to back up to nodes that were skipped before, or will expect a differently marked expression e.g. "now go to" vs. "go back to" if they do. If the dialogue does not just consist of traversing a menu tree, but each node may be itself a partial description of a scenario, the intermediate nodes will have their own structure. This gets us to focus stacks (in Grosz and Sidner's terms) i.e. not just a path of atomic nodes, but a path through a series of focus spaces (each of which may contain some objects, properties or relations) which comprises the attentional state, or "focus of attention".

Focus of attention is often seen as closely related to part-whole relationships. For example, the focus changes when the participants are discussing a particular component of an object, or a particular step in a plan. However, authors such as McCoy and Cheng (1991), working in text generation, have also discussed focus of attention moving from a kind of action to a specialisation of the action. This corresponds to moving down an is-a hierarchy, similar to the task or disease clarification examples discussed above. McCoy and Cheng also emphasised that it is not just the objects that must be taken into account, but also the perspective taken on the objects. In the following description, the entity in focus is always the balloon, but the switching of perspective between colour and size makes the text infelicitous:

(7) ??The balloon was red and white striped. Because this balloon was designed to carry men, it had to be large. It had a silver circle at the top to reflect heat. In fact, it was larger than any balloon John had ever seen.

McCoy and Cheng build a focus tree rather than a focus stack. A focus stack could be derived by taking the right frontier of the tree, but keeping the full tree structure is more general, allowing for operations which access nodes that are not on the right frontier. The tree has a fine-grained structure, for example, when describing the person "John" they get the following:

(8) [John [physical [[brown hair] [blue eyes]]]
[interests[[plays football][collects stamps]]]]

In our approach, a dialogue is coherent if each dialogue participant acts independently to preserve coherence to the best of their ability. For example, if they want to ask a clarificatory or elaborative question (according to the participant's ontology) they do so immediately. Coherence is determined by considering the

information states of the participants, rather than from the perspective of an independent observer. Thus the decision of the system to ask about bilateral discharge has similar motivation to the close attachment of elaboration utterances in well-structured texts (Mann and Thompson, 1988), but that does not mean that coherent multi-party dialogues will necessarily have elaborations right next to their heads. For example, we predict the following dialogue to be coherent

(8) S: What are the patient details?
U: Female, severe nipple discharge, family history of cancer
S: And, is it a bilateral nipple discharge?
U: Yes
S: Can you give details of the family history?
U: Her mother was diagnosed with ovarian cancer at age forty-five

For the respondent, the grouping of 'female', 'severe nipple discharge' and 'family history' is acceptable since these are all pieces of information relevant to the referral decision and the respondent does not know at this stage that the system requires further information about the discharge. This suggests that the linguistic structure of dialogue arises from the interaction between the intentional/informational structures of the dialogue participants, but need not correspond to either. In contrast, in monologue, since there is a single intentional/informational structure, we would expect the speaker to elaborate immediately at the appropriate points. We would therefore predict the following to be infelicitous:

(9a) ??I have a female patient. She has nipple discharge. She has a family history of cancer. She suffers from bilateral discharge. Her mother was diagnosed with ovarian cancer at age forty-five.

In both monologue and dialogue, the intentional/informational structure can be violated by using marked constructions. For example, in the following monologue (Florian Wolf p.c.) the phrase "as far as family history is concerned" marks the fact that we are revisited an earlier topic.

(9b) I have a female patient with nipple discharge and a family history of cancer. The nipple discharge is bilateral. As far as family history is concerned, her mother was diagnosed with ovarian cancer at age fourty-five.

The discussion so far suggests that for dialogue systems we need a context which is at least as detailed as the focus trees of McCoy and Cheng (1991). For example, the dialogue context corresponding to the dialogue in example (6) might be as follows:

(10) [Diagnostic Signs [Patient Details [Sex=Female][Age=55]] [Clinical Details [Nipple Discharge=Yes] [Bilateral Nipple Discharge=No]]]

Attachment without any marked syntax or cue words can be performed not only at the right frontier, but at any nodes newly introduced by the other dialogue participant, as in example (8) above, where both "family history of cancer" and "severe nipple discharge" require further clarification.

Traversal down is-a hierarchies is represented similarly. For example, the structure corresponding to the user choosing "cinema booking" as an entertainment option is as follows:

(11) [Service [Entertainment [Cinema_Booking [Film=?][Time=?][No.People=?]]]

The context provides an appropriate abstraction of the dialogue history and current user/system goals. This, together with known ontological relationships, allows the system to decide on the next move. It should not be confused with a more detailed dialogue history which would be necessary for e.g. pronominal anaphora. As we have discussed, this might be differently structured. For example, a single value filled in for the "time" slot in the context above may have been established after several discontinuous turns, first specifying e.g. "morning" then "10 o'clock".

5 Implementation

Most of the components in the theoretical approach outlined above have been implemented, but currently not all within a single system. The

dialogues 1 - 3 and 6 are all real dialogues with the systems.

Focus of attention is used for generating prompts, for interpreting utterances in context, and for restricting the possible hypotheses of the speech recogniser. For example, in the home information demonstrator, the initial focus of attention at the topmost "service" node results in the question: "which service do you require?". At this point the speech recogniser grammar contains the expected service options, and possible hypernyms or hyponyms. The assumption is that users will say something that will be coherent with respect to the current state of the dialogue, or one of a limited number of marked utterances (e.g. "back to the top"). In order to generate the grammars we use the system's ontology. Strictly according to the approach above, it should be according to the user's ontology not the system. However, in both of the domains considered we can provide the system with a rich ontology for which it is reasonable to assume the user's ontology will be a subset.

In the cancer demonstrator, grammars are generated for the currently open nodes as well as for the right frontier. Furthermore, grammars are generated dynamically at the start of each dialogue segment in order to ensure the language model is consistent with the current high-level context. In the home domain, a narrowing of the focus of attention is achieved by going down not just the part-whole and is-a hierarchies, but also through the 'in' hierarchy (e.g. from kitchen to cooker). This position in the hierarchy provides a situation in which we can evaluate definite references (c.f. Milward 1995). For example, if the current position in the hierarchy is a particular room in the house, "the light" in "turn on the light" will be taken to be the light in this room.

6 Relation to Other Work

As described earlier our approach here has similar motivations to those behind the design of AT&T's HMIHY system (Abella and Gorin, 1999), except that HMIHY only makes use of a task hierarchy whereas we extend this to include domain ontological knowledge also.

Lascarides and Asher (1999) similarly make use of both intentional and discourse relations in order to interpret (as opposed to generate) dialogue. They employ a Question Elaboration relation *Q-Elab*(α, β) which "holds if β is a question whose answers all specify part of a plan to bring about an SARG [Speech Act Related Goal] of α". This is demonstrated in (12) below (Schlangen and Lascarides, 2002) in which *Q-Elab*(U_1, U_2) because all possible answers to U_2 specify part of a plan to bring about the SARG of U_1 (to arrange to meet next week).

(12) [U_1] A: Let's meet next week
[U_2] B: (OK.) Thursday at three pm?

This relation is therefore intentionally-based in that its definition refers to partial satisfaction of goals (rather like the dominance relation of Grosz and Sidner (1986)). However, this relation doesn't seem to be applicable to the examples we have described so far, e.g. (13) below.

(13) [U_1] S: What is the patient's sex?
[U_2]U: Female and she has some nipple discharge
[U_3] S: And is it a bilateral nipple discharge?
[U_4] U: no

Here, we don't seem to be able to claim that U_3 is a coherent continuation of U_2 because *Q-Elab*(U_2, U_3). The SARG of U_2 is presumably that S believe that the patient is female and has some nipple discharge, and the answers to U_3 don't appear to specify part of a plan to achieve that goal. In fact U_3 implicitly indicates that the goal of U_2 has already been achieved.

The elaboration relation that we have used is instead an informational relation, similar to the subject-matter elaboration relation of RST (Mann and Thompson, 1988). This means that we consider that U_2 elaborates U_1 if U_2 presents additional detail about the situation or some element of subject matter (e.g. a particular entity) introduced in U_1. Hence U_3 ELABORATE U_2 in (13) above by virtue of that fact that they refer to the common entity 'nipple discharge'.

Ginzburg (in press) also uses discourse relations such as elaboration to order the contents of QUD (Questions Under Discussion (Ginzburg, 1995)). This is used to account for the order in which questions are typically answered. How-

ever, only successive queries within a single turn are considered, such as (14) below (adapted from Ginzburg (in press)).

(14) [Q_1] A: Who have you invited?
[Q_2] Have you invited Jill?
B: yes
A: Aha
B: I'm also inviting…

Here the elaboration relation between Q_1 and Q_2 leads to the expectation that Q_2 will be answered before Q_1 and hence Q_2 should be maximal in QUD after A's initial utterance. Such an approach does not really apply to the examples we have discussed here, however, as we are not dealing with successive queries in a single turn.

In (13) above, for example, the system's QUD would be updated with a question Q_1 regarding the patient's sex at U_1, then when the answer A_1 is received in U_2 a question ?A_1 ("whether A_1") would be added to QUD but not pursued (since A_1 is accepted) and so both it and Q_1 would be downdated from QUD. Similarly, at U_3 QUD will be updated with a question Q_2 regarding whether or not the nipple discharge is bilateral, and after the answer A_2 in U_4 the question ?A_2 will added, but both ?A_2 and Q_2 will then be downdated from QUD since A_2 is accepted. Hence, there is never more than one question (plus a ?A question regarding the answer, which is never pursued) in QUD at any given time, and so the ordering of QUD does not seem to account for the coherence of U_3 as a continuation of U_2.

In fact, for both of the systems discussed here (in which answers are assumed to always be accepted) QUD would need to contain at most a main question, and a single other help-type question should the user ask for help (in which case the help question would be maximal in QUD until the system provided an answer and it was downdated).

In order to account for the coherence of (13) we need instead to impose an ordering over pending questions, but these do not form part of the Dialogue Game Board (DGB) in Ginzburg (1995; in press). Instead, they are presumably part of the dialogue participant's (DP's) unpublicised mental situation UNPUB-MS(DP).

The GoDiS system (Larsson et al., 2001), a dialogue system based on Ginzburg's QUD, does represent pending questions via a PLAN field in their information state which specifies "a list of dialogue actions that the agent wishes to carry out" (Larsson et al., 2001, p.1). However, our dialogue context corresponding to the PLAN field is more structured than this: it is a tree (rather than a list) and it is structured, not only according to a task hierarchy, but also following ontological relations. In GoDiS, for example, the only alteration of the sequence of actions in PLAN is to accommodate the user answering more than one question at a time. In our approach, however, we re-order the pending actions in the dialogue context according to ontological relations between them and the user's last utterance. This ensures that the system's next planned action is maximally relevant to the ongoing dialogue topic.

7 Conclusions

In this paper we have shown how ontological information can be used for clarification dialogues and to order questions to maximise coherence. This has extended the use of rhetorical relations from their traditional role in text analysis and generation to multi-party spoken dialogues, and has started to explore the distinctions between monologue and dialogue. We have implemented parts of this work in spoken dialogue systems for home information and control, and a system that advises doctors on whether to refer patients with suspected breast cancer to a specialist

Acknowledgements

The work on medical systems described in this paper has been funded by the European Union under the 5th Framework Project Homey (http://turing.eng.it/pls/homey/). The work on home control has had support from the EU 6th Framework Project, TALK (http://www.talk-project.org) and the UK Department of Trade and Industry Next Wave Programme. We would like to thank our partners in Homey and TALK for many useful discussions.

References

Alicia Abella and Allen Gorin. 1999. Construct Algebra: Analytical Dialog Management. *Proc. of ACL*, June, Washington D.C., USA.

Harald Aust, Martin Oerder, Frank Seide, and Volker Steinbiss. 1995. The Philips Automatic Train Timetable Information System. *Speech Communication*, 17:249-262.

Martin Beveridge and David Milward. 2003. Combining Task Descriptions and Ontological Knowledge for Adaptive Dialogue. In V. Matoušek and P. Mautner (Eds.) *Proc. of Text, Speech and Dialogue (TSD'03)*, September, České Budějovice, Czech Republic, pp. 341 – 348.

Werner Ceusters and Barry Smith. 2003. *A Realist Ontology for Natural Language Understanding*. Abstract, Colloquium on Knowledge Management, Erasmus University, Nijmegen, The Netherlands (http://www.landcglobal.com)

John Fox, Martin Beveridge, and David Glasspool. 2003. Understanding Intelligent Agents: Analysis and Synthesis, *AI Communications*, 16:139–152.

Jonathon Ginzburg. 1995. Interrogatives: Questions, Facts and Dialogue. In S. Lappin (Ed.) *Handbook of Contemporary Semantic Theory*. Blackwell, Oxford.

Jonathon Ginzburg. In press. Querying and Assertion in Dialogue. *Questions and the Semantics of Dialogue*, Chapter 4. Forthcoming from CSLI Publications and University of Chicago Press.

Kathleen McCoy and Jeanette Cheng. 1991. Focus of attention: Constraining what can be said next. In C.L. Paris, W.R. Swartout, and W.C. Mann (Eds.) *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer Academic Publishers, pp.103-124.

Barabara Grosz and Candace Sidner. 1986. Attention, Intention and the Structure of Discourse. *Computational Linguistics*, 12(3):175-204.

Staffan Larsson, Robin Cooper, and Stina Ericsson. 2001. menu2dialog, In *Proc. of the IJCAI'01 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, August, Seattle, USA.

Alex Lascarides and Nicholas Asher. 1999. Cognitive States, Discourse Structure and the Content of Dialogue. *Proc. of the Amstelogue Workshop on the Semantics and Pragmatics of Dialogue*, May, Amsterdam, The Netherlands.

Stephen Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.

William Mann and S. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8(3):243-281.

David Milward. 1995. Integrating Situations into a Theory of Discourse Anaphora. In *Proc. of the 10th Amsterdam Colloquium* pp. 519-538.

David Milward and Martin Beveridge. 2003. Ontology-Based Dialogue Systems. In *Proc. of the IJCAI'03 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, August, Acapulco, Mexico.

Megan Moser and Johanna Moore. 1993. Investigating Discourse Relations. *Proc. of the ACL Workshop on Intentionality and Structure in Discourse Relations*, Columbus, USA.

Megan Moser and Johanna Moore. 1996. Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics*, 22(3):409-419.

David Schlangen and Alex Lascarides. 2002. Resolving Fragments Using Discourse Information. *Proc. of the Edilog Workshop on the Semantics and Pragmatics of Dialogue*, Sept., Edinburgh, UK.

CLARIE: the Clarification Engine

Matthew Purver
Department of Computer Science
King's College London, Strand, London WC2R 2LS, UK
matthew.purver@kcl.ac.uk

Abstract

This paper describes the CLARIE system, a prototype information-state-based text dialogue system designed to deal with many types of *clarification requests* (CRs) by using a highly contextualised semantic representation together with a suitable grounding process. This allows it to interpret and respond to user CRs, and generate its own CRs in order to clarify unknown reference and learn new words, with both integrated within the standard dialogue update processes.

1 Introduction

CLARIE is a prototype information-state (IS)-based dialogue system designed to generate, interpret and respond to many types of *clarification requests* (CRs), allowing it to clarify problematic features of utterances – including unknown or surprising reference and meaning – and allowing users to do the same. This is achieved via a view of utterances as *contextual abstracts* requiring a *grounding* process to fully specify their content; a highly contextualised semantic representation including a view of ellipsis as abstraction; and a simple set of pragmatic contextual operations implemented as IS update rules. The system itself is implemented using the TrindiKit (Larsson et al., 2002), building upon the GoDiS dialogue system (Larsson et al., 2000) and SHARDS ellipsis reconstruction system (Ginzburg et al., 2001). Being a prototype, it is currently text-based and has only a small narrow-coverage grammar and a toy domain. This paper will concentrate on the novel semantic representation and the grounding process which enable its clarificational capabilities.

Motivation CRs (questions about a previous (sub-)utterance's meaning or form) are common in dialogue (3-4% of human-human dialogue turns according to a corpus study (Purver et al., 2003)) but are often not paid a great deal of theoretical or implementational attention. Dialogue systems generally have the capability of indicating inability to recognize or understand an entire user turn (or inability to do so to a reasonable degree of confidence), and will usually be able to produce outputs like *"I did not understand what you said. Please rephrase"* or *"You want to go to Paris, is that right?"* (from IBiS, (Larsson, 2002)). However, they are not usually able to clarify problems in a finer-grained way (e.g. at the word or phrase level, as argued for by (Gabsdil, 2003)), nor to understand and respond to CRs generated by the user. While recent advances have led to some systems that can highlight problematic words (Hockey et al., 2002), or ask about NP reference (Traum, 2003), these are so far restricted to particular phenomena, and tend to treat CRs and clarificational dialogue as governed by different rules from standard questions and standard dialogue.[1] As the following imagined example (Stone, 2003) shows,

[1] Hockey et al. (2002) use a separate module to highlight a problematic word and suggest reformulation; Traum (2003) sees CRs and their answers as pairs of dedicated dialogue moves *request-repair* and *repair*, and restricts them to wh- or alternative-questions about NP reference.

clarification need not be restricted to NPs, may involve extended sequences, and will ideally be seamlessly integrated within the dialogue.

(1)
Q:	What do I do next?
A:	Slide the sleeve onto the elbow.
Q:	What do you mean sleeve?
A:	That tube around the pipe at the joint.
Q:	What do you mean slide?
A:	Just push the sleeve gently over along the pipe.
Q:	What do you mean onto?
A:	The sleeve can hang there safely out of the way while you complete the repair.

Background While there has been extensive research into the possible levels of information which CRs can query, e.g. (Larsson, 2002; Gabsdil, 2003; Schlangen, 2004) but going back at least to (Clark, 1996), there has been little which examines the precise relation between their surface form and the question they ask.[2] A suitable analysis of CRs must provide two things: it must give a representation to normal utterances that explains how and why they can cause CRs; and it must allow CRs themselves (including their often elliptical forms) to be given a suitable representation.

Ginzburg and Cooper (2004) (hereafter G&C) provide a HPSG analysis of CRs that promises both. Utterances are represented as encoding *meaning* rather than *content*: functions from context to fully specified content. Contextually dependent parameters such as the reference of proper names (as well as speaker, hearer and utterance time) are abstracted to a set expressed in HPSG terms as a C-PARAMS feature, but shown here as the abstracted set in a simultaneous λ-abstract.[3]

An utterance *"I want to go to Paris"* would be given a representation such as (3), or simplifying by removing the parameters for speaker and addressee as will be done hereafter, as in example (4):[4]

(2) $\lambda\Big\{ABS\Big\}[RESTR].BODY$

(3) $\lambda\Big\{a, b, x\Big\}[speaker(a),\ addressee(b),\ name(x, Paris)].assert(a, b, go\_to(a, x))$

(4) $\lambda\Big\{x\Big\}[name(x, Paris)].assert(a, b, go\_to(a, x))$

This abstract must then be *grounded* – the abstract applied to the context – in order to fully instantiate the content by finding a suitable referent x which satisfies the given restriction $name(x, Paris)$. If this cannot be done (the hearer may not know what/where Paris is, or perhaps instantiating x to Paris leads to this new assertion being inconsistent with previous beliefs), the utterance cannot be grounded and this can lead to a CR concerning the intended reference of the problematic parameter $[x : name(x, Paris)]$.

This CR may take many forms, one of the most common being an elliptical *reprise fragment*, an echo *"Paris?"* (although others are also possible including reprise sluices *"Where?"* and full reprise sentences *"You want to go to Paris?"* or *"You want to go where?"*). G&C analyse such reprises using a question-under-discussion (QUD)-based approach to ellipsis (Ginzburg et al., 2001): briefly, reprises and other elliptical fragments are given a content which depends on the maximal QUD and a salient utterance, encoded as MAX-QUD and SAL-UTT features which are taken to be provided by context. In the case of standard non-reprise fragments (such as a bare answer to a wh-question), values for these contextual features will be provided by standard dialogue mechanisms triggered by the prior asking of the question. In the case of CRs, failure of grounding for a particular parameter licenses one of a set of *coercion operations* which produce a context in which the values of MAX-QUD and SAL-UTT allow the fragment to be resolved as a question concerning the constituent associated with the problematic parameter.

They give two specific such operations, termed *parameter focussing* and *parameter identification*, which lead to different contexts and thus eventually lead to different reprise readings. The first

[2] Although see (Rodríguez and Schlangen, 2004) in this volume.

[3] More specifically, they are interpreted as *simultaneous abstracts with restriction* as shown in (2): $\{ABS\}$ is the set of abstracted indices, $[RESTR]$ a set of restrictions which must be satisfied during application, and $BODY$ the body of the abstract (in this case, the semantic content). For further formal details, see (Ginzburg and Sag, 2000).

[4] The representation of examples (3) and (4) is simplified for clarity; in particular the use of a go_to predicate ignores details of the representation of verbs and modification.

will be used in cases where grounding produces a surprising or inconsistent content: the new context makes the question *"For which X did you say you want to go to X?"* under discussion, resulting in an elliptical CR *"Paris?"* being resolved as asking the yes/no question *"Is it really **Paris** you are saying you want to go to?"*, what they call the *clausal* reading. The second will be used in cases where no referent for *Paris* can be found: it produces a context where the QUD, and the resolved content of the CR, is the *wh*-question *"What do you intend the word 'Paris' to refer to?"*, what they call the *constituent* reading. In this second case, the elliptical fragment must be given an *utterance-anaphoric* analysis, allowing it to refer to the previous utterance *Paris* and ask a question about its intended content.

2 Utterance Representation

The analysis of G&C applies (explicitly at least) only to proper names. The general approach has now been extended to cover a wide range of word and phrase types and a wide range of CR forms, together with an integrated account of ellipsis and reprises.

Contextual Abstraction Given the view of clarification as querying contextual parameters, a suitable semantic representation must require *all* those elements of an utterance with clarificational potential (i.e. that can function as *sources* of CRs) to be included in the abstracted set. This leads to a highly contextualised representation. As shown in (5) for a simple utterance *"The dog snores"*, the abstracted set must include not only the referents of proper names, but the referents of definite NPs, and the denotations of common nouns, verbs and even function words such as determiners, as any of these can be subsequently clarified:

$$\text{(5)}\ \ \lambda\Big\{w, Q, P, S\Big\}.[w = Q(P),\ Q = the',\ name(P, dog),\ name(S, snore)].\ assert(a, b, S(w))$$

As detailed in (Purver and Ginzburg, 2003; Purver and Ginzburg, 2004), nouns and verbs are taken to denote named predicates,[5] while determiners denote logical relations. NPs are given a lower-order representation, denoting sets of individuals (rather than generalised quantifiers) which for definites must be made part of the abstracted set, but for indefinites and other quantifiers are existentially quantified within the utterance.[6] The overall representation is built up compositionally by a HPSG grammar; space precludes details here, and the use of HPSG is not essential, but it is important to note that the output of the grammar (here, an HPSG sign) must associate each subconstituent with all and only the contextual parameters which it contributed, thus ensuring that clarifying a particular word or phrase can ask only about its contributions to the utterance.

Ellipsis & Reprises via Abstraction The treatment of reprises and other elliptical fragments is based on that of SHARDS (as assumed by G&C): their content is specified by the grammar as being identified with features of the context, specifically the features MAX-QUD and SAL-UTT. However, these features are now also taken to be members of the utterance's abstracted set. A fragment *"Paris"* is therefore given an abstracted representation such as that in (6): its content will be an assertion of a proposition concerning some object x named Paris, but first not only x but a maximal QUD question Q and a salient utterance S must be found in context to fully specify that proposition and the role of x in it:

$$\text{(6)}\ \ \lambda\Big\{x, Q, S\Big\}.[name(x, paris),\ max\_qud(Q) \wedge Q = \lambda\{\ldots\}.P,\ sal\_utt(S) \wedge content(S, x)].\ assert(a, b, P(\ldots x \ldots))$$

This has several advantages. Firstly, it avoids some potential problems with the SHARDS approach (see (Schlangen, 2003)): the representation of the fragment is now a well-defined object (a simultaneous abstract) rather than being underspecified (with the potential problems that can lead to when implementing within a standard grammar and/or parser), and is derived entirely compositionally, with the non-abstracted parts derived entirely from the constituent words and the ab-

[5] Mass nouns and bare plurals are more complex, seen as ambiguous between predicates (or kinds) and existentially quantified individuals.

[6] Quantifier scope is treated via a functional analysis, and monotone decreasing quantifiers via a representation as pairs of sets – see (Purver and Ginzburg, 2004) for more details.

stracted set expressing only its contextual dependence, specifying the type of context that the abstract can be applied to. Secondly, resolution no longer has to be performed by a separate module, as with SHARDS: as all contextual dependence is now expressed together, resolution can be part of the grounding process, instantiating all parameters together to obtain the fully specified content.[7]

All fragments show this kind of abstraction; so do all CRs (which depend on their source SAL-UTT utterance, and on MAX-QUD if elliptical or reprise). The representation in example (6) above is for a standard declarative fragment, where the word *Paris* is taken as denoting an object named Paris, and the overall content of the fragment is an assertion. Other versions are also possible (and required for certain CR types): firstly equivalent interrogative fragments; secondly utterance-anaphoric fragments, where the word *Paris* is taken to denote a previous salient utterance *'Paris'*, as in (7):

(7) $\lambda\{Q, S\}.[max\_qud(Q), sal\_utt(S) \wedge phon(S, \texttt{paris})]. ask(a, b, Q(\ldots S \ldots))$

As (7) shows, CRs are treated as standard interrogative *ask* moves (rather than special e.g. *request-repair* moves). Their CR nature comes only from the question asked (concerning some feature of the source utterance). They also have contextual parameters which must be grounded, and CRs-of-CRs are therefore possible (and do occur in corpora – see (Purver, 2004)).

3 Utterance Processing & Grounding

The system's ability to handle clarificational dialogue centres around the grounding process: application of the abstract to the current context (the IS), finding suitable referents for each of the abstracted parameters (including values for the MAX-QUD and SAL-UTT features) such that the utterance then receives a fully specified content.[8] It is the inability to ground a particular parameter in context (or to ground it in a way that is consistent with what is already known in context) that gives rise to system CRs; it is the grounding of parameters in a suitable way that allows user CRs (particularly elliptical forms) to be interpreted correctly.

In CLARIE this process is implemented in as simple a way as possible. Rather than using general reasoning or inference, a set of logical constraints and preferences that govern the process are defined as TrindiKit IS update rules. Prolog backtracking is then used to find an assignment for the abstracted set such that all constraints are satisfied. The constraints are expressed as preconditions on particular rules, and express general requirements on the way parameters are instantiated: for example, to ensure that utterances are interpreted in such a way that their content is internally consistent and consistent with what is already known (where possible). The preferences are expressed in the ordering of the update rules, and ensure that utterances are grounded in a maximally relevant way: e.g. that an ambiguous utterance be instantiated as an answer to a question currently under discussion if possible, and only as a CR if not.

Disambiguation It is therefore the grounding process which performs disambiguation between all the possible moves that the utterance can make. The abstracted representation means that lexical ambiguity, as well as ambiguity of reference and elliptical resolution, is now represented as contextual dependence (to be fixed by grounding). Any other ambiguity (i.e. multiple possible parses returned by the grammar) means that more than one possible abstract will be available,[9] and again it is the grounding process that must choose between them based on their consistency and relevance. This therefore allows all IS information to be used in disambiguation: not only the possible referents for parameters, but the current state of the dialogue (QUDs, beliefs etc.).

[7] Note that this use of abstraction in ellipsis is not the same as the higher-order abstraction approach of (Dalrymple et al., 1991) for VP ellipsis, in which abstracts are formed from the antecedent and used in resolving the ellipsis. Here, the elliptical fragment is the abstract, to be applied to the context.

[8] The term *grounding* is often used in a wider sense to incorporate the general process of understanding and addition to the common ground, including acknowledgement and acceptance. Here it is used narrowly to refer to the fixing of an utterance's content in context.

[9] The grammar will in fact *always* assign more than one possible parse – a typical fragment will be given at least four representations, two in each of the dimensions declarative/interrogative and standard/utterance-anaphoric – see above.

Utterance Processing The CLARIE IS is shown in AVM (8). Like GoDiS, it is divided into two parts, with PRIVATE for system plans and beliefs that have not been explicitly introduced into the dialogue, and SHARED representing the system's view of the common ground:

(8)
$$\begin{bmatrix} \text{PRIVATE} & \begin{bmatrix} \text{AGENDA} & [\textit{stack(action)}] \\ \text{PLAN} & [\textit{stackset(action)}] \\ \text{BEL} & [\textit{set(proposition)}] \\ \text{BG} & [\textit{set(parameter)}] \end{bmatrix} \\ \text{SHARED} & \begin{bmatrix} \text{COM} & [\textit{set(proposition)}] \\ \text{BG} & [\textit{set(parameter)}] \\ \text{QUD} & [\textit{stack(question)}] \\ \text{SAL-UTT} & [\textit{stack(sign)}] \\ \text{UTT} & [\textit{nstackset(4,sign)}] \\ \text{PENDING} & [\textit{stack(set(sign))}] \end{bmatrix} \end{bmatrix}$$

In the shared part, COM is a set of commitments and BG a set of descriptions of referents that have been explicitly introduced in the dialogue. QUD and SAL-UTT are stacks of QUDs and salient utterances respectively, used for ellipsis resolution and answerhood. UTT is an utterance record, a stack of utterances in linear dialogue order which is used to find CR sources, allowing the questions asked by user CRs to be fully interpreted, and their answers to be determined. It is therefore important that its members are *signs*, including all attendant phonological, syntactic and semantic information which may be clarified, rather than just semantic representations such as *moves*. It has a limited length, currently 4 utterances, as Purver et al. (2003) found that CRs beyond this distance are rare. PENDING holds ungrounded utterance abstracts during the grounding process.

Utterance processing is based on the protocol proposed by G&C and proceeds as follows: the utterance abstracts produced by the parser are pushed onto the PENDING and UTT stacks while grounding is attempted:

(9)
$$\begin{bmatrix} \text{AGENDA} & \langle \dots \rangle \\ \text{QUD} & \langle \dots \rangle \\ \text{UTT} & \langle U, \dots \rangle \\ \text{PENDING} & \langle U \rangle \end{bmatrix}$$

Grounding can be achieved via three sets of rules, tested in order. Firstly, standard *integration* rules attempt to ground the utterance given the current IS (and in particular the current top members of QUD and SAL-UTT). Secondly, *accommodation* rules can be used to achieve the same effect using a new QUD determined from a relevant but as yet not explicitly asked question from the plan (see (Larsson et al., 2000)). If neither of these succeed, a third set of *coercion* rules attempt to ground the utterance as a user CR, by using contextual coercion operations which produce new CR-related values of QUD and SAL-UTT. In all cases, the newly grounded utterance (with all parameters now fully instantiated) is removed from PENDING and its update effects applied to the IS (e.g. for a question, raising a new QUD Q and an action to respond):

(10)
$$\begin{bmatrix} \text{AGENDA} & \langle \texttt{respond}(Q), \dots \rangle \\ \text{QUD} & \langle Q, \dots \rangle \\ \text{UTT} & \langle U, \dots \rangle \\ \text{PENDING} & \langle \rangle \end{bmatrix}$$

If all grounding rules fail, the utterance cannot be grounded in the current IS, and a set of *clarification* rules use its ungrounded parameters to form a suitable CR and add an agenda action to ask it. Importantly, the ungrounded utterance is left in PENDING so that grounding can be re-attempted once a user answer to the CR has provided sufficient new information:

(11)
$$\begin{bmatrix} \text{AGENDA} & \langle \texttt{clarify}(U), \dots \rangle \\ \text{QUD} & \langle \dots \rangle \\ \text{UTT} & \langle U, \dots \rangle \\ \text{PENDING} & \langle U \rangle \end{bmatrix}$$

Grounding Non-CRs The *integration* rules use the current IS unchanged and only general constraints apply: the parameters corresponding to MAX-QUD and SAL-UTT must be identified with the current heads of the QUD and SAL-UTT IS stacks. Parameters concerning individual referents (e.g. of names and definites) must be identified with unique antecedents in the BG set or available from the domain model. Parameters concerning predicates (from nouns and verbs) must again refer to explicit antecedents in BG or be available

from the lexicon.[10] Logical relations (contributed by closed-class function words) are always available from the lexicon. A simple consistency check is also applied, preventing grounding of e.g. assertions that conflict with previously held beliefs (conflicting assertions will lead to clarification – see below). The protocol is as follows:

1. Interpret as answering a question which is under discussion (currently maximal in QUD).
2. Interpret as asking a question which is relevant to the current IS.
3. Interpret as a greeting, closing or thanking move.

Accommodation rules follow similar lines, but use a new accommodated QUD to ground the MAX-QUD parameter. Coercion rules are described in the next section.

4 Processing CRs

A Sample Dialogue The dialogue below shows some of CLARIE's current capabilities:

```
S1> Where do you want to go to?
U1> pardon?
S2> Where do you want to go to?
U2> i want to go to that city
S3> Which city?
U3> paris
S4> How do you want to travel?
U4> no i want to go to london
S5> London?
U5> yes london
S6> How do you want to travel?
U6> by pullman
S7> 'Pullman'?
U7> train
S8> The trip will cost fifty pounds.
U8> sixty?
S9> No, fifty.
U9> fifty?
S10> Yes.
```

Note that both system CRs (S3, S5, S7) and user CRs (U1, U8, U9) are included, as are questions about utterance form (U2), NP reference (S3) and noun predicate reference (S7), and both clausal (S5, U8, U9) and constituent (S7) reprises.[11]

User CRs The *coercion* rules allow user CRs to be grounded (and thus fully interpreted) by modifying the context: using coercion operations in order to ground the MAX-QUD and SAL-UTT parameters. Firstly, for all CRs, the *source* utterance (the utterance being asked about) must be identified by examining constituents from the UTT record until a suitable one is found which meets the constraints associated with the CR. This corresponds to grounding the SAL-UTT parameter. Some CRs only require this step – those which have their propositional content specified directly by the grammar, thus not requiring a MAX-QUD question to fill in, merely requiring utterance reference to be established, e.g. non-reprise CRs such as *"Did you say 'Paris'?"*, *"What do you mean by 'pullman'?"* or conventional expressions such as *"What?"*, *"Pardon?"*. For these, then, a simple coercion operation which provides possible SAL-UTT values from the UTT record suffices.

However, the most common forms of CR are of a reprise and/or elliptical nature and therefore also require a MAX-QUD parameter to be grounded to fully specify their propositional content. For these, there are currently four different possible coercion operations which not only take a possible source constituent from UTT but also use it to produce particular new CR-related MAX-QUDs: G&C's clausal and constituent versions, plus two further questions about lexical form (one querying the identity of an echoed word, one a *gap* question querying the identity of the word following an echoed word). These operations and the order and constraints of the rules which apply them are determined by corpus and experimental studies (see (Purver, 2004)). Constraints include factors such as source word/phrase category (e.g. no function words can lead to constituent readings as their meaning is mutual knowledge; only some function words such as number determiners seem likely

[10] This allows a degree of *alignment* behaviour (Pickering and Garrod, 2004) – once words and phrases have been grounded to particular referents, these become present in BG and will be used again for grounding before looking in the lexicon.

[11] The current behaviour is designed to demonstrate the elliptical CR interpretation and generation capabilities, hence the highly elliptical forms used. Many system CRs may benefit from less elliptical form in practice (e.g. S7 might be less ambiguously realised as *"What do you mean by 'pullman'?"* or *"What is a 'pullman'?"*) – this behaviour can be controlled by a user-settable flag.

to be given clausal readings), source parallelism (constituent reprise fragments require phonological parallelism) and common ground (words previously grounded in the dialogue do not cause constituent readings).[12]

1. Coerce SAL-UTT only and interpret as a conventional or non-reprise CR – see U1.
2. Perform *parameter identification* and interpret as a constituent fragment reprise *if* the source is the first mention of a content phrase fragment.
3. Perform *parameter focussing* and interpret as a clausal reprise (sentence, fragment or sluice) *if* the source is a content phrase or number determiner – see U8, U9.
4. Perform *gap identification* and interpret as a lexical reprise gap.
5. Perform *lexical identification* and interpret as a lexical reprise.

As CRs are *ask* moves, these coercion rules will all include in their effects the addition of a new question to QUD, and an agenda action to answer it. Answers can now be established from the relevant features of the antecedent utterance in the UTT record: in the case of clausal and constituent CRs, directly from the semantic content of the source utterance; in the case of lexical or gap questions, from the identity of the source utterance itself. This process is specified as a set of selection rules which produce corresponding *assert* moves – these are then passed to a generation module which uses the grammar to generate in the same way as answers to normal questions.[13]

System CRs If a user utterance cannot be grounded in any way, the *clarification* rules produce a system CR. Particular grounding problems lead to particular questions being asked (again derived from empirical findings – out-of-vocabulary nouns and verbs lead to constituent wh-questions, parameters which cause inconsistency lead to clausal yes/no "check" questions), and particular source types will lead to particular forms being used (definites and demonstratives are clarified using sluices, nouns using fragments). The protocol is as follows:

1. Parse failure (no move to ground): constituent CR about whole utterance meaning.
2. Unknown parameter (no unique referent can be found): clausal wh-question about source constituent *if* a definite or pronoun (leading to a sluice – see S3); constituent wh-question otherwise (leading to a fragment – see S7).
3. Inconsistent parameter (can only be grounded in a way which causes inconsistency with previous beliefs): clausal yn-question leading to a reprise fragment – see S5.
4. Inconsistent moves (can only be grounded inconsistently) and irrelevant moves (ungroundable MAX-QUD or SAL-UTT parameter): constituent question about whole intended utterance meaning.

Note that there is significant correspondence between these grounding problem types and the levels identified by (Larsson, 2002; Schlangen, 2004), but that the association of problematic parameters with their source words/phrases allows specific CR forms which target those phrases.[14]

As system CRs (being *ask* moves) introduce a new question to QUD, subsequent user answers (elliptical or not) can be interpreted according to standard answerhood rules – no special treatment is required – and that as long as such answers provide the required information, the problematic PENDING utterance will now be groundable without requiring repetition.

5 Summary

This paper shows how a basic dialogue system can be implemented which can handle many forms of CR, using them to clarify unknown reference and meaning, and allowing users to do the same. The grammar can parse and generate a wide range of

[12] While these coercion operations can perhaps be seen as a form of reasoning about context, they are highly constrained and far from unrestricted inference.

[13] There are two special cases: when answering CRs which ask about word meaning, alternative descriptions are used wherever possible rather than the original problematic form; when answering yn-questions negatively, an over-answer is produced by answering the coerced MAX-QUD as well as the explicit CR question – see S9.

[14] Being text-based, the level of *perception* is currently ignored, but must be taken into account if a speech interface is to be added – see (Gabsdil, 2003).

CR forms from a wide range of source types. Importantly, this is achieved without having to use heavyweight inference about utterances or their relation to each other, or modelling the user's beliefs or context. User CRs are assigned straightforward (although heavily contextually dependent, and often ambiguous) representations; and the grounding process then gives them a full interpretation by instantiating their abstracted parameters in context. Problems with the grounding process, and with particular abstracted parameters, lead to system CRs. CRs are not treated in a significantly different way from other utterances: they are parsed by the same grammar and given a standard interrogative interpretation as *ask* moves which raise new questions for discussion – it is just that these questions concern other utterances.

Acknowledgements

The author is currently supported by the ESRC (RES-000-22-0355), although most of this work was carried out as part of the ROSSINI project (EPSRC GR/R04942/01). Thanks are due to the Catalog reviewers, and to Jonathan Ginzburg for enormous help and support as the supervisor of my thesis (Purver, 2004), of which this paper summarizes chapter 6.

References

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

M. Dalrymple, S. Shieber, and F. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4):399–452.

M. Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*.

J. Ginzburg and R. Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–365.

J. Ginzburg and I. A. Sag. 2000. *Interrogative Investigations: the Form, Meaning and Use of English Interrogatives*. CSLI Publications.

J. Ginzburg, H. Gregory, and S. Lappin. 2001. SHARDS: Fragment resolution in dialogue. In *Proceedings of the 4th International Workshop on Computational Semantics (IWCS-4)*, pages 156–172.

B. Hockey, J. Dowding, G. Aist, and J. Hieronymus. 2002. Targeted help and dialogue about plans. In *ACL-02 Companion Volume to the Proceedings of the Conference*.

S. Larsson, P. Ljunglöf, R. Cooper, E. Engdahl, and S. Ericsson. 2000. GoDiS - an accommodating dialogue system. In *Proceedings of ANLP/NAACL-2000 Workshop on Conversational Systems*.

S. Larsson, A. Berman, L. Grönqvist, and F. Kronlid. 2002. TrindiKit 3.0 manual. In *SIRIDUS Deliverable 6.4*.

S. Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University.

M. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, forthcoming.

M. Purver and J. Ginzburg. 2003. Clarifying noun phrase semantics in HPSG. In *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar (HPSG-03)*.

M. Purver and J. Ginzburg. 2004. Clarifying noun phrase semantics. To appear in *Journal of Semantics*, 21(3).

M. Purver, J. Ginzburg, and P. Healey. 2003. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse & Dialogue*. Kluwer Academic Publishers.

M. Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, London.

K. Rodríguez and D. Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*.

D. Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh.

D. Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*.

M. Stone. 2003. Specifying generation of referring expressions by example. In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*.

D. Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *Proceedings of the International Workshop on Computational Semantics*.

Incrementality, Alignment and Shared Utterances

Matthew Purver[†] and **Ruth Kempson**[‡]
Departments of [†]Computer Science and [‡]Philosophy
King's College London, Strand, London WC2R 2LS, UK
{matthew.purver,ruth.kempson}@kcl.ac.uk

Abstract

This paper describes an implemented prototype dialogue model within the Dynamic Syntax (DS) framework (Kempson et al., 2001) which directly reflects dialogue phenomena such as alignment, routinization and shared utterances. In DS, word-by-word incremental parsing and generation are defined in terms of *actions* on semantic tree structures. This paper proposes a model of dialogue context which includes these trees and their associated actions, and shows how alignment and routinization result directly from minimisation of lexicon search (and hence speaker's effort), and how switch of speaker/hearer roles in shared utterances can be seen as a switch between incremental processes directed by different goals, but sharing the same (partial) data structures.

1 Introduction

Study of dialogue has been proposed by (Pickering and Garrod, 2004) as the major new challenge facing both linguistic and psycholinguistic theory. Two of the phenomena which they highlight as common in dialogue, but posing a significant challenge to and having received little attention in theoretical linguistics, are *alignment* (including *routinization*) and *shared utterances*. Alignment describes the way that dialogue participants apparently mirror each other's patterns at many levels (including lexical word choice and syntactic structure), while routinization describes their convergence on set descriptions (words or sequences of words) for a particular reference or sense. Shared utterances are those in which participants shift between the roles of parser and producer midway through an utterance:[1]

(1)
Daniel:	Why don't you stop mumbling and
Marc:	Speak proper like?
Daniel:	speak proper?

(2)
Ruth:	What did Alex …
Hugh:	Design? A kaleidoscope.

These are especially problematic for theoretical or computational approaches in which parsing and generation are seen as separate disconnected processes, even more so when as applications of a grammar formalism whose output is the set of wellformed strings:[2] the initial hearer must parse an input which is not a standard constituent, and assign a (partial) interpretation, then presumably complete that representation and generate an output from it which takes the previous words and their syntactic form into account but does not produce them. The initial speaker must also be able to integrate these two fragments.

In this paper we describe a new approach and implementation within the Dynamic Syntax (DS) framework (Kempson et al., 2001) which al-

[1] Example (1) from the BNC, file KNY (sentences 315–317).

[2] Although see (Poesio and Rieser, 2003) for an initial DRT-based approach.

lows these phenomena to be straightforwardly explained. By defining a suitably structured concept of context, and adding this to the basic word-by-word incremental parsing and generation models of (Kempson et al., 2001; Otsuka and Purver, 2003; Purver and Otsuka, 2003), we show how alignment phenomena result directly from minimisation of effort on the part of both hearer and speaker independently (implemented as minimisation of lexical search in parsing and generation), and how the switch in roles at any stage of a sentence can be seen as a switch between processes which are directed by different goals, but which share the same incrementally built data structures.

2 Background

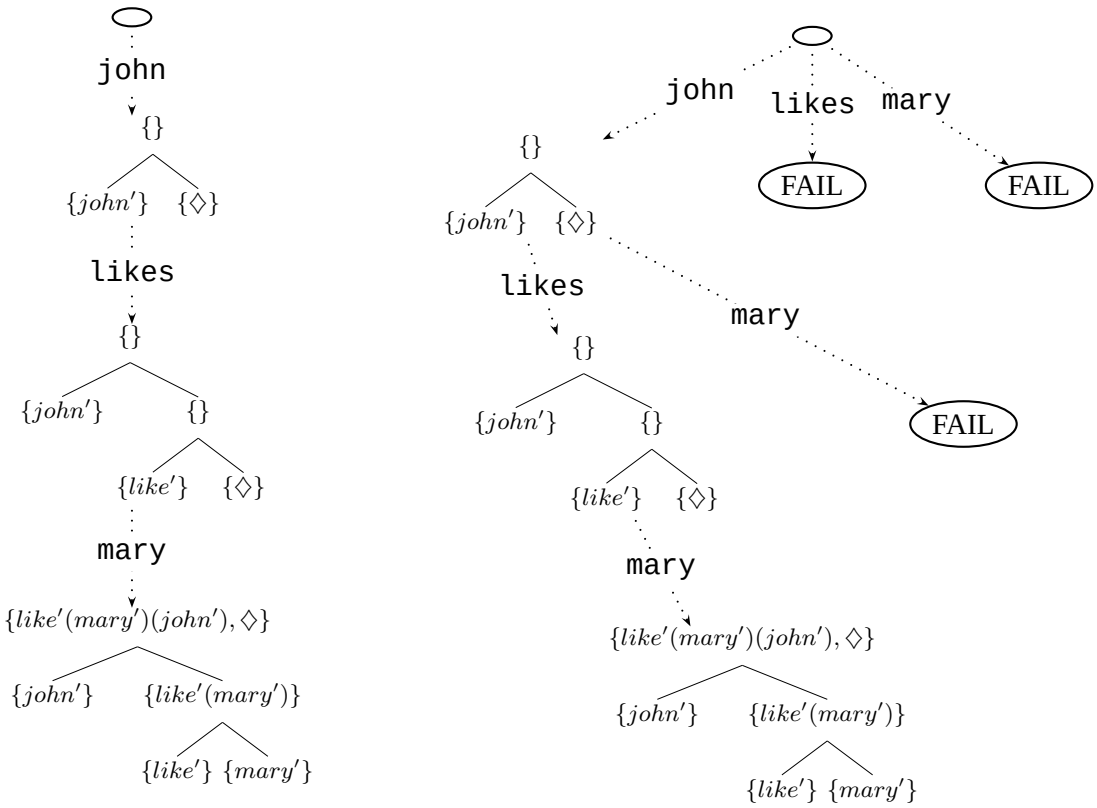
DS is a parsing-directed grammar formalism in which a decorated tree structure representing a semantic interpretation for a string is incrementally projected following the left-right sequence of the words. Importantly, this tree is not a model of syntactic structure, but is strictly semantic, being a record of how some formula representing interpretation assigned to the sentence in context is compiled, with the topnode of the tree being decorated with some (type t) formula, and dominated nodes with subterms of that formula. In this process, sequences of linked trees may be constructed, sharing decorations through anaphoric processes, e.g. for relative clause construal. In DS, grammaticality is defined as parsability (the successful incremental construction of a tree-structure logical form, using all the information given by the words in sequence), and there is no central use-neutral grammar of the kind assumed by most approaches to parsing/generation. The logical forms are lambda terms of the epsilon calculus (see (Meyer-Viol, 1995) for a recent development), so quantification is expressed through terms of type e whose complexity is reflected in evaluation procedures that apply to propositional formulae once constructed, and not in the tree itself. The analogue of quantifier-storage is the incremental build-up of sequences of scope-dependency constraints between terms under construction: these terms and their associated scope statements are subject to evaluation once a propositional formula of type t has been derived at the topnode of some tree structure.[3] With all quantification expressed as type e terms, the standard grounds for mismatch between syntactic and semantic analysis for all NPs are removed; and, indeed, all syntactic distributions are explained in terms of this incremental and monotonic growth of partial representations of content, hence the claim that the model itself constitutes a NL grammar formalism.

Parsing (Kempson et al., 2001) defines parsing as a process of building labelled semantic trees in a strictly left-to-right, word-by-word incremental fashion by using computational and lexical actions defined (for some natural language) using the modal tree logic LOFT (Blackburn and Meyer-Viol, 1994). These actions are defined as transition functions between intermediate states, which monotonically extend tree structures and node decorations. Words are specified in the lexicon to have associated lexical actions: the (possibly *partial*) semantic trees are monotonically extended by applying these actions as each word is consumed from the input string. Partial trees will be underspecified in one or more ways, each being associated with a requirement for subsequent update: the tree may lack a full set of nodes; some relation between nodes may be only partially specified (as in the parsing of long-distance dependency effects); some node may lack a full formula specification (as in the parsing of anaphoric/expletive expressions); and the sequence of scope constraints may be incomplete. Once all requirements are satisfied and all partiality and underspecification resolved, trees are *complete*, parsing is successful and the input string is said to be grammatical. For the purposes of the current paper, the important point is that the process is monotonic: the parser state at any point contains all the partial trees produced by the portion of the string so far consumed which remain candidates for completion.

Generation (Otsuka and Purver, 2003; Purver and Otsuka, 2003) (hereafter O&P) give an initial method of context-independent tactical generation based on the same incremental parsing process, in which an output string is produced according to an input semantic tree, the *goal tree*. The generator

[3] For formal details of this approach to quantification see (Kempson et al., 2001) chapter 7.

Figure 1: **Parsing *"john likes mary"* ...** **... and generating *"john likes mary"***



incrementally produces a set of corresponding output strings and their associated partial trees (again, on a left-to-right, word-by-word basis) by following standard parsing routines and using the goal tree as a subsumption check. At each stage, partial strings and trees are tentatively extended using some word/action pair from the lexicon; only those candidates which produce trees which subsume the goal tree are kept, and the process succeeds when a complete tree identical to the goal tree is produced. Generation and parsing thus use the same tree representations and tree-building actions throughout.

3 A Model of Context

The current proposed model (and its implementation) is based on these earlier definitions but modifies them in several ways, most significantly by the addition of a model of context: while some notion of context was assumed no formal model or implementation was given.[4] The contextual model we now assume is made up not only of the semantic trees built by the DS parsing process, but also the sequences of words and associated lexical actions that have been used to build them. It is the presence of (and associations between) all three, together with the fact that this context is equally available to both parsing and generation processes, that allow our straightforward model of dialogue phenomena.[5] For the purposes of the current implementation, we make a simplifying assumption that the length of context is finite and limited to the result of some immediately previous parse (although information that is independently available

[4]There are other departures in the treatment of linked structures (for relatives and other modifiers) and quantification, and more relevantly to improve the incrementality of the generation process: we do not adopt the proposal of O&P to speed up generation by use of a restricted multiset of lexical entries selected on the basis of goal tree features, which prevents strictly incremental generation and excludes modification of the goal tree.

[5]In building n-tuples of trees corresponding to predicate-argument structures, the system is similar to LTAG formalisms (Joshi and Kulick, 1997). However, unlike LTAG systems (see e.g. (Stone and Doran, 1997)), both parsing and generation are not head-driven, but fully (word-by-word) incremental.

can be represented in the DS tree format, so that, in reality, larger and only partially ordered contexts are no doubt possible): context at any point is therefore made up of the trees and word/action sequences obtained in parsing the previous sentence and the current (incomplete) sentence.

Parsing in Context A parser state is therefore defined to be a set of triples $\langle T, W, A\rangle$, where T is a (possibly partial) semantic tree,[6] W the sequence of words and A the sequence of lexical and computational actions that have been used in building it. This set will initially contain only a single triple $\langle T_a, \emptyset, \emptyset\rangle$ (where T_a is the basic axiom taken as the starting point of the parser, and the word and action sequences are empty), but will expand as words are consumed from the input string and the corresponding actions produce multiple possible partial trees. At any point in the parsing process, the context for a particular partial tree T in this set can then be taken to consist of: (a) a similar triple $\langle T_0, W_0, A_0\rangle$ given by the previous sentence, where T_0 is its semantic tree representation, W_0 and A_0 the sequences of words and actions that were used in building it; and (b) the triple $\langle T, W, A\rangle$ itself. Once parsing is complete, the final parser state, a set of triples, will form the new starting context for the next sentence. In the simple case where the sentence is unambiguous (or all ambiguity has been removed) this set will again have been reduced to a single triple $\langle T_1, W_1, A_1\rangle$, corresponding to the final interpretation of the string T_1 with its sequence of words W_1 and actions A_1, and this replaces $\langle T_0, W_0, A_0\rangle$ as the new context; in the presence of persistent ambiguity there will simply be more than one triple in the new context.[7]

Generation in Context A generator state is now defined as a pair (T_g, X) of a goal tree T_g and a set X of pairs (S, P), where S is a candidate partial string and P is the associated parser state (a set of $\langle T, W, A\rangle$ triples). Initially, the set X will usually contain only one pair, of an empty candidate string and the standard initial parser state, $(\emptyset, \{\langle T_a, \emptyset, \emptyset\rangle\})$. However, as both parsing and generation processes are strictly incremental, they can in theory start from *any* state. The context for any partial tree T is defined exactly as for parsing: the previous sentence triple $\langle T_0, W_0, A_0\rangle$; and the current triple $\langle T, W, A\rangle$. Generation and parsing are thus very closely coupled, with the central part of both processes being a parser state: a set of tree/word-sequence/action-sequence triples. Essential to this correspondence is the lack of construction of higher-level hypotheses about the state of the interlocutor. All transitions are defined over the context for the individual (parser or generator). In principle, contexts could be extended to include high-level hypotheses, but these are not essential and are not implemented in our model (see (Millikan, 2004) for justification of this stance).

Anaphora & Ellipsis Anaphoric devices such as pronouns and VP ellipsis are analysed as decorating tree nodes with metavariables to be updated from context using terms established, or, for ellipsis, the (lexical) tree-update actions. Strict readings of VP ellipsis result from taking a suitable semantic formula directly from a tree node in context; sloppy readings involve reuse of actions. This action re-use approach, combined with the representation of quantified elements as terms, allows even ellipsis phenomena which are problematic for other e.g. abstraction-based approaches (see (Dalrymple et al., 1991) for discussion):

(3)
A: A policeman who arrested Bill read him his rights.
B: The policeman who arrested Harry did too.

Here re-use of the actions associated with *read him his rights* allows *Harry* to be selected as antecedent for the metavariable projected by these re-used actions, given the new context, leading to a new term and a sloppy reading. Other forms of ellipsis such as bare fragments involve taking a previous structure from context as a starting point for parsing (here *wh*-expressions are analysed as particular forms of metavariables, so parsing the question yields an open formula which the term

[6] Strictly speaking, scope statements should be included in these n-tuples – for now we consider them as part of the tree.

[7] The current implementation of the formalism does not include any disambiguation mechanism. We simply assume that selection of some (minimal) context and attendant removal of any remaining ambiguity is possible by inference.

presented by the fragment updates):

(4)
A: What did you eat for breakfast?
B: Porridge.

4 Alignment & Routinization

The parsing and generation processes must both search the lexicon for suitable entries at every step (i.e. when parsing or generating each word). For generation in particular, this is a computationally expensive process in principle: every possible word/action pair must be tested – the current partial tree extended and the result checked for goal tree subsumption. As proposed by O&P (though without formal definitions or implementation) our model of context now allows a strategy for minimising this effort, as it includes previously used words and actions. If a first search through context finds a subset of such actions which can be re-used in extending the current tree, full lexical search can be avoided altogether. Even given a more complex model of the lexicon which might avoid searching all possible words during generation (e.g. by activating only certain subfields of the lexicon based on the semantic formulae and structure of the goal tree), searching through the immediate context will still minimise the effort required.

High frequency of elliptical constructions is therefore expected, as ellipsis licenses the use of context, either in providing some term directly or in licensing re-use of actions which context makes available; the same can be said for pronouns, as long as they (and their corresponding actions) are assumed to be pre-activated or otherwise readily available from the lexicon.

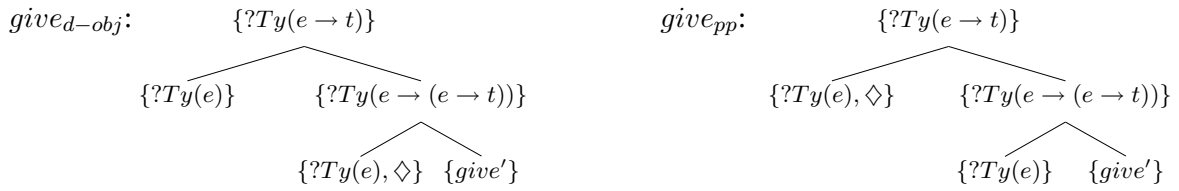
Lexical Alignment As suggested by O&P, this can now lead directly to a model of alignment phenomena, characterisable as follows. For the generator, if there is some action $a \in (A_0 \cup A)$ suitable for extending the current tree, a can be re-used, generating the word w which occupies the corresponding position in the sequence W_0 or W. This results in *lexical* alignment – repeating w rather than choosing an alternative but as yet unused word from the lexicon.

In this connection, re-use of the actions associated with the construction of semantic trees is importantly distinct from re-use of these trees and the terms that decorate them. For example, the actions associated with the parse of a pronoun decorate a node with a metavariable which must then be provided with a fully specified value from a term in context (see section 3); subsequent re-use of this action will introduce a new metavariable, rather than merely copying the previous value, and so the resulting value in this case may differ from the value given previously. Such re-use of actions is essential to construal of indexical pronouns, such as *I* and *you*, as their actions will require values to be assigned from the current context (which must contain information about the identity of the current speaker and addressee) rather than copying values from previous uses.

This re-use of actions applies also to quantifying expressions, e.g. indefinites. By definition, on the DS approach, the construal of a quantified noun phrase introduces a new variable as formula decoration, and re-use of these actions will not therefore license introduction of the same term. This is in contrast to the approach of (Lemon et al., 2003) in which strings are re-used in a way that licenses the same construal, necessitating a special rule to prevent a generator from re-using indefinite NPs with the same interpretation as the antecedent occurrence.

Syntactic Alignment Apparent alignment of *syntactic* structure also follows in virtue of the procedural action-based specification of lexical content. (Branigan et al., 2000) showed that syntactic structure tends to be preserved, with semantically equivalent double-object forms *"give the cowboy a book"* or full PP forms *"give a book to the cowboy"* being chosen depending on previous use. Most frameworks would have to reflect this via activation of syntactic rules, or perhaps preferences defined over parallelisms with syntactic trees in context, both of which seem problematic. In DS, though, this type of alternation is reflected not as a difference in the output of parsing (the semantic tree structure) but as a difference in the lexical actions used during parsing to build up this output: a word such as *give* has two possible lexical actions a' and a'' corresponding to the two

Figure 2: Output of alternative lexical actions for $give$



alternative forms (figure 2). A previous use will cause either a' or a'' to be present in $(A_0 \cup A)$; re-use of this action will cause the same form to be repeated.

Repetition of adjective structures as attributive or in a predicative relative-clause (*a green book* vs. *a book which is green* (Cleland and Pickering, 2003)) can be explained in the same way. Adjective construal in DS is distinguished by whether a linked tree structure is constructed before the head noun (by the lexical actions associated with attributive adjectives) or after the head (by the actions associated with a relative pronoun); and re-use of these actions will cause repetition of form. So again the two distinct tree-building strategies, despite producing the same logical form, nevertheless lead us to expect parallelism following the sequence of actions already in context.

Semantic Alignment & Routines The same approach can be applied for the parser, with contextual re-use of actions bypassing the need to test all possible actions associated in the lexicon with a particular word. A similar definition holds: for a word w presented as input, if $w \in (W_0 \cup W)$ then the corresponding action a in the sequence A_0 or A can be used without consulting the lexicon. Words will therefore be interpreted as having the same sense or reference as before, modelling the *semantic* alignment described by (Garrod and Anderson, 1987). These characterisations can also be extended to sequences of words – a sub-sequence $(a_1; a_2; \ldots; a_n) \in (A_0 \cup A)$ can be re-used by a generator, producing the corresponding word sequence $(w_1; w_2; \ldots; w_n) \in (W_0 \cup W)$; and similarly the sub-sequence of words $(w_1; w_2; \ldots; w_n) \in (W_0 \cup W)$ will cause the parser to use the corresponding action sequence $(a_1; a_2; \ldots; a_n) \in (A_0 \cup A)$. This will result in sequences or phrases being repeatedly associated by both parser and generator with the same sense or reference, leading to what Pickering and Garrod (2004) call *routinization* (construction and re-use of word sequences with consistent meanings).

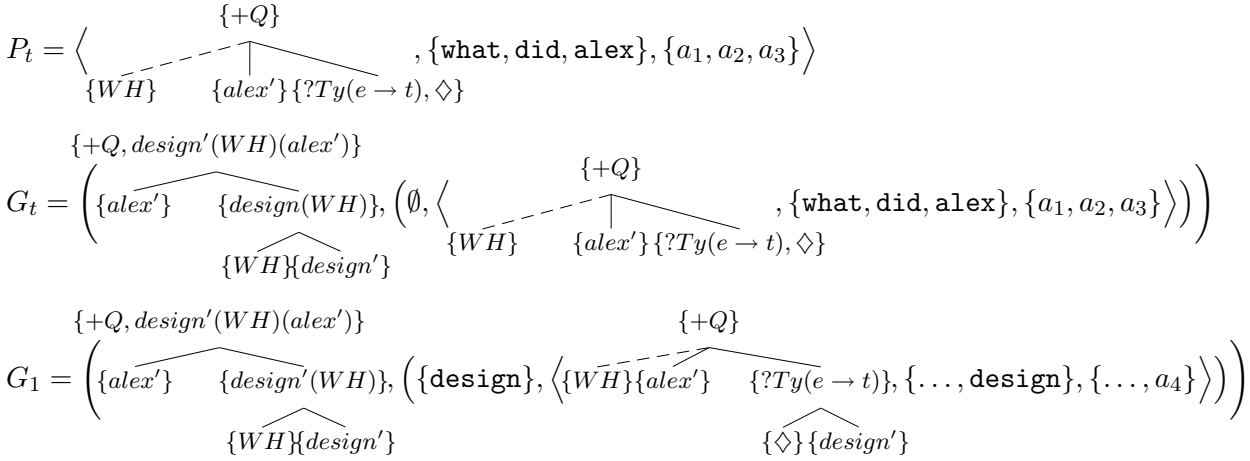
It is notable that these various patterns of alignment, said by Pickering and Garrod (2004) to be alignment across different levels, are expressible without invoking distinct levels of syntactic or lexical structure, since context, content and lexical actions are all defined in terms of the same tree configurations. Note also that this context-based approach models both speaker and hearer actions without any need for meta-level calculations about their interlocutor.

5 Shared Utterances

O&P suggest an analysis of shared utterances, and this can now be formalised given the current model. As the parsing and generation processes are both fully incremental, they can start from any state (not just the basic axiom state $\langle T_a, \emptyset, \emptyset \rangle$). As they share the same lexical entries, the same context and the same semantic tree representations, a model of the switch of roles now becomes relatively straightforward.

Transition from Hearer to Speaker Normally, the generation process begins with the initial generator state as defined above: $(T_g, \{(\emptyset, P_0)\})$, where P_0 is the standard initial "empty" parser state $\{\langle T_a, \emptyset, \emptyset \rangle\}$. As long as a suitable goal tree T_g is available to guide generation, the only change required to generate a continuation from a heard partial string is to replace P_0 with the parser state (a set of triples $\langle T, W, A \rangle$) as produced from that partial string: we call this the *transition state* P_t. The initial hearer A therefore parses as

Figure 3: Transition from hearer to speaker: *"What did Alex . . . / . . . design?"*



usual until transition,[8] then given a suitable goal tree T_g, forms a transition generator state $G_t = (T_g, \{(\emptyset, P_t)\})$, from which generation can begin directly – see figure 3 as a display of the interpretation process for example (2).[9] Note that the context does not change between processes modulo information about identity of current speaker and addressee.

For generation to begin from this transition state, the new goal tree T_g must be subsumed by at least one of the partial trees in P_t (i.e. the proposition to be expressed must be subsumed by the incomplete proposition built so far by the parser). Constructing T_g prior to the generation task will often be a complex process involving inference and/or abduction over context and world/domain knowledge – Poesio and Rieser (2003) give some idea as to how this inference might be possible – for now, we make the simplifying assumption that a suitable propositional structure is available.

Transition from Speaker to Hearer At transition, the initial speaker B's generator state G'_t contains the pair (S_t, P'_t), where S_t is the partial string output so far, and P'_t is the corresponding parser state (the transition state for B).[10] In order for B to interpret A's continuation, B need only use P'_t as the initial parser state which is extended as the string produced by A is consumed.

As there will usually be multiple possible partial trees at the transition point, A may continue in a way that does not correspond to B's initial intentions – i.e. in a way that does not match B's initial goal tree. For B to be able to understand such continuations, the generation process must preserve all possible partial parse trees (just as the parsing process does), whether they subsume the goal tree or not, as long as at least one tree in the current state *does* subsume the goal tree. A generator state must therefore rule out only pairs (S, P) for which P contains no trees which subsume the goal tree, rather than thinning the set P directly via the subsumption check as proposed by O&P.

Transition Effects Just as with alignment, the change in reference of the indexicals *I* and *you* across the speaker/hearer transition (example (5)) emerges straightforwardly from the nature of their lexical actions, with their use at any point involving reference to the speaker or addressee at the time of use:

(5)
A: Have you read . . .
B: Your latest chapter?

Note that there is no constraint on when in

[8] We have little to say about exactly *when* transitions occur. Presumably speaker pauses and the availability to the hearer of a possible goal tree both play a part.

[9] Figure 3 contains several simplifications to aid readability, both to tree structure details and by showing parser/generator states as single triples/pairs rather than sets thereof.

[10] Of course, if both A and B share the same lexical entries and communication is perfect, $P_t = P'_t$, but we do not have to assume that this is the case.

the utterance the transition point can occur, as might be the case in head-driven approaches where transition prior to the sentential head would be problematic. In addition, as quantifier scope-dependency constraints form part of the contextual tree under construction and are not evaluated until a complete type t formula has been derived, dependencies between the portions either side of transition are unaffected, even when some quantifying expression is taken to be dependent on a quantifying term introduced after the role switch:

(6)
A: Did a nurse ...
B: See every patient?

This latter case turns on the (Kempson et al., 2001) account of quantification, in which indefinites are exceptional in projecting a metavariable in their scope-dependency statement allowing choice of term on which to be construed as dependent, even, parallelling expletive pronouns, including some term subsequently constructed.

6 Summary

The left-to-right incrementality and monotonicity of DS, together with the close coupling of parsing and generation processes, allow shared utterances to be modelled in a straightforward fashion. Alignment phenomena can be predicted given a suitable model of context already motivated by the DS treatment of anaphora and ellipsis. A prototype system has been implemented in Prolog which reflects the model given here, demonstrating shared utterances and alignment phenomena in simple dialogue sequences.

Acknowledgements

This paper is an extension of joint work on the DS framework with Wilfried Meyer-Viol, on defining a context-dependent formalism with Ronnie Cann, and on DS generation with Masayuki Otsuka. Each has provided ideas and input without which the current results would have differed, although any mistakes here are ours. Thanks are also due to the anonymous reviewers. This work was supported by the ESRC (RES-000-22-0355) and (for the second author) by the Leverhulme Trust.

References

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

H. Branigan, M. Pickering, and A. Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75:13–25.

A. Cleland and M. Pickering. 2003. The use of lexical and syntactic information in language production. *Journal of Memory and Language*, 49:214–230.

M. Dalrymple, S. Shieber, and F. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4):399–452.

S. Garrod and A. Anderson. 1987. Saying what you mean in dialogue. *Cognition*, 27:181–218.

A. Joshi and S. Kulick. 1997. Partial proof trees as building blocks for a categorial grammar. *Linguistics and Philosophy*, 20:637–667.

R. Kempson, W. Meyer-Viol, and D. Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.

O. Lemon, A. Gruenstein, R. Gullett, A. Battle, L. Hiatt, and S. Peters. 2003. Generation of collaborative spoken dialogue contributions in dynamic task environments. In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*.

W. Meyer-Viol. 1995. *Instantial Logic*. Ph.D. thesis, University of Utrecht.

R. Millikan. 2004. *The Varieties of Meaning*. MIT Press.

M. Otsuka and M. Purver. 2003. Incremental generation by incremental parsing. In *Proceedings of the 6th CLUK Colloquium*.

M. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, forthcoming.

M. Poesio and H. Rieser. 2003. Coordination in a PTT approach to dialogue. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck)*.

M. Purver and M. Otsuka. 2003. Incremental generation by incremental parsing: Tactical generation in Dynamic Syntax. In *Proceedings of the 9th European Workshop in Natural Language Generation*.

M. Stone and C. Doran. 1997. Sentence planning as description using tree-adjoining grammar. In *Proceedings of the 35th Annual Meeting of the ACL*.

Pointing in Dialogue

Hannes Rieser
SFB 360 "Situated Artificial Communicators"
Bielefeld University
Postfach 10 01 31,
D-33501 Bielefeld, Germany
Hannes.Rieser@Uni-Bielefeld.de

Abstract

A new approach based on experiments aiming at the integration of content originating from pointing plus definite descriptions (objects called "CDs") in dialogue is presented. We develop it against the background of the early semiotic positions of Wittgenstein, Peirce, and Quine, the intentionalism of Kaplan, *Neo*-Peirce-Wittgenstein-Quine approaches and "mixed" points of view. Our experimental data show that pointing gestures are polysemous and polymorphic entities. Polysemy of CDs is due to their different functions, pointing to objects ("object demonstration") and pointing to regions ("restrictor demonstration"), polymorphism originates from different positions wrt the utterance. Gesture information and expression meaning are integrated into a syntax-semantics interface using constraint-based syntax and type-logical semantics. Finally, it is shown that an underspecification account for the syntax-semantic interface can be set up along the lines of Logical Description Grammars.

1 Overview and Introduction

Ch. (1) deals with constraints of pointing gestures and introduces the "↘"-notation for gesture strokes. (2) overviews Peircian to *post*-Kaplan approaches on demonstration and reference. (3) describes gesture experiments. (4) is on "object demonstration" and "restrictor demonstration". The set-up of the interface combining constraint-based grammar and type-logics for the integration of multimodal content is specified. (5) deals with the logical form of CDs. (6) shows that an underspecification account for the syntax-semantic interface can be set up along the lines of Logical Description Grammars. Discussion and future research come in (7).

Demonstration is bound up with reference (see e.g. Levinson 1995). Demonstrations (characteristically pointings) can accompany simple or complex referring expressions. We represent the stroke of hand gestures (see Mc Neill 1992) and similar devices by "↘". Up to section (4) the nature of the↘ sign will be left to intuition. It occurs at the position indicated in the string and marks gesture stroke occurrence.

Examples of CD-expressions:

(1) Grasp↘this/that.

(2) *grasp.

(3) *↘

(4) Grasp ↘this/that yellow bolt.

(5) * Grasp this/that yellow bolt.

(6) Grasp the yellow bolt.

(7) ↘This yellow bolt, grasp it.

(8) All the bars get fixed by ↘this yellow bolt.

(9) ↘This yellow bolt doesn't fix all the bars.

(10) ↘This yellow bolt must fix all the bars left of it.

(7) shows a CD-expression taken up by an anaphora; *it* comprises the content provided by *this yellow bolt* and the "↘" together. (8), (9) and (10) show scope interactions of CDs and either quantifier phrases ((8) and (9)), negation or modals ((9) and (10)).

(11) $\searrow_1$This/that is different from $\searrow_2$this/that.

(12) $\searrow_1$This/that, $\searrow_2$this/that, and $\searrow_3$this/that goes into the box.

(11) and (12) have different occurrences of ↘. Anaphora, scope-like effects and multiple occurrences of ↘s are among the most convincing cases for an integrated treatment of demonstratives and demonstrations. Three things have to be considered if we want to get a fuller understanding of CDs: (a) demonstrations and their timing wrt to speech, (b) the structure of verbal expressions going into CDs, and (c) the interaction of demonstrations and expressions, i.e. what they individually contribute to the semantic or pragmatic information provided by CDs *in toto*.

2 Related Research: From Peirce to Kaplan and Beyond

A unified account of CDs will opt for a compositional semantics to capture the information coming from the verbal and the visual channel. Peirce (1932, p. 166) and Wittgenstein (1958, p. 109) consider pointing as part of the symbol. Quine (1960) is committed to a similar point of view.

At present one can distinguish three mainstream philosophical attitudes towards CDs: The line of thought farthest off the Peirce-Wittgenstein-Quine line is the intentionalism associated with Kaplan's late work (1989b). There demonstration is taken as a mere externalisation of intention. It is intention that determines reference. Later on Reimer (1992), Dever (2001), King (2001) and Borg (2000) have supported this line.

Neo-Peirce-Wittgenstein-Quinians (*neo*-PWQians) exist as well. A case in point is McGinn (1981). He holds that in establishing reference the gesture functions as part of the language. Larson and Segal (1995), Hintikka (1998) and ter Meulen (1994) also sympathize with this view. However, there is no *neo*-PWQian approach explicitly representing pointing gestures and providing a semantics for them.

Still, there is a group of "in-betweeners", stressing the contribution of intention and demonstration in fixing demonstrative reference: Among these are the early Kaplan (1989a), D. Braun (1994, 1996) and Lepore and Ludwig (2000). Of these only D. Braun explicitly represents demonstrations in his 1996 approach.

The literature referred to rests almost exclusively on intuitions concerning pointings to single, visible objects. However, pointing is a more varied phenomenon as experiments show.

3 Gesture Experiments Using Simple Reference Games

Reliable intuitions for demonstration are hard to come by. Therefore we use experimental data called "simple reference games" (see Kühnlein and Stegmann (2004)). These are set up in the following way: We have two subjects, description-giver and object-taker. The description-giver must give sufficient information to the object-taker to make him identify one of the objects on a table between them.

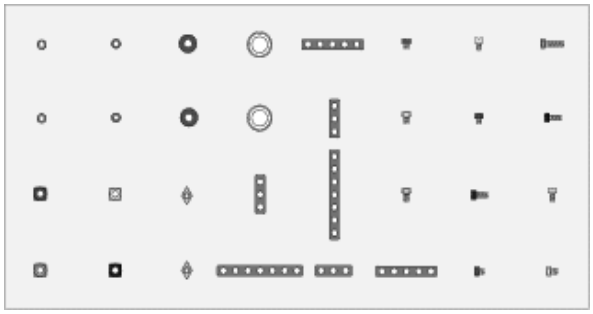


Fig.1: One type of clustering used for gesture experiments in simple reference games (Kühnlein and Stegmann (2004)).

The description-giver selects objects providing the identification information by verbal or gestural means. The object-taker may grasp the object singled out and lifts it from the table. Every game was video-taped from two perspectives, the description giver's as well as a neutral one (fig. 2). There were two types of clusterings, sameness of colour *vs* sameness of form. The games in two video-films have been annotated using the TasX-annotator (Milde and Thies (2002)) (fig. 3); location of gesture stroke, syntax and semantics of NL-expressions and the structure of discourse have been considered. Lack of space prevents us from

Fig. 2: Video-taped pointing gesture from two perspectives

going into descriptions of graspings and of dialogue structure here. Also, discussion of interesting statistical details must be left out (but see Lücking, Rieser, Stegmann (2004)).

4 CDs and Definite Descriptions: Object Demonstration and Restrictor Demonstration

There is a debate on whether definite NPs plus demonstrations can be regarded as definite descriptions (Kaplan 1989 a,b; Rieber 1998). A plea for taking CDs as definite descriptions comes from Quine (1960, ch. III). For us, definite NPs are definite descriptions to which demonstrations add content, either by specifying an object independently of the definite description or by narrowing down the description's restrictor. We call the first technique "object demonstration" and the second one "restrictor demonstration". Graspings are the clearest cases of object demonstration.

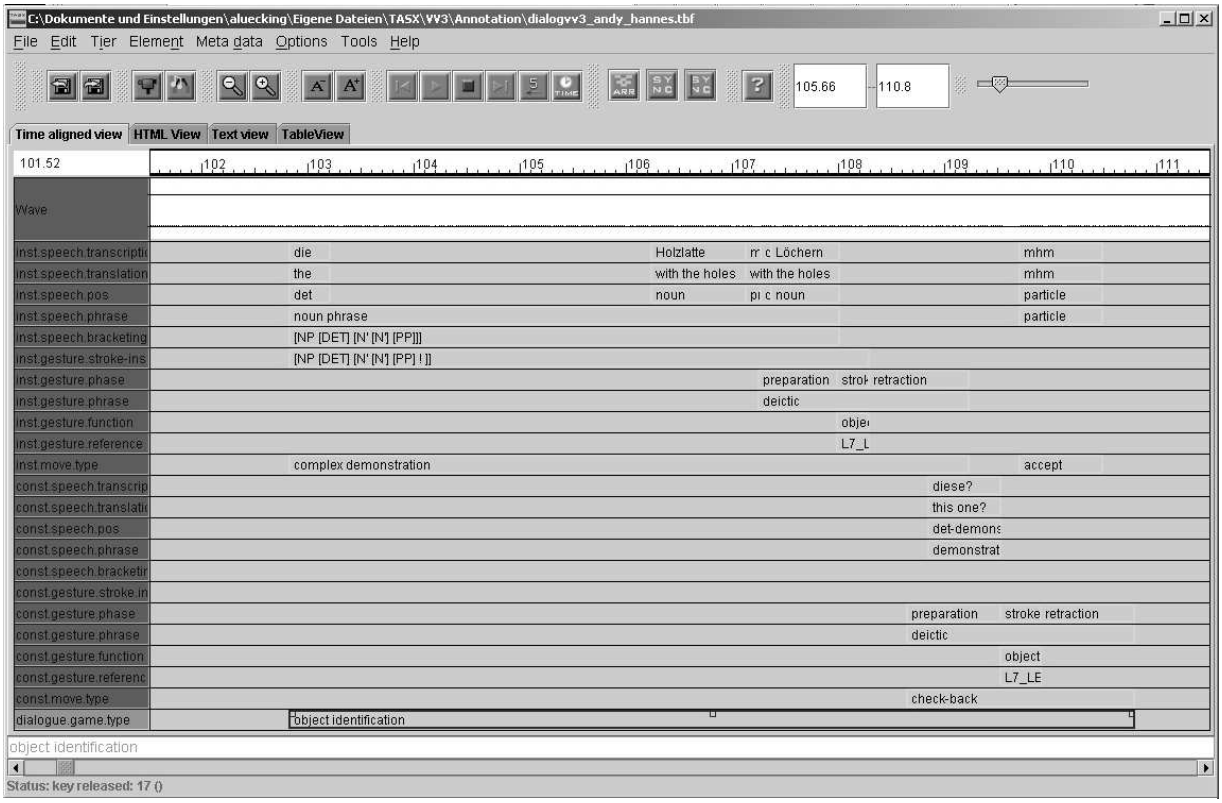


Fig.3: TasX-annotated dialogue game *object identification* comprising instructor's complex demonstration, constructor's check-back and instructor's acceptance

The Syntax-semantics Interface Used

Figs. 4 and 5 show the components of the interface used. Fig. 4 sketches the interpreted grammar, Fig. 5 its empirical coverage.

Following Sag and Wasow (1999), the interface uses constraint based grammar. It combines syntactic and semantic information in one AVM-format. Because of technical reasons we use type-

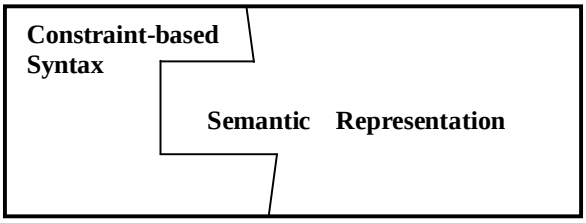


Fig. 4: Components of the constraint-based syntax-semantics interface

logics for semantic representation, i.e. in the values of SEM-attributes, and interpret Sag and Wasow's ⊗-operator (Sag and Wasow (1999), p. 116) as functional application. Due to limits of space, we only represent logical forms here and neglect linking proper.

Following Searle and Vanderveken (1989), directives consist of an illocutionary role indicator and the proposition. They do not have truth conditions in the classical sense. The interpretation specifies satisfaction conditions for them, i.e. it singles out successful directives wrt models. The generalised notion of satisfaction used here is Recanati's (1993). In this context the relation of the definite NP to "its" demonstration is of decisive importance: If both serve their referential tasks, one condition for the satisfaction of the directive is met. The model provides conditions both for the illocutionary role and the associated proposition. Translation of the type-logical format into dynamic semantics is possible in principle (see Eijk and Kamp (1997)).

The model handles canonical uses of CDs, even cases where the demonstration follows the definite NP. Its coverage subsumes real world experimental data as well as *VR*-data.The working of the semantic component will be illustrated here discussing the toy example *Grasp the yellow bolt!*. The meaning of directives is identified with their illocutionary forces.

The imperative is represented by the illocutionary role marker F_{dir} operating on the open formula *grasp(u, v)*, the difference between imperatives and other finite forms being expressed by "F_{dir}". $\lambda P.P$ (you) $\otimes$ $\lambda u(F_{dir}$ (grasp(u, $\iota z(yb(z)))))$ gives us the representation of the whole directive F_{dir}(grasp(you, $\iota z(yb(z))))$ to paraphrase as "There is exactly one yellow bolt, grasp it, addressee!"

So far we have not integrated demonstrations.

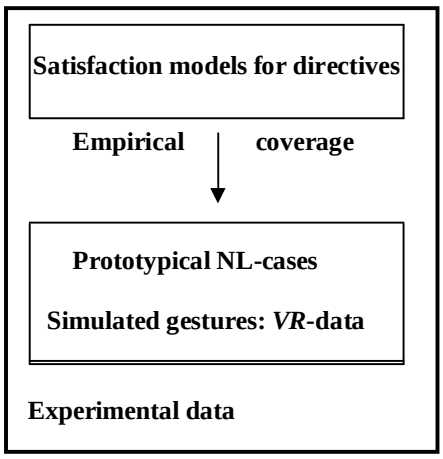


Fig. 5: Models for the constraint-based interface and their empirical coverage

5 Logical Forms for Multi-modal Content

5.1 Integrating Demonstrations into Descriptions

Before we show how to represent demonstrations together with descriptions, we specify our main hypotheses concerning their integration. These are related to content, compositionality, *i.e.* role in building up truth-functional content for the embedded proposition, and scope of gesture. Hypothetically then, demonstrations (a) act like verbal elements in providing content, (b) interact with verbal elements in a compositional way, (c) may exhibit forward or backward dynamics, (d) involve a continuous movement over a time interval, comparable to suprasegmentals, and (e) can be described using discrete entities like the "↘".

Demonstrations introduce objects independently of the definite description ("object demonstration") or act as restrictors of descriptions

("restrictor demonstration"). Intuitively, this will invest demonstrations with two functions. However, this does not yet entail that they are ambiguous between two readings, regardless of the position of the stroke. There still could emerge arguments for a division of labour concerning semantics and pragmatics. Before we enter modelling gesture stroke, we report on the findings concerning stroke position from the empirical data. All findings are corroborated by statistical material (see Lücking, Rieser, Stegmann (2004)):

(1.) Stroke positions can be *pre*-N', *on*-N' or *post*-N'. Here data exhibit greater variation than commonly assumed: Demonstration does not occur before referring expressions unexceptionally. The proto-typical stroke position is on-N'. (2.) Demonstrations can fail. Descriptions they are associated with can denote nevertheless. In particular: satisfiable object demonstrations and corresponding non-satisfiable descriptions yield false propositional content. (3.) Object demonstration and restrictor demonstration are clearly separable and seem to cover together the classifiable data. (4.) Stroke positions do not indicate object demonstration or restrictor demonstration preferences. (5.) A failing description can be completed by a restrictor demonstration. We can have elliptical descriptions in CDs. (6.) In case the description is satisfied on its own, a successful restrictor demonstration is redundant. (7.) Non-classifiability can arise with respect to stroke, direction or role of demonstration, description or completed description.

The central problem is of course how to interpret demonstrations. This question is different from the one concerning the ↘'s function tied to its position in the string. We base the discussion on the following examples showing different empirically found ↘ positions and turn first to "object demonstration":

(13) Grasp ↘this/that yellow bolt.

(13a) Grasp this/that ↘yellow bolt.

(13b) Grasp this/that yellow ↘bolt.

(13c) Grasp this/that yellow bolt↘.

5.2 Object Demonstration

Our initial representation for the propositional frame of the demonstration-free expression is

(14) $\lambda P\, \lambda u(P\ \ \lambda v\ F_{dir}\ (\text{grasp}(u, v)))$.

The ↘ provides new information. If the ↘ is independent from the reference of the definite description the only way we can express that is by extending (14) with $v = y$:

(15) $\lambda P\ \lambda u\ \lambda y(P\ \ \lambda v\ \ F_{dir}\ (\text{grasp}(u, v)\ \wedge\ (v = y)))$.

The idea tied to (15) is that the reference of *v* and the reference of *y* must be identical, regardless of the way in which it is given. Intuitively, the reference of *v* will be given by the definite description *ιz(yb(z))* and the reference of *y* by the ↘. We could also work with a free variable, which, however, would have a different effect (see below).

The Compositionality Problem Concerning Strokes

(15) or the free variable solution may be interesting options for type-logical expressions integrating referential expressions and demonstrations. However, an intuition frequently put forth is that demonstrations to objects act like constants in standard logical notation. Whichever route we want to follow, one thing is common to the three solutions: demonstrations are taken as referring terms, that is, we can represent them as either

(16) (a) $\lambda P \lambda x.P(x)$ (bound variable)

(b) $\lambda P.P(x)$ (free variable)

(c) $\lambda P.P(a)$ (constant)

(a), (b) and (c) do different things: (a) and (b) contribute content *via* an assignment, whereas (c) does so *via* the model's interpretation function.

In order to get a logical form for the whole directive, we must decide on the position of the ↘ in the string. We opt for (13), *Grasp ↘this/that yellow bolt.*, which intuitively indicates that the reference of the ↘ is independent of the reference of the definite description *this/that yellow bolt*.

The bracketing assumed for the string is roughly

(17) [grasp [↘ this/that yellow bolt]].

This implies we have to find a representation for *grasp* which combines with $\lambda P.P(a)$ first, followed by the definite description. A workable solution for this problem is (18), as the derivation based upon it shows:

(18) $\lambda Q \lambda P\, \lambda u\, (P\ (Q\ (\lambda y\ \lambda v F_{dir}\ (\text{grasp}(u, v) \wedge (v = y)))))\ \lambda P.P(a)$ /*[grasp ↘]

(19) $F_{dir}\ (\text{grasp}(\text{you}, \iota z(yb(z))) \wedge \iota z(yb(z)) = a)$.

What can one say about (18)? There the reference is coded twice, once via the pointing gesture $\lambda P.P(a)$ and once via the description $\iota z(yb(z))$. The information exchange, so to speak, maintains a security principle. In most empirical data, however, demonstrations and verbal information show a sort of "division of labour". We now turn to these cases.

5.3 Restrictor Demonstration

(13a) and (13b) above are the prototypical cases where demonstration is embedded into the description, hence the only thing that matters there is the set up of the description. Object demonstration case and restrictor demonstration case are similar insofar as information is added. In the object demonstration case, this is captured by a conjunct with identity statement; in the restrictor demonstration case the ↘ contributes a new property narrowing down the verbally expressed one. The bracketing we assume for the string is roughly

(20) [[grasp] [this/that [↘yellow bolt]]].

As a consequence, the format of the description has to change. This job can be done by

(21) $\lambda D \lambda F \lambda P.P(\iota z(F(z) \wedge D(z)))$.

The demonstration "↘" in (13a) will then be represented simply by

(22) $\lambda y(y \in D)$,

where *D* intuitively indicates the demonstrated region in the domain. We use the $\in$-notation here in order to point to the information from the other channel. Under "⊗" this winds up to

(23) $\iota z(yb(z) \wedge z \in D)$.

Intuitively, (23), the completed description, indicates "the demonstrated yellow bolt".

Different stroke positions come with different compositionality problems.

6 Polymorphism of ↘ Captured in an Underspecification Account

To see what the real problems are if we want to get a stab at multimodal semantics, consider the possible stroke positions marked in the labelled bracketing of example (13):

(24) $[_S [_S [_{VP} [_{Vinf}$ grasp$]$ ↘$_{pre\text{-}NP}[_{NP} [_{Dem}$ this/that$]$ ↘$_{pre\text{-}N'}$ $[_{N'}$ ↘$_{pre\text{-}Adj}[_{Adj}$ yellow$]$ ↘$_{pre\text{-}N}[_N$ bolt$]$ ↘$_{post\text{-}N}]$ ↘$_{post\text{-}N'}]$ ↘$_{post\text{-}NP}]$ ↘$_{post\text{-}VP}]]$ *↘$_{post\text{-}S}]$

We have pre-occurrences and post-occurrences of ↘. The pre-occurrences are ↘$_{pre\text{-}NP}$, ↘$_{pre\text{-}N'}$, ↘$_{pre\text{-}Adj}$, ↘$_{pre\text{-}N}$; these are the post-occurrences: ↘$_{post\text{-}N}$, ↘$_{post\text{-}N'}$, ↘$_{post\text{-}NP}$, ↘$_{post\text{-}VP}$. ↘$_{post\text{-}S}$ we consider as not well-formed. At the same time, every occurrence can be paired with at least two readings, that is where the polysemy comes from. Seen from the point of view of our type-logical formulas for "object demonstration" and "restrictor demonstration",

(25) $\lambda P.P(a)$ and (26) $\lambda y\ (y \in D)$

we get the problem that there emerges a clash between the "natural" context-free syntactic category of ↘ and ist semantic function. We won't solve that entirely here. Clearly, all the "post"-occurrences of ↘ are problematic in a way, nevertheless they do occur. By way of solution, we can take up a suggestion of Sag and Wasow's (1999) concerning underspecification and distinguish between descriptions, feature structures and models as follows: Descriptions can be underspecified, feature structures are complete in relevant respects and serve as models for linguistic entities. Underspecified descriptions are satisfied by sets of structures. Seen from this perspective, our discussion so far dealt entirely with the semantic side of structures. Now, we move on to descriptions.

The underspecification model nearest the formalisms used here is the *Logical Description Grammars* (LDGs) account of Muskens (2001), which has evolved *inter alia* from Lexicalised Tree Adjoining Grammars and D-Tree Grammars. The structures derived within LDGs are compatible with those we get in constraint based formalisms using AVMs, hence there is no big methodological

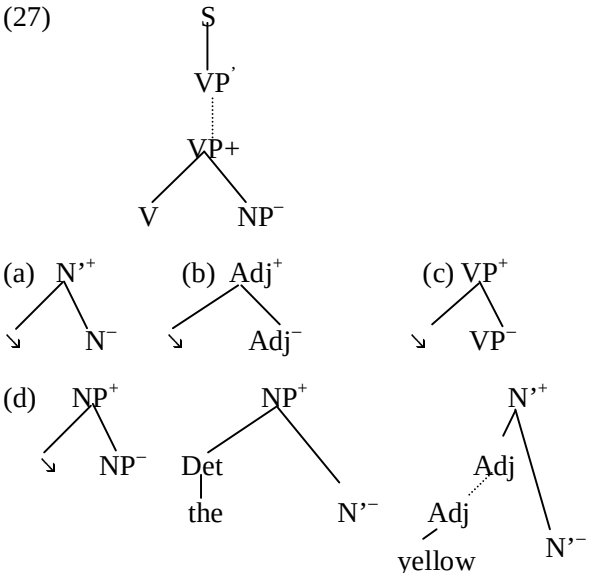
difference between the assumptions made about the theory of grammar here and LDGs. The intuitive idea behind LDGs is that, based on general axioms capturing the structure of trees, we work with a *logical description of the input*, capturing linear precedence phenomena, *lexical descriptions for words* and *elementary trees*. Then a *parsing-as-deduction* method is applied yielding semantically interpreted structures.

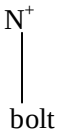
We provide the main steps of an LDG-reconstruction of the readings of (24) below.

A graphical representation of the input is given in (27). '+' respectively '-' indicate components which can substitute ('+') as against nodes to be substituted ('-'). Dotted lines represent dominance and solid ones direct dominance. Models for the description in (27) are in the parsing-as-deduction approach derived by pairing off + and – nodes in a one-to-one fashion and by identifying the nodes thus paired. "I.e., each +node must be identified with a –node and *vice versa*, but not two +s and no two –s can be identified". (Muskens (2001), p. 424). Words can come with several lexicalisations. The ↘-positions in (27) (a) to (d) have to be regarded as alternatives.

The *logical description of the input* has to provide the linear precedence regularities for our example *Grasp the yellow bolt!* Observe that these will be different from (27), which contains alternatives (a) to (d) for ↘-positions. (28) shows some precedence possibilities; the subindices on ↘ are provided to facilitate understanding.

(27)





(28)

(a) grasp < $\searrow_{\text{pre-NP}}$ < this/that < yellow < bolt.

(b) grasp < this/that < $\searrow_{\text{pre-N'}}$ < yellow < bolt.

(c) grasp < this/that < $\searrow_{\text{pre-Adj}}$ < yellow < bolt.

(d) grasp < this/that < yellow < $\searrow_{\text{pre-N}}$ < bolt.

(e) grasp < this/that < yellow < bolt < $\searrow_{\text{post-N}}$.

(f) grasp < this/that < yellow < bolt < $\searrow_{\text{post-N'}}$.

(g) grasp < this/that < yellow < bolt < $\searrow_{\text{post-NP}}$.

(h) grasp < this/that < yellow < bolt < $\searrow_{\text{post-VP}}$.

The description of the input must fix the underspecification range of the ↘. It has to come after the imperative verb, but that is all we need to state; in other words, that covers all the models depicted in (28).

The *lexical descriptions for words* will have to contain the type-logical formulas for compositional semantics. From the descriptions of the *elementary trees* we will get the basics for the "pairing-off" mechanism. It is easy to see that we can establish a proof for the NP with $\searrow_{\text{pre-NP}}$ yielding (28)(a). (27) (a), (b) allow us to extend the NP with $\searrow_{\text{pre-N'}}$ and $\searrow_{\text{pre-Adj}}$, respectively. The "post"-versions could be generated by lexical anchors roughly similar to (27)(a) to (d). Lack of space prevents us from explaining here what has to be done at the type-logical level to ensure compositionality and well-formedness.

7 Discussion and Future Research

One of the central questions is of course whether there is an alternative to the *neo*-PWQian point of view and the ensuing methodology. A PWQian approach leads quite naturally to an integrated theory. A viable alternative might be to try an approach stressing the difference (!) between NL-expression and demonstration and to capture the role of demonstration in a different way, perhaps solely *via* the semantic model for the formal description chosen. Seen from this perspective, demonstration is an object with semantic impact

but it is not part of the language. By and large this would be a Kaplan point of departure.

Keeping within *neo*-PWQian assumptions, the following points concerning the approach described here seem worthy of mentioning and need further detailed study: How can polymorphism/polysemy of demonstration be handled best? Will Logical Description Grammars do all there needs to be done? And, which division of labour between semantics and pragmatics is the correct one for setting up a theory of CDs? In addition, describing the simple reference games familiar from the data in a real discourse games approach is a worthwhile target but the other problems have to be sorted out first.

References

Almog, J., Perry, J., Wettstein, H. (eds.): 1989, *Themes from Kaplan*. New York, Oxford: OUP

Borg, E.: 2000, Complex Demonstratives. In: *Philosophical Studies 97*: 229 – 249

Braun, D.: 1996, Demonstratives and Their Linguistic Meanings. In: *Nous* Vol. 30, Nr. 2, pp. 145- XX

Braun, D.: 1994, Structural characters and complex demonstratives. In: *Philosophical Studi*es, Vol. 74, Nr. 2, pp. 193-221

Dever, J.: 2001, Complex Demonstratives. In: *Linguistics and Philosophy*, Vol. 24, Nr. 3. pp. 271-330

Hintikka, J.: 1998, Perspectival Identification, Demonstratives and "Small Worlds". In: *Synthese* 114; pp. 203 – 232

Kaplan, D.: 1989a, Demonstratives. In Almog et al. eds., pp. 481-563

Kaplan, D.: 1989b, Afterthoughts. In Almog et al. eds., pp.565-614

King, J.C.: 2001, *Complex Demonstratives. A Quantificational Account.* Cambridge, Mass.: MIT Press

Kühnlein, P. and Stegmann J.: 2004, *Referring to Objects in Simple Identification Tasks*. Technical report of the SFB 360

Larson R. and Segal, G.: 1995, Pronouns and Demonstratives. Ch. 6 of *Knowledge of Meaning*. The MIT Press: Cambridge, Mass, pp. 197 – 227

Lepore, E.: 2000, Semantics and Pragmatics of Complex Demonstratives. In: *Mind*, Oxford, Vol. 109, Nr. 434, pp. 199-240

Levinson, St. C.: 1995, *Pragmatics*. Cambridge: CUP

Lücking, A, Rieser, H. and Stegmann, J.: 2004, Statistical Support for the Study of Structures in Multi-Modal Dialogue. *Catalog04 Proceedings*, pp.

Mc Neill, St.: 1992, *Hand and Mind*. The University of Chicago Press: Chicago and London

Milde, J.-T. and Thies, A.: 2002, *The TasX environment: Owner's Manual*. Ms, Bielefeld, Univ.

Muskens, R.: 2001, Talking about Trees and Truth-Conditions. In: *Journal of Logic, Language and Information*, pp. 417 - 455

Quine, W. van O.: 1960, *Word and Object*. MIT Press, ch. III, The Ontogenesis of Reference

Recanati, F.: 1993, *Direct Reference. From Language to Thought*. Oxford, UK: Blackwell

Rieber, St.: 1998, Could demonstratives be descriptions? In: *Philosophia*, Vol. 26, Nr. 1-2, pp. 65-79

Sag, I. A. and and Wasow, Th.: 1999, *Syntactic Theory: a formal introduction*. Stanford, Calif.: CSLI Public.

Searle, J. and Vanderveken, D.: 1989, *Foundations of Illocutionary Logic*. Cambridge: CUP

Stanley, J. and Gendler, S.: 2000, On Quantifier Domain Restriction. In: *Mind & Language*, Vol 15, pp. 219-261

Ter Meulen, A. G. B.: 1994, Demonstratives, indications and experiments. In: *The Monist*, Vol. 77, Nr. 2, pp. 239-256

Van Eijk, J. and Kamp, H: 1997, Representing Discourse in Context. In: Van Benthem *et alii* (eds.), *Logic and Language*. North-Holland, pp. 179 – 239.

Wexelblat, A.: 1998, Research Challenges in Gesture: Open Issues and Unsolved Problems. In: Wachsmuth, I. Fröhlich, M. (eds.): *Gesture and Sign Language in Human-Computer Interaction*. International Gesture Workshop Bielefeld, Germany, September 1997 Proceedings, pp. 1-13

Form, Intonation and Function of Clarification Requests in German task-oriented spoken dialogues

Kepa Joseba Rodríguez and **David Schlangen**
University of Potsdam
Department of Linguistics / Applied Computational Linguistics
P.O. Box 601553
D-14415 Potsdam — Germany
{rodriguez|das}@ling.uni-potsdam.de

Abstract

We present a classification-scheme for describing the form (including intonation) and function of clarification requests (CRs) that is more fine-grained than extant classifications, and a study of a corpus of German task-oriented dialogues where we used this scheme to annotate the occuring CRs. Among the correlations between form and function we found was a hitherto undescribed correlation between intonation of CRs and their interpretation, which could possibly aid dialogue systems in interpreting CRs.

1 Introduction

Clarification requests (CRs), as exemplified by B's utterances in the mini-dialogues in (1), are of eminent theoretical as well as practical interest.

(1) a. A: Well, I've seen him.
B: Sorry, you *have* or you *haven't*?

b. A: Did you talk to Peter?
B: Peter Miller?

c. A: Did you bring a 3-5 torx?
B: What's that?

They are of theoretical interest because they are a prime example of a dialogue move that is concerned more with dialogue management than with conveying propositional information, and hence goes beyond what formal semantics was invented to model. Arguably even stronger is the practical interest in modelling CR, since practical dialogue systems are constantly confronted with situations where it would be beneficial if they could clarify their understanding of a user's utterance, or where they must interpret a clarification requested by the user.[1] (To give an impression of the frequency of this phenomenon even in human-human dialogue, in our corpus we found that around 5.8% of all turns were CRs.)

In this paper we hope to further both lines of inquiry, by offering a theoretically motivated and practically usable classification of CR uses and of CR forms, and by investigating the link between the two in a corpus of German spoken dialogues. While we replicate (for a different language) some of the results of earlier studies (Purver et al., 2001; Purver et al., 2003), we argue for, and show the use of, an analysis of form and function that is more fine-grained than that underlying those studies. We also make use of the fact that we had available information about intonation in our corpus—a feature that significantly influences the interpretation of CRs, as we show, and that could be used in practical dialogue systems to disambiguate CRs.

The remainder of this paper is organised as follows. In the next section we describe our multidimensional classification of form and function of

[1] The semantics of (some kinds of) CRs is modelled for example in (Ginzburg and Cooper, 2001; Larsson, 2003); see below for some remarks on the former analysis. There is a vast literature on dealing with clarifications in spoken dialogue systems, some very recent examples taking a more theoretical perspective include (Gabsdil, 2003; Larsson, 2003; Schlangen, 2004).

CRs and compare it to earlier classification attempts. In Section 3 we give details about the corpus study we conducted, whose results we present and discuss in Section 3.2. In particular, we discuss the links between form-features and function that are present in the corpus. We close with a discussion of the overall result, and of possible further work.

2 Classification of CRs

2.1 Earlier work

In a number of papers, the most recent of which is (Purver et al., 2001) (henceforth PGH), Jonathan Ginzburg and colleagues have put forward a scheme for classifying form and function of CRs, which we will now discuss and relate to the one proposed in this paper.

PGH classify CR-forms using the classes shown on the left in Table 1. While these classes achieve good coverage on the corpus (PGH report that only 0.5% of CRs were classified as `other`), we wanted to explore the influence of individual features of the form on the interpretation in more detail, and hence we further analysed these classes and devised a multi-dimensional classification. We will describe our schema in detail below, but to give an example of how it relates to PGH's, in our schema we 'factor out' the component "reprise" that is found in several of PGH's classes into a feature "relation to the antecedent" (`rel-antec`), which can take the values `repetition`, `reformulation` and `independent`, independently from other features. This allows us to make finer distinctions, for example between "Paris?" and "The capital of France?", which as a reply to "I'm going to Paris." would both be classified as `frg` by PGH. Our multi-dimensional approach also allows us to emphasise *similarities* between forms; for instance, PGH's classes `frg` and `lit` have in common in our approach the value for a certain feature (both are literal repetitions of material from the antecedent utterance), while having different values for other features. Using such fine-grained features, we can test for more fine-grained correlations between form and function.

While PGH's classification of CR-forms seems to be generally correct (just not as fine-grained as possible), their classification of CR functions (or *readings*, as they call them), shown on the right in Table 1, seems more problematic. In particular what they call the *clausal* reading of CRs seems to be difficult in practice to delineate from the other readings they define. For instance, given a situation as shown in (2), it is not clear why the *clausal* reading should not be able to play the function the authors assign to the *constituent* reading, namely to clarify a referent. (The other direction is more clearly distinguished: unlike the *clausal* reading, the *constituent* reading *cannot* clarify an acoustic problem.)

Class	Description	Example
non	Non-Reprise	*"What did you say?"*
wot	Conventional	*"Pardon?"*
frg	Reprise Fragment	*"Paris?"*
slu	Reprise Sluice	*"Where?"*
lit	Literal Reprise	*"You want to go to Paris?"*
sub	Wh-substituted Reprise	*"You want to go where?"*
gap	Gap	*"You want to go to ...?"*
fil	Gap Filler	*"... Paris?"*
oth	Other	Other

Class	Description	Paraphrase
cla	Clausal	"Are you asking/telling me that ...X..?"
con	Constituent	"What/who do you mean by 'X'?"
lex	Lexical	"Did you utter 'X'?"
corr	Correction	"Did you intend to utter X (instead of Y)?"
oth	Other	Other

Table 1: CR forms and readings as classified by (Purver *et al.* 2001)

(2) A: Did Bo leave?
B: Who?
clausal: For which x are you asking whether x left?
constituent: Who's Bo?

Moreover, it seems difficult to integrate CRs asking for clarification of intentions into this scheme:

(3) a. A: Push the red button.
B: Why?
b. A: Turn it on.
B: By pushing the red button?

To summarize, the problem seems to be that the readings defined by PGH still abstract over different reasons why one might want to make a CR—

they are still too close to the ambiguity of "what did you say?".[2] For these reasons we will in the next section propose a different classification of CR functions; first, however, we turn again to the *form* of CRs.

2.2 Surface form of CRs

We now go through the features we use to describe the form of the CRs.[3] A few selected examples for the different types are shown at the end of this section in (4).

Mood The possible values of the attribute `mood` are: a) *declarative*; canonical declarative word order or fragment without a verb with falling end-boundary tone.[4] b) *polar question*; fully realised syntactic polar interrogatives. c) *alternative question*; d) *wh-question*; e) *imperative*; f) *other*.

Completeness The possible values for the attribute `completeness` are: a) *particle*; or conventional phrase, e.g. "pardon?". b) *partial*; a syntactic fragment, normally a phrase. c) *complete*; a syntactically 'complete' sentence.

Relation to the antecedent The possible values for the attribute `rel-antec` are: a) *repetition*; parts of the problematic utterance are repeated *literally*. b) *addition*; something is added to a literal repetition (most often a *wh*-word). c) *reformulation*; a phrase is uttered that is co-referent to elements of the original utterance, but is not a literal repetition. d) *independent*; no elements of the problematic utterance are repeated or reformulated.

We also classify CRs according to the intonation with which they are uttered. Specifically, we look at the end-boundary tone, marking it use an encoding that is related to ToBI (Silverman et al., 1992), but somewhat simplified.

Boundary tone The values are: a) *rising* and b) *falling*, which correspond to (X)H% and (X)L%, respectively (X being an arbitrary tone).

A few examples for CRs of the types described above are shown below, with the classification according to the above scheme shown in `typewriter font`.

(4) a. K.: na hinten.
I.: vorne oder hinten?
K.: hinten.
(K.: well, to the back – I.: to the front or to the back? – K.: to the back)
`mood:alt-q,`
`completeness:partial,`
`rel-antec:addition,`
`bound-tone:falling`

b. I.: hm ist doch (ei)n Klacks für dich.
K.: hä?
(I.: hm, that shouldn't be a problem for you – K.: eh?)
`mood:other,`
`completeness:particle,`
`rel-antec:indep,`
`bound-tone:rising`

c. K.: ich hab(e) aber noch zwei Stäbe.
I.: du hast noch zwei Stäbe?
(K.: But I still have two bars. – I.: you still have to bars?)
`mood:decl,`
`completeness:complete,`
`rel-antec:repet,`
`bound-tone:rising`

d. I.: [...] und der grüne sitzt obendrauf.
K.: obendrauf?
(I.: [...] and the green one sits on top of it – K.: on top of it?)
`mood:decl,`
`completeness:partial,`
`rel-antec:repet,`
`bound-tone:falling`

2.3 Function of CRs

We also classify the function of each CR instance according to a multi-dimensional schema. The most important dimension is the one specifying

[2] These readings are realised technically by a straightforward formalisation of these paraphrases in an HPSG framework, using an *illocutionary-act* relation for the clausal reading and a relation *content* for the clausal readings, where both relations take signs as arguments. Since the formalisation is so close to the paraphrases (and is in any case not backed up by a formal semantics of the predicates used), we don't think we miss crucial details by using just the paraphrases in this discussion here.

[3] We initially also used word order as a classification feature, but since it turned out not to have any predictive power as to the possible function of a CR, we do not include it here.

[4] The name of this value is slightly misleading: it covers all cases of non-interrogative word order, i.e. both declarative sentences and fragments, and so a more appropriate (but less immediately understandable) name would be "non-interrogative".

	Level of action	Kind of problem	Example
1	execution / attention	channel	"huh?"
2	presentation / identification	Acoustic problem	"Pardon?"
3	signal / recognition	Lexical problem	"What's a double torx?"
		Parsing problems	"Did *you* have a telescope, or the man?"
		Reference resolution problem:	
		• NP-reference	"Which square?"
		• Deictic-reference	"Where is 'there'?"
		• Action-reference	"What's to kowtow?"
4	proposal / consideration	Problem with recognising or evaluating the intention	"Why?" "You want me to give you this?"

Table 2: Levels of action and associated problems

the likely source of the problem that lead to the need for clarification. This dimension is related to PGH's *readings*, but, as discussed above, needs to be more fine-grained and better defined. As the basis of our classification we use the well-known models of (Clark, 1996) and (Allwood, 1995), to which we add some further (sub-)levels. The other dimensions specify the *extent* and *severity* of the problem, as described below. Lastly, we also group under this heading a classification of the reaction to the CR.

Source of the problem The models of (Clark, 1996) and (Allwood, 1995) describe four levels of action involved in communication, each of which is a possible locus for communication problems. In Table 2 they are represented schematically, together with a specification of the kinds of problems that can occur on these levels, and some examples. As this specification shows, the levels can be further subclassified, and this we have done for our classification.[5]

The possible values for this feature correspond to the column "kind of problem" in the table. For reasons of space, we can only give the constructed examples in the last column of the table here.

Extent This feature describes whether the CR points out a problematic element in the problem utterance (e.g., "To Paris?", "I didn't hear the second word.") or not; its possible values are *yes* and *no*. Note that this is a *function*-feature, which may or may not be strongly connected to the *form*-feature "fragmental", but is logically independent, as the second example above, a full sentence that points out a problematic element, shows.

Expectation / Severity This dimension describes which action the CR initiator requests from the other dialogue participant, or, to look at it from another perspective, it describes how severe the problem was. The possible values are: a) *repetition/elaboration of previous material*; the CR initiator asks for a repetition/reformulation of material from the move to be clarified, possibly triggered by a complete understanding failure. b) *confirmation of the hypothesis*; the CR initiator asks for a confirmation of her/his understanding about the content of the move to be clarified. I.e., a hypothesis could be drawn, but agent is not confident about its correctness. (4-b) above is an example for the former, (4-a) for the latter.

Reply to the CR This feature classifies the reply to the CR, not the CR itself. Its possible values are: a) *y/n-answer*; b) *repetition*; an answer that repeats an element of the problem utterance literally. c) *reformulation*; an anwer that reformulates an element. d) *elaboration*; an answer that elaborates on (an element of) the problem utterance, adding information. e) *word definition*; an answer to a lexical question ("what does x mean?"). f) *no reaction*; the CR addressee did not react.

Satisfaction of the CR-initiator This feature records the reaction of the CR initiator to the reaction of the CR addressee. The possible values are: a) *happy*; the CR initiator seems satisfied with the reply; this can be taken as an indication that the interpretation of the CR addressee was correct. b) *unhappy*; CR initiator renews request for clarification.

[5] (Gabsdil, 2003) and (Larsson, 2003) similarly use these models to classify CRs, and they are roughly at the same level of fine-grainedness. (Schlangen, 2004) uses a more fine-grained classification that is motivated by a formal semantic / pragmatic processing model, but to strike a balance between detailed analysis of the phenomenon and making annotation possible, we have decided on the fewer levels described here.

3 The Corpus study

3.1 Material and Method

Material We used the Bielefeld Corpus of German task-oriented human-human dialogue (SFB-360, 2000) (the scenario is that one dialogue participant (DP) gives instructions to the other DP to build a model plane), which consists of 22 dialogues, with 3962 dialogue turns and 35813 words.

Method In a first step, we identified the turns containing CRs, which we then annotated for form and function, using the MMAX-tool (Müller and Strube, 2001). Annotation of the form classification features was straightforward, as their values can easily be read off of the surface form of the CR, or, in the case of `rel-antec`, from CR and problem utterance. The *function* of a CR of course cannot as easily be seen from the form—to find whether there is a reliable link is one goal of the present study, after all. We used the reply of the CR addressee, and the reaction of the CR initiator to that reply as a guide for the interpretation that was chosen by the DPs. Hence what we annotated as 'function' could more properly be called "mutually agreed upon interpretation of the CR"—and that is not necessarily what the CR initiator might initially have had in mind. Since "overanswering" in certain configurations systematically addresses several different problem sources (for example, a reformulation of content answers both acoustic understanding problems as well as reference resolution problems), this is a real methodological problem for finding a link between form and problem source. We circumvented this problem by defining ambiguity classes for use in the cases where we could not make a decision; this weakens the overall correlations we report below, but makes the ones we did find between form features and unambiguously identified functions more valid.[6]

[6]This strategy is more cautious than the one chosen by PGH. As they say, in cases of ambiguity "the response(s) of the other DPs were examined to determine which reading was chosen by them. The ensuing reaction of the CR initiator was then used to judge whether this interpretation was acceptable." However, this method is not infallible, as their own example shows:

(i) George: [...] with a piece of spunyarn in the wire.
Anon1: Spunyarn?
George: Spunyarn, yes.
Anon1: What's spunyarn?
George: Well that's like er tarred rope.

PGH use this as an example where the original interpretation was incorrect; however, in our opinion an interpretation seems equally likely where Anon1 first wanted to clarify acoustic understanding, and, once this was accomplished, clarified lexical understanding. To be on the safe side, we annotated such cases with superclasses combining the subclasses it is ambiguous in.

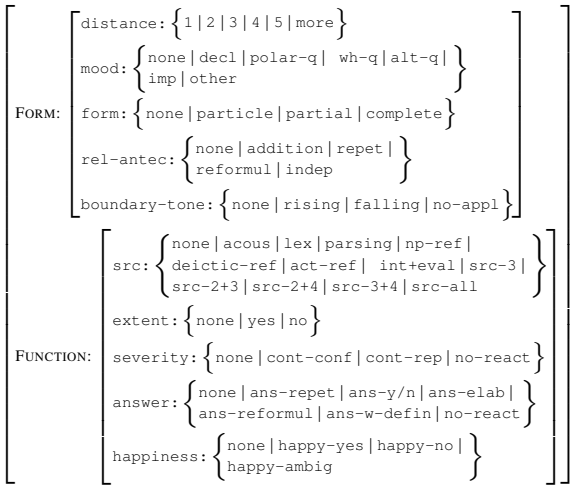


Figure 1: Annotation scheme

Annotation scheme The annotation schema basically just implements the distinctions described in Section 2, with ambiguity classes for function as discussed above. It is shown in Fig. 1. Note that we also recorded the distance between the CR and the problem utterance.

3.2 Results

We identified 230 CRs in the 3962 turns we looked at; this indicates that with 5.8% of turns this is a rather frequent phenomenon in our corpus. (PGH: just under 4%, but their corpus contained general conversation, which might account for the difference.) The results of classifying these instances and of testing for dependence between features are reported in this section.

3.2.1 Distribution

Clarification seems to be a very local phenomenon: in our corpus, 95% of all clarifications target the immediately preceding utterance (PGH: 85%). This high number might reflect the

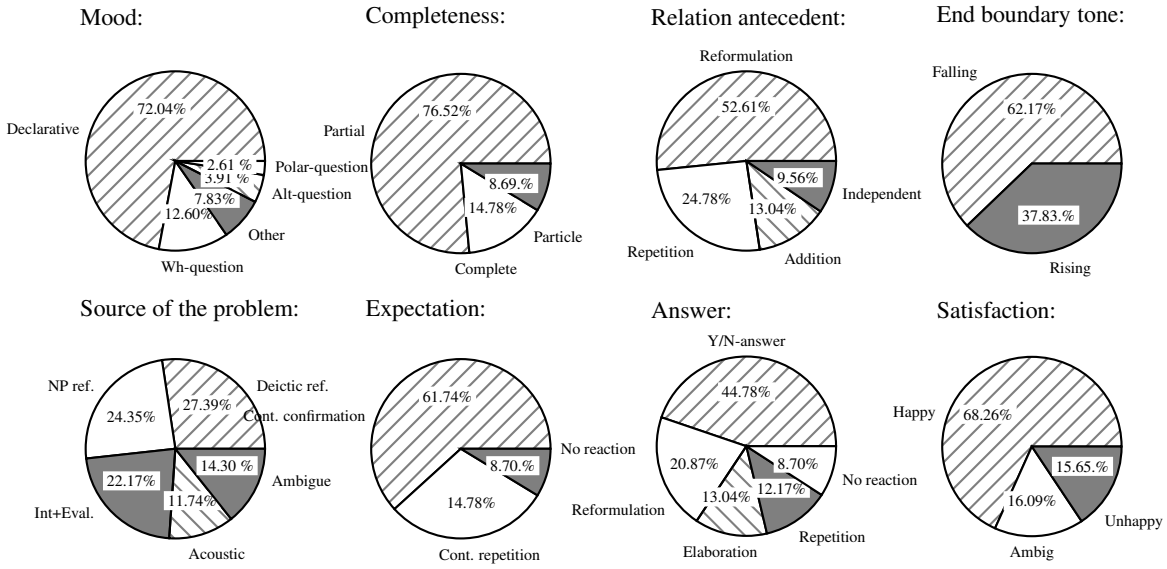


Figure 2: Distribution of values for form-features (top) and function-features (bottom)

task-oriented nature of the corpus, where grounding presumably is more cautious compared to free conversation, and potential problems are clarified immediately.

Distribution of forms The frequencies for values of the form-features are given by the pie-charts in the top row of Figure 2. As this figure shows, the overwhelming majority (76.5%) of the CRs in our corpus were fragmental in form (PGH: 42.4%). Since we separated our analysis into several dimensions, we can further analyse the class of fragmental CRs: 62.6% of them were reformulations of previous content, 24.8% were repetitions. Another distinction not made in earlier studies is that between rising and falling intonation. Using these features, we can access different (sub-)types of what PGH collectively call "reprise fragments". Indeed, four of the five most frequent types of CRs were classified as syntactically partial (i.e., the value for `completeness` is `partial`), with either `falling` or `rising` as value for `bound-tone` and either `repet` or `reform` as value for `rel-antec`. (The one other type in the "top-five" being that of conventional CRs.) We come back to these distinctions when we report the correlations between form and function we found.

Another interesting observation is that most CRs take up material from the problematic utterance in some form, with only 9.6% of CRs being fully independently formulated. Overall, these numbers seem to confirm the findings of PGH regarding distribution of forms, showing that at least for speakers of English and German behaviour with respect to clarification seems comparable—useful to know for designers of dialogue managers for multi-lingual dialogue systems.

Distribution of functions The distribution of values for the function-features can be seen in the bottom row of Figure 2 (with the exception of the feature `extent`, whose two values `yes` and `no` were chosen 87.8% and 12.2% times, respectively).

As this figure shows, the most frequent problems were related to resolving references (just above 50%, with 27.4% clarifying deictic references, and 24.4% clarifying NP reference). 14.3% of the CRs were annotated with a super-class, meaning that their function was ambiguous in the context. However, most instances, and most types, could be classified unambiguously. The distribution of super-classes is instructive, showing for example that the different kinds of problems at level 3 could be distinguished fairly well.

We found only one instance of a lexical problem, making our corpus non-representative for this type of CR. We speculate that the reason is that the vocabulary in this domain is very restricted and domain specific, and known to the DPs. This might also explain the relatively low frequency of

acoustic problems, since a restricted vocabulary may make recognition easier.

Most CRs in our corpus point out a specific element in the problem utterance—this of course correlates with the high number of fragmental CRs found. Only 8.7% of CRs in our corpus failed to elicit a response ((Purver et al., 2003): 17%); again, this seems to be a difference between task-oriented dialogue, where the task demands that problems be clarified, and free conversation.

3.2.2 Correlations

We used χ^2 to test for (in)dependence between features of the surface form and function of CRs (we used Yates' correction to account for cases where due to data sparseness there were expected values below 5), and if there was a significant dependence, Pearson's ϕ to determine the strenght of the correlation. The results of this test are shown in Table 3, where the rows are the form dimensions and the columns those of the function, and the cells show the results of testing for (in)dependence between these variables (showing χ^2, χ^2 with Yates' correction, and Pearson's ϕ). Note that all tests are significant at P=0.001. For reasons of space we can only pick out the most relevant findings here for further discussion.

One very interesting result is that intonation seems to disambiguate fairly reliably between CRs clarifying reference and those clarifying acoustic understanding, with rising boundary tones being significantly more often used to clarify acoustic problems and less often than expected to clarify reference resolution problems, and complementary correlations for falling tones. (The confusion matrix is shown in Table 4.) A similar distinguishing tendency is shown by reformulations vs. repetitions, with the former being significantly often NP reference resolution questions and the latter acoustic clarifications.

Looking at `mood` vs. `answer`, one can see that declaratives in general prompt yes/no-answers (and hence confirmations of hypotheses) more than reformulations of content, which in turn is the most likely reaction to *wh*-questions. These are nice results, showing that despite the fact that both readings are in principle available for fragments (cf. PGH), more clarity is achieved if a dialogue system for example produces such forms only if it wants to get a hypothesis confirmed, and *wh*-questions if it needs more information about an element of the problem utterance. Moreover, if the hypothesis is one about the referent of an NP, a reformulation is the best bet; if it is one about acoustic understanding, a literal repetition might be better.

	`rising`	`falling`	
`int+eval`	24 (21.18)	32 (34.82)	56
`deictic-ref`	**8** (20.43)	**46** (33.57)	54
`np-ref`	**8** (18.91)	**42** (31.09)	50
`acous`	**23** (9.08)	**1** (14.92)	24
`src-2+3`	3 (2.27)	3 (3.73)	6
`src-2+4`	**11** (6.81)	**7** (11.19)	18
`src-all`	**1** (0.76)	1 (1.24)	2
`lex`	0 (0.38)	**1** (0.62)	1
`src-3+4`	9 (7.19)	10 (11.81)	19
	87	143	230

χ^2 Total: 63.23 (YC: 56.59); df = 8; $\phi = 0.52$

Table 4: `src x bound-tone`

3.3 Reliability

Although the complete annotation was only performed once (by one of the authors), we did test for reliability of what is intuitively the most problematic feature, namely `source of the problem`. This feature was annotated by a second annotator, resulting in a κ (Carletta, 1996) of 0.70. While this is not great (values between .67 and .8 are often seen to allow only tentative conclusions), it is comparable to the results reported by PGH (0.75), and reflects the difficulty of the task.

Where we cannot report reliability yet is for the task of identifying CRs in the first place. This is not a trivial problem, which we will address in future work.[7]

4 Summary and Further Work

We have presented a fine-grained classification scheme for form and function of clarification requests. This scheme was used to annotate a corpus of task-oriented dialogues, where about 4% of all turns were found to be CRs—this confirms the observation that clarification is a quite frequent phenomenon. Our fine-grained annotation scheme,

[7] As far as we can see, PGH have not tested for reliability of doing this task either.

	source	severity	extent	answer
mood	*indep.*	χ^2-Σ: 106.52/96.58; df = 8; $\phi = 0.48$	χ^2-Σ: 112.04/101.31; df = 4; $\phi = 0.70$	χ^2-Σ: 72.90/72.64; df = 20; $\phi = 0.28$
bound-tone	χ^2-Σ: 63.23/56.59; df = 8; $\phi = 0.52$	*indep.*	χ^2-Σ: 14.85/13.29; df = 1; $\phi = 0.25$	*indep.*
rel-antec	χ^2-Σ: 142.85/114.62; df = 24; $\phi = 0.46$	χ^2-Σ: 66.16/59.87; df = 6; $\phi = 0.38$	χ^2-Σ: 98.49/90.55; df = 3; $\phi = 0.65$	*indep.*
completeness	*indep.*	χ^2-Σ: 35.88/31.50; df = 4; $\phi = 0.28$	χ^2-Σ: 94.54/86.98; df = 2; $\phi = 0.64$	*indep.*

Table 3: χ^2 values for combinations of form- and function-features

and the fact that we annotated intonation, allowed us to find correlations that have hitherto been unnoticed, such as that described above between intonation of CRs and their relation to the antecedent utterance (repetition or reformulation) on the one hand and reference resolution function or acoustic clarification on the other hand. Information like this could be of much use in dialogue systems that are faced with the task of interpreting CRs by the user which in theory are often multiply ambigous.

In further work we plan to connect our findings to general theories of the interpretation of intonation in discourse (e.g. (Gunlogson, 2001)), and we also plan to collect more data, with which then automatic classifiers could be trained. Another interesting extension of the research presented here would be to also annotate features such as "quality of the communication channel", or "frequency of clarified word", which could further aid interpretation.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

References

Jens Allwood. 1995. An activity based approach to pragmatics. Gothenburg Papers in Theoretical Linguistics 76, Göteborg University, Göteborg, Sweden.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

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

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, Stanford, USA.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *Proceedings of the 39th Meeting of the ACL*, Tolouse, France.

Christine Gunlogson. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Ph.D. thesis, University of California, Santa Cruz, CaliforniaUSA, December.

Staffan Larsson. 2003. Interactive Communication Management in an Issue-based Dialogue System. In Ivana Kruijff-Korvayová and Claudia Kosny, editors, *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue. Diabruck*, pages 75 – 82, Saarbrücken, Germany, September. Universität des Saarlandes.

Christoph Müller and Michael Strube. 2001. MMAX: A Tool for the Annotation of Multi-modal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 45–50, Seattle, USA, August.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the Means for Clarification in Dialogue. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue (SIGdial01)*, pages 116–125, Aalborg, Denmark, September. Association for Computational Linguistics.

Matthew Purver, Patrick G.T. Healey, James King, Jonathan Ginzburg, and Greg J. Mills. 2003. Answering clarification questions. In Alexander Rudnicky, editor, *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue(SIGdial03)*, Sapporo, Japan, July.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th Workshop of the ACL SIG on Discourse and Dialogue (SIGdial04)*, Boston, USA, April.

SFB-360. 2000. Bielefeld Corpus. URL: http://www.sfb360.uni-bielefeld.de.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, October.

Case-based Natural Language Dialogue System using Facial Expressions

Satoko SHIGA
Fujitsu Laboratories LTD.
1-1, Kamikodanaka 4-chome,
Nakahara-ku, Kawasaki,
211-8588, Japan
shiga.satoko@jp.fujitsu.com

Seishi OKAMOTO
Fujitsu Laboratories LTD.
1-1, Kamikodanaka 4-chome,
Nakahara-ku, Kawasaki,
211-8588, Japan
seishi@jp.fujitsu.com

Abstract

A new method for case-based natural language dialogue system is presented. This system deals with not only the utterance sentences that are used in usual case-based dialogue systems, but also facial expression information to express past cases. As a result, it can improve the appropriateness of response and present the system's utterance along with facial expression information to the user. We show the advantage of our system over other systems by using examples of dialogue provided by our system.

1 Introduction

A natural language dialogue is one of the best ways for creating a man-machine interface. Although many approaches for dialogue systems have been proposed, including a template-based approach (Weizenbaum, 1966) and a plan-based approach (Allen et al., 1994; Carberry, 1990), in this paper, we apply a case-based approach.

Case-based reasoning (CBR) is a reasoning model that solves a new problem by using previous observations. Past cases, which consist of pairs of problems and their solutions, are stored in a case-base. The system recalls a similar case to the new problem, and then the solution of the selected case is modified to adjust to any difference between the new problem and the past problem. Finally, the system puts forward the modified solution as the solution to the new problem. CBR has the following advantages over other approaches (e.g.,(Leake, 1996)):

- The cost of knowledge acquisition is low, because the system only has to record facts that actually happen as cases.
- Knowledge maintenance is easy because the system learns incrementally. The cases are added automatically, and it is unnecessary to take into account the consistency of knowledge.
- Quality of solutions is increased even though the domain is ill-defined, because the system can treat phenomena that are difficult to formalize.
- Problem-solving efficiency is increased because the system gets shortcut to the successful solution by reusing the case.

In applying the CBR model to dialogue systems, a past dialogue history is stored as a case in a case-base. To generate a response, the system retrieves a similar utterance to the current context from the case-base, and modifies the response utterance of the case to suit to the current situation.

In making a dialogue system, the advantages of CBR are important for the following reasons:

- A large quantity of complicated templates or planning rules must be used in the template-based or plan-based dialogue system. It is, however, quite difficult to make an enough quantity manually. The case-based approach reduces the cost of the knowledge acquisition

and makes possible the system construction easily.

- Knowledge maintenance is a thorny issue in other approach. To develop the system's vocabulary, for example, it is often necessary to revise the whole rule (because adding one rule often means rewriting a large part of the rules). In case-based approach, we just have to add cases of utterances including the new word.

- There are various ways to respond to one utterance, and it is difficult to formalize them as rules. The case-based approach is suitable for such domain to provide the high-quality solution.

- The template-based or plan-based systems can not deal with unexpected dialogues. In contrast, the case-based system has robustness because they can always respond by modifying a similar case.

Several dialogue systems have been developed under the case-based approach. Murao et al. proposed a spoken dialogue system to provide shopping information to a person driving a car (Murao et al., 2003). To generate the response to an utterance, this system uses hand-annotated dialogue cases collected by the Wizard of OZ (WOZ) (Fraser and Gilbert, 1991) system in advance. Okamoto et al. proposed a dialogue agent for web guidance (Okamoto et al., 2001). This system is based on the WOZ method, but it is combined with case-based method for automatic response generation to reduce gradually the burden on the operator (wizard). The wizard checks each generated response and corrects it only when it is inappropriate. However, these systems cause the problem that manual operation is required. This means the advantage of the case-based approach (namely, low construction cost) is lost.

On the other hand, as a case-based system without hand control, a general-purpose chat system was proposed by Inui et al. (Inui et al., 2001; Inui et al., 2003). This system uses dialogue cases that are collected through all interactions with users and annotated automatically. The case is defined as a sequence of utterances and its response. However, their system involves the problems described below.

The first problem is that the similarity measure only depends on the information obtained from surface sentences of an utterance. As a result, the system can not distinguish two utterances that are the same sentence but have different intention. The meaning of an utterance changes according to how the word is expressed. For example, the response to the utterance "Pardon?" in a normal, puzzled, or angry manner should be just repeating the sentence, by saying it again with paraphrase, or by saying something different. In this way, natural human communication uses various modes of information. According to the published findings from psychological research (Mehrabian, 1972), only 7 percent of information is communicated verbally (through words), while the remaining 93 percent is communicated nonverbally (38 percent through the use of the voice, and 55 percent through facial expressions, body posture, gestures etc.). We believe nonverbal information is therefore necessary for dialogue systems to interpret the user's utterances more correctly.

The second problem is that the system's self-learning is only addition of the cases. For example, when the system tries to respond to "What's your name?", the following two past dialogues are put forward as similar cases:

Case 1:

A: "What's your name?"
B: "Today is my birthday."

Case 2:

A: "What's your name?"
B: "My name is Mary."

Although Case 1 is a system's inappropriate automatic response, Inui et al.'s system chooses it at a probability of 1/2. Moreover, if the inappropriate case is selected, another failed case (current generated dialogue) is added to the case-base, and increases the probability of a miss selection to 2/3 in the next selection. This is caused by the lack of learning mechanism of distinction between successful and failed cases.

The third problem is that case selection depends on only the similarity with current context, and the system does not care for the following turn. For example, there is a following two similar cases to the user's utterance "I lost my dear necklace":

Case 3:

A: "I lost my dear necklace."
B: "You're so careless."
A: "... Terrible!"

Case 4:

A: "I lost my dear necklace."
B: "That's too bad. Cheer up."
A: "Thank you."

Both Case 3 and Case 4 have the same utterance to user's input, and Inui et al.'s system chooses Case 3 at a probability of 1/2. However, the system's response in Case 3 angers the user, in comparison with comfort in Case 4. As shown in this example, it is important for case selection to consider the following user's reaction.

In light of the above-described problems, we propose a new method for a case-based natural language dialogue system. Although our system is based on the system proposed by Inui et al., it provides one solution to the problems described above by using a user's facial expressions that accompany their utterances. Our system uses the facial expressions for the following purposes:

- To improve the accuracy of similar case retrieval.
- To evaluate the appropriateness of similar cases for optimal case selection.
- To enhance the system's utterances to the user.

2 Case-based Dialogue System using Facial Expressions

The outline of our system is shown in Figure 1. In this system, a user and the system give utterances alternately, and one utterance consists of several sentences and one facial expression. When a user inputs one utterance, at first, the system extracts

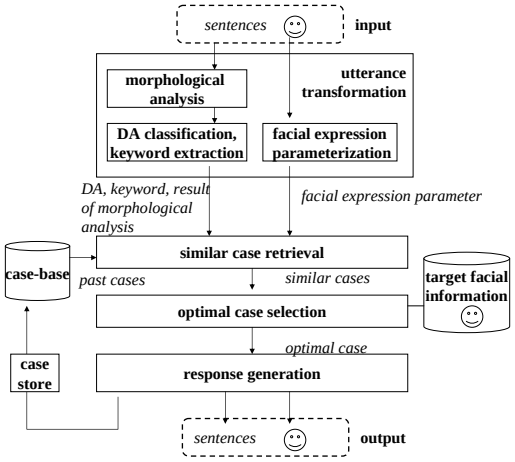


Figure 1: System overview

the linguistic and facial information. Secondly, the system considers the current context as a new problem, and selects similar cases by using linguistic and facial similarity measures. In the next step, the appropriateness of each similar case is evaluated by using facial information of user's reaction, and the optimal case is selected. Then, the selected case is adapted to the current context to generate the response to the user. Finally, the current user's input utterance and system's output utterance are added to the case-base.

2.1 Case Expressions

The case-base contains the time-series utterance history. The form of one utterance is as follows:

ID Number

- Utterance number

Sentences(For each sentence:)

- String
- Result of morphological analysis
- Dialogue act (DA)
- Keywords (a noun,a verb,an adjective)

Facial expression

- Parameters to represent a facial expression

Utterance number is a sequential serial number of the utterance, and a DA is a type of sentence indicating user's intention. Keywords are meaningful words indicating the topic of an utterance.

2.2 Utterance Transformation

When the user inputs one utterance, the utterance transformation module transforms it to the same form as with case expression.

First, the input sentences are divided into individual sentences. A morphological analyzer (Inui and Kotani, 1999) is used to analyze them into a series of words and parts of speech, and passes the results to the DA classifier (Inui et al., 2001) and keywords extractor (Inui et al., 2001). The DA classifier, trained from a DA-tagged corpus, determines a DA for each sentence. There are 17 types of DA, as listed in Table 1.

Table 1: Dialogue acts

greet	*request_comment*	*reject*
bye	*request (Y/N)*	*deliberate*
opinion	*confirm*	*apologize*
will	*request*	*surprise*
explain_fact	*suggest*	*thank*
give_reason	*accept*	

Meanwhile, keyword extractor computes the weight of each word with heuristic rules which focus on "parts of speech", "kinds of characters" (*kanji*, *katakana*, *hiragana* in Japanese), "position in the sentence" (as a substitute for syntactic analysis), and so on. Then, a triplet of a noun, a verb, and an adjective is extracted as the keywords from each sentence.

The facial expression is represented by 18 parameters. There are 15 characteristic points on the eyebrows, eyes and mouth of the face, and the value of each parameter is given as the distance between two different characteristic points. The parameters and the characteristic points are shown in Figure 2. The nose has no characteristic points, since change of the facial expression hardly ever appears in the nose. An example result of an utterance transformation is shown in Figure 3.

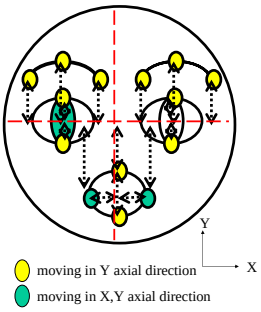
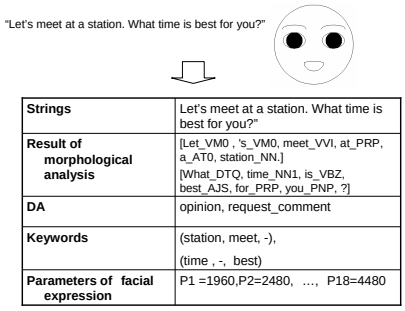


Figure 2: Facial characteristic points and parameters



Strings	Let's meet at a station. What time is best for you?"
Result of morphological analysis	[Let_VM0 , 's_VM0, meet_VVI, at_PRP, a_AT0, station_NN.] [What_DTQ, time_NN1, is_VBZ, best_AJS, for_PRP, you_PNP, ?]
DA	opinion, request_comment
Keywords	(station, meet, -), (time , -, best)
Parameters of facial expression	P1 =1960,P2=2480, …, P18=4480

Figure 3: Example result of utterance transformation (original is in Japanese)

2.3 Similar Case Retrieval

The similar case retrieval module considers the sequence of the past M utterances during the current dialogue as the current context, and selects similar cases in contrast to the sequence of the utterances in the case-base. In this paper, we set M to two.

Throughout this paper, we represent the current context as $\mathbb{P} = \langle p_1, p_2 \rangle$, where p_1 is the last system's output and p_2 is the current user's input. On the other hand, a case is a sequence of time-series J utterances in the case-base, and it is expressed as $\mathbb{C}_i = \langle c_i, c_{i+1}, c_{i+2}, ..., c_{i+J-1} \rangle$, where c_i is an utterance with the utterance number i in the case-base.

For calculating similarity between current context and each case in the case-base, we use the following three methods:

1. DA-based matching
2. Keyword matching
3. Calculation of facial expression similarity

The techniques that Inui et al. developed for DA-based matching and keyword matching are used in this module. DA-based matching is used for the case filtering based on a type of sentences. Keyword matching is the cost calculation based on the number of matched terms. Refer to (Inui et al., 2001) for further information. In this paper, the similarity between two sets of keywords, s and t, is expressed as $Sim_{\text{key}}(s,t)$.

The similarity between two facial expressions, x and y, is calculated from Formula (1) using 18 parameters of distance between the characteristic points.

$$Sim_{\text{face}}(x,y) = \sum_{i=1}^{18} W_i \frac{\{x[i]-y[i]\}^2}{\{\text{MAX}_i\}^2}, \quad (1)$$

where
W_i:weight of i-th parameter,
MAX_i:maximum value of i-th parameter,
$x[i]$ and $y[i]$:value of i-th parameters of facial expression x and y respectively

Then, the similarity between two utterances, u and v, is calculated by using both similarity for keywords set and similarity for facial expression:

$$Sim(u,v) = \alpha Sim_{\text{key}}(key[u], key[v]) + \beta Sim_{\text{face}}(face[u], face[v]), \quad (2)$$

where
$key[u]$:a set of keywords of utterance u,
$face[u]$:facial expression of utterance u,
α, β:constant values

The similar case retrieval module calculates the total similarity $Sim_{\text{ret}}(\mathbb{P}, \mathbb{C}_i)$ by using Formula (3), and retrieves K most similar cases to the current context $\mathbb{P}$.

$$Sim_{\text{ret}}(\mathbb{P}, \mathbb{C}_i) = Sim(p_1, c_i) + Sim(p_2, c_{i+1}) \quad (3)$$

2.4 Optimal Case Selection

After similar cases have been retrieved from the case-base, the optimal case selection module selects an optimal case from them and uses it to generate the response to the user.

As mentioned in Section 1, it is important that a case-based dialogue system guesses the following dialogue and selects the case by measuring the appropriateness of the look-ahead section of each case in order to generate the appropriate response.

To measure the appropriateness, the user's feedback information about the quality of a system's response is useful. As feedback information, our system utilizes the facial expression of the N utterances uttered right after the system's response. The number of N utterances is fixed at one, that is, the system only uses the user's utterance uttered right after the system response. The case is, therefore, expressed as a quadruplet of the utterances.

The calculation of the appropriateness is explained informally as follows. As shown in Figure 4, after similar cases are obtained, utterance c_3 immediately following utterances c_1 and c_2, which are similar to the current context, is the candidate for the system's response. The facial expression of a utterance c_4 right after the candidate utterance c_3 is used as user feedback information, and the system compares it with the target facial expression. The appropriateness of case $\mathbb{C}_i = \langle c_i, c_{i+1}, c_{i+2}, c_{i+3} \rangle$ for the target facial expression q is formally represented as the following Formula(4), by using the similarity for facial expression between c_{i+3} and q.

$$App(\mathbb{C}_i, q) = \begin{cases} Sim_{\text{face}}(face[c_{i+3}], q) & \text{(if } q \text{ is desirable)} \\ 1 - Sim_{\text{face}}(face[c_{i+3}], q) & \text{(if } q \text{ is undesirable)} \end{cases} \quad (4)$$

As the target facial expression, either a desirable or an undesirable facial expression can be set. If we set a desirable target, the similar case has priority for selection; and if it is an undesirable target, the priority of similar case is low. We adopted a strategy of using a "smiling face" for a desirable target and an "angry face" for an undesirable target. However, desirable facial expressions will vary according to various factors, such as the domain of the system and the duration of a dialogue. We therefore presume that users can dynamically specify the target facial expression, and that more

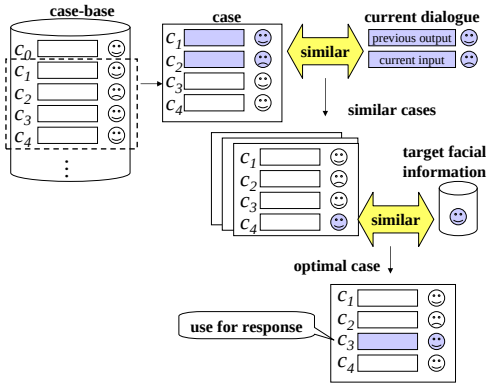


Figure 4: Optimal case selection

than one facial expression can be set and switched dynamically according to policy settings.

The optimal case selection module considers two factors comprehensively to choose one optimal case; one is the similarity to the current context, and another is the appropriateness of cases using the similarity to the target facial expression. The optimality between current context $\mathbb{P}$ and the case $\mathbb{C}_i$ is calculated as shown in Formula (5). Then, the case $\mathbb{C}_i$ which has a minimum score of $Opt(\mathbb{P}, \mathbb{C}_i)$ is chosen as the optimal case.

$$Opt(\mathbb{P}, \mathbb{C}_i) = \gamma Sim_{\text{ret}}(\mathbb{P}, \mathbb{C}_i) + \delta App(\mathbb{C}_i, q), \quad (5)$$

where

γ, δ:constant values

2.5 Response Generation

After an optimal case is selected, the system uses third utterance in the quadruplet expression of the optimal case as a template for the response utterance. The adaptation of the template to the current context is done as follows. The output sentences are generated by replacing each keyword of the template with the corresponding keyword in the current context. To replace the keyword, we use the Inui et al.'s technique (Inui et al., 2001), which uses the keyword correspondence table made in keyword matching process, is applied. On the other hand, the facial expression of the optimal case can be directly used as the system's output.

2.6 Case Store

The case store module stores the pair of the user's input utterance and the system's output utterance in the case-base. As the dialogue is repeated, the input and output utterances are accumulated in the case-base in chronological order.

3 Empirical Evaluation

We made a prototype of the system for testing. Compared with Inui et al.'s system (Inui et al., 2001), the appropriateness of responses in our system was confirmed to be better. Some examples of actual dialogues that represent the advantage of our system over Inui et al.'s system are given in the following.

Dialogue example 1 (see Figure 5) shows the advantage of using facial expression information for similar case retrieval. Two input dialogues containing the same sentences but different facial information are considered. Although Inui et al.'s system generated the same response for these inputs, our system generates more appropriate responses according to the input facial information.

On the other hand, dialogue example 2 (see Figure 6) shows the advantage of optimal case selection. Case 1 and Case 2 are selected as similar cases to the current context, since utterances U1 and U2 are similar to those of the current context. Inui et al.'s system chooses Case1 as a similar case, although the user is angry in Case 1 because the system's response U3 is inappropriate. However, in our system, the facial expression shown in Figure 7 was set as an undesirable target in this experiment. The appropriateness of Case 1 is much lower than that of Case 2, because the similarity between the facial expression of U4 and the undesirable target facial expression is much higher. Therefore, overall, the system uses Case 2 to generate the responses shown in Figure 6.

4 Conclusion

A new method for case-based natural language dialogue system was developed. To generate an appropriate response, this system obtains the user's facial expressions and uses them to retrieve similar cases to the current context. Moreover, the system uses the user's facial information to evaluate the

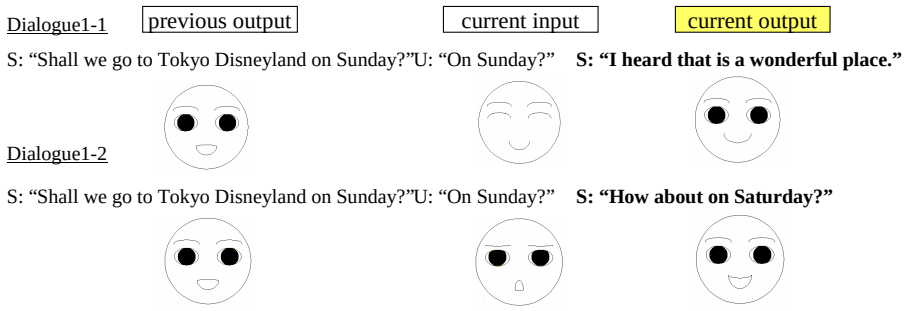


Figure 5: Dialogue example 1 (original dialogue is in Japanese)

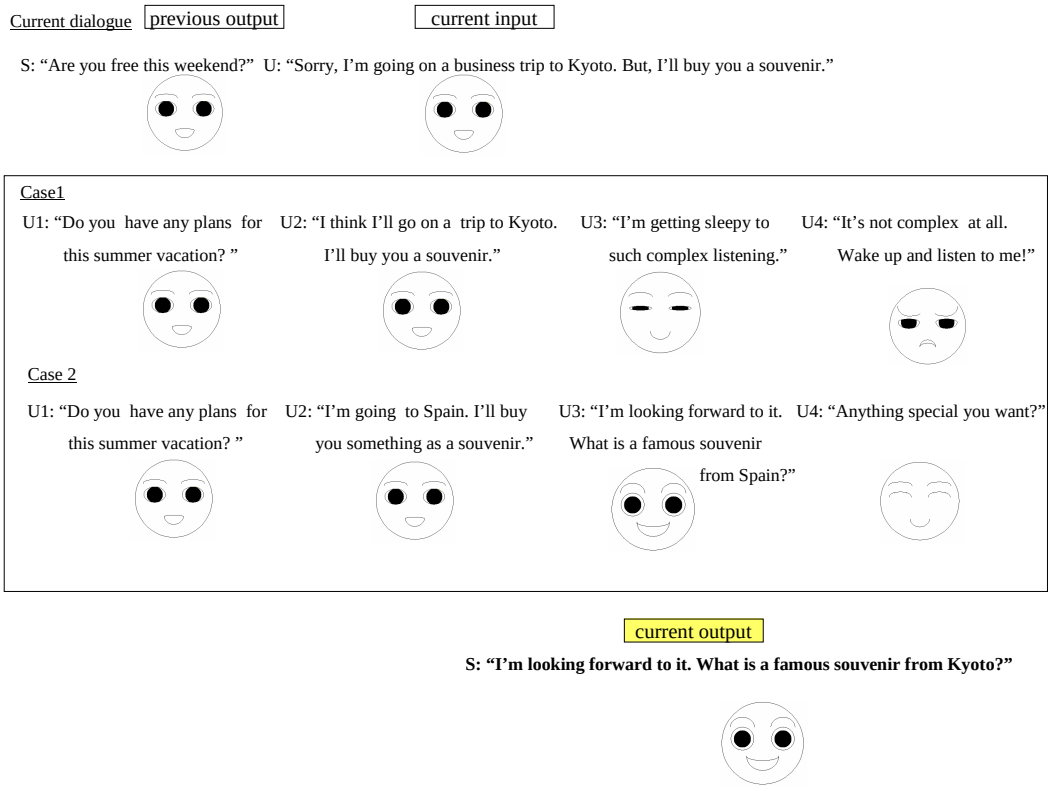


Figure 6: Dialogue example 2 (original dialogue is in Japanese)

Figure 7: Undesirable target facial expression

appropriateness of each case and to choose the optimal case. We plan to provide a more detail evaluation of our current system. After much experimentation, we would like to show the advantage of our system over other systems. We also plan to adopt an automated recognition technique of facial expression (Mase, 1991) to reduce the user's task because our current prototype system requires the user to input the facial expression manually.

References

J. Allen, L. Schubert, and et al. 1994. The TRAINS project: A case study in building a conversational planning agent. TRAINS Technical Note 94-3, Univ. of Rochester.

S. Carberry. 1990. *Plan Recognition in Natural Language Dialogue*. The MIT Press, Cambridge MA.

N. Fraser and G. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5(1):81–99.

N. Inui and Y. Kotani. 1999. Finding the best state for HMM morphological analyzer. *NLPRS'99*, pages 44–49.

N. Inui, T. Ebe, B. Indurkhya, and Y. Kotani. 2001. A case-based natural language dialogue system using dialogue act. *IEEE International Conference on Systems, Man, and Cybernetics*, pages 193–198.

N. Inui, T. Koiso, J. Nakamura, and Y. Kotani. 2003. Fully corpus-based natural language dialogue system. *2003 AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pages 58–64.

D. B. Leake. 1996. CBR in context:the present and future. In *Case-Based Reasoning - Experiences,Lessons, & Future Directions*, chapter 1, pages 3–30. AAAI Press & The MIT Press, Menlo Park California & Cambridge MA & London England.

K. Mase. 1991. Recognition of facial expressions for optical flow. *IEICE Trans. on Information Systems*, E-74(10):44–49.

A. Mehrabian. 1972. *Nonverbal Communiation*. Aldine-Atherton, Chicago.

H. Murao, N. Kawaguchi, and et al. 2003. Example-based spoken dialogue system using WOZ system log. *4th ACL SIGDIAL Workshop on Discourse and Dialogue(SIGDIAL-2003)*, pages 140–148.

M. Okamoto, Y. Yang, and T. Ishida. 2001. Wizard of Oz method for learning dialog agents. *International Workshop on Cooperative Information Agents (CIA-2001)*, pages 20–25.

J. Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communications between men and machines. *CACM*, 9(1):3–45.

Information-Seeking Chat: Dialogue Management by Topic Structure

Manfred Stede and **David Schlangen**
University of Potsdam
Department of Linguistics
Applied Computational Linguistics
P.O. Box 601553
D-14415 Potsdam — Germany
{stede|das}@ling.uni-potsdam.de

Abstract

In this paper we describe the dialogue sub-genre "information-seeking chat", which is distinguished from other kinds of information-seeking dialogue (e.g. travel information) by its more exploratory and less (single) task-oriented nature. We present an approach to modelling this kind of dialogue, based on the notion of *weighted topic structures* — a single data structure that represents both the domain knowledge and the dialogue history, and we sketch an implementation of this approach in a typed dialogue system.

1 Introduction

Both theoretical analyses of dialogue and implemented dialogue systems have so far mostly focused on two main dialogue genres: strictly task-oriented dialogue (as in the travel agent domain, call routing applications, or collaborative problem solving domains), or tutorial dialogue. In this paper we describe another type of dialogue, which we call "information-seeking chat". This genre is distinguished by its more exploratory and less task-oriented nature, while still being more structured than general free conversation.

Our thesis is that this kind of dialogue can be modelled with a simple taxonomy of dialogue moves and a dialogue management (DM) strategy based on *topic structure*, where the main task of the dialogue manager is to guide the user through the pre-defined topic map. This topic map is a declarative domain model (similar to an ontology) that serves both as a representation of the domain knowledge and as a repository for the discourse history. (The model represents the discourse history insofar as during the course of the dialogue it is annotated with information about which topics have been broached or have been exhausted.) Moreover, it is the only discourse *planning* device the system uses, since it also records the effect of each utterance on the decision of which bit of information to relay, which topic to explore next. This surprisingly simple information structure can successfully model this important kind of dialogue, as we argue here, and it also makes it relatively easy to implement new applications covering other domains in this style—information about companies, for example, or more generally about structured fields of knowledge.

The remainder of this paper is organised as follows. Section 2 elaborates on the peculiarities of "information-seeking chats", and presents our taxonomy of dialogue acts. A discussion of the problems extant approaches to dialogue modelling would have with this kind of dialogue leads over to Section 3, where we describe our approach, and the prototypical implementation. After discussing in Section 4 related attempts to reduce dialogue management to representations of task-knowledge, Section 5 sketches how our dialogue manager fits in with the other modules of the system that is under development. Section 6 finally discusses evaluation issues, and further work.

2 The Dialogue Genre

2.1 Information-Seeking Chat

Imagine that you are planning next year's vacation, and that you want to find a destination that offers both cultural attractions and leisure activities. Once a candidate emerges, you now ideally want to have a conversation with a local representative, ask her a few specific questions about activities you have in mind, but also be alerted by her towards attractions you did not think about. This conversation will wander from one aspect to another, sometimes come back to a topic already mentioned, and possibly sometimes digress to the weather, or local football teams. The purpose in any event is for you to form an impression of an area that previously you knew only very little about, guided by your interests and by what you learn.

This kind of dialogue is rather different from dialogues aiming at finding *one particular* piece of information or executing one particular transaction. While it does fall under the general rubric "Inquiry Oriented Dialogue" defined by (Larsson, 2002) as shown below in (1), it is distinguished from those kinds of inquiry activities by not being driven by specific goals that can easily be decomposed into hierarchically ordered subgoals (like "find a specific train connection", which can be decomposed into "get destination, start date, etc.."). Rather, it is, at least at the beginning of the dialogue, only driven by relatively unspecific high-level goals (e.g., "tell me something about city XY.") that are not as easily decomposed.[1]

(1) "[. . .] the term Inquiry Oriented Dialogue, or IOD, will henceforth be taken to refer to any dialogue whose sole purpose is the transference of information, and which does not involve any DP assuming (or trying to make another DP assume) commitments or obligations concerning any non-communicative actions outside the dialogue." (Larsson, 2002, p. 17)

Hence, this kind of dialogue does not naturally lend itself to an approach of goal-directed hierarchical planning. Instead, it shows similarities to "smalltalk" that drifts from one aspect of the topic to another, while still being more constrained than that by having a specific, albeit very general, purpose. It is also much more a "mixed-initiative dialogue" than for example a dialogue in the well-known travel domain (as modelled in the GoDiS system (Traum and Larsson, 2003) and the DARPA Communicator systems), since both interlocutors can quite freely open up new subtopics, declare one as closed, digress, or hesitate.[2] Accordingly, this kind of dialogue requires implementation strategies quite different from those established for task- or transaction-oriented dialogue. We will propose one such strategy in Section 3, but first we briefly summarise results of a corpus study, which lends further motivation to this choice of strategy.

2.2 Dialogue Flow and Dialogue Acts

We conducted a small corpus study, collecting text dialogues (using a web-based chat tool) between a domain expert and an information-seeker. The only instruction we gave to the expert was to open the dialogue with a standard opening ("Welcome. This is Potsdam Tourist Information. How may I help you?"); the scenario for the inquirer was that they will be in Potsdam for a conference and have to decide whether they should take the weekend after the conference off and stay. All inquirers were non-locals and not familiar with Potsdam. We collected 13 dialogues, of around 17 turns each.[3] An excerpt of one of the dialogues is shown in Figure 1 (annotated with the dialogue-act-types that will be described presently).[4]

[1] Note that it is not excluded that such a dialogue might develop into a more focused, traditional task-oriented dialogue. For example, during an information-seeking chat, the inquirer might become interested in a particular offering and want to book a ticket. In our final system we plan to build in interfaces that can hand over control to other dialogue managers designed for this kind of dialogue.

[2] However, there *are* differences between the dialogue participants, and the boundaries of the roles of inquirer and expert are clear. For example, in our corpus (see next subsection) there is a tendency for the expert adviser not to cut off a topic that is being talked about, and so topic-changes mostly initiate with the user. In principle, however, both interlocutors seem to have the same range of moves available at any point.

[3] Which makes them rather short; this is possibly due to the "simulation"-nature, but hopefully does not affect the representativity of the dialogues.

[4] The collected dialogues are in German, but for reasons of space we only show here an English translation. Since the focus of the paper is on dialogue-management (and not for example on dialogue act recognition), the observations should largely be language-independent.

dialogue 008			
1	E:	Hello. This is Potsdam Tourist Information. How may I help you?	opening
2	U:	Well, I'd like to know whether a weekend in Potsdam is worthwhile?	switch-topic
3	E:	A weekend in Potsdam is always worth your while, there are lots of nice things you can do here.	tell-topic-general
4	U:	Like what?	ask-more-specific
5	E:	For example, you can make a boat trip, or a sightseeing tour. Always popular with tourists is Park Sanssouci.	tell-more-specific
6	U:	Is all of this in the inner city? Can you walk to these things, or do you have to use some kind of transport?	ask-more-attribute
7	E:	It's probably best to take the bus to the park, since it's not really in the centre. By foot it's about 20 minutes. The busses leave from the central station. This is also where the sightseeing tours start. The harbour is also only 5 minutes away from the station.	tell-spec-attribute
8	U:	Sounds good. Are there any reasonably priced hotels near the station?	switch-topic
9	E:	There's the XY-hotel close to the station. But in the centre there are also many other places where you can stay. [...]	tell-topic-general
12	U:	Are there things to do for kids as well?	switch-topic
13	E:	A popular attraction for kids is the filmpark in Babelsberg. You can get there with the tram.	tell-topic-general
14	U:	How far away is that? Which tramline?	ask-more-attribute
15	E:	It's only two stops from the station. You take the S1 to Griebnitzsee, and then you walk, it's only 10 minutes. But you can also take the bus.	tell-spec-attribute
16	U:	Only two stops, great. Is the entrance fee very high?	ask-more-attribute
17	E:	Entrance is 17 Euro for adults and 10 Euro for children.	tell-spec-attribute
18	U:	Not exactly cheap. What's on offer there?	ask-more-specific
19	E:	You can visit the Ufa-filmstudio and take part in the shows. Apart from that, it's like a theme park, with roller-coasters and stuff.	tell-spec-attribute
20	U:	Sounds good! I think I'll stay in Potsdam for the weekend then. Thank you very much for the information.	bye
21	E:	My pleasure. Have fun in Potsdam.	closing

Figure 1: An example dialogue, with dialogue-act annotation

We made the following observations:

• Firstly, the dialogues mostly seem to follow a recursive pattern of dialogue moves: the user asks for (further) information about a topic, which the expert gives, thereby proposing alternative ways of further exploring that topic. Then either one of these alternatives is taken up, and a new sequence is started beginning with this sub-topic, or else the user jumps to a different topic and begins a sequence there. This pattern is illustrated by turns 4 to 8 in Figure 1: turn 5 offers several alternative answers to the question in turn 4, and the user replies by inquiring more details. Then, in turn 8, the user ends this subsequence and jumps to a new sub-topic, which is then briefly explored. Turns 12 to 19 give an example of a sub-topic that is explored in more detail.

• Connected to this pattern is the observation that most questions (and, since this is the preferred device for changing the topic, most topic shifts) are initiated by the user: 89% of all questions come from the user, with a proportion of 35% of all turns being questions.

These are the requirements for modelling this kind of dialogue, then: a) the dialogue manager must allow for systematic exploration of a topic, while b) allowing at all times user-initiated topic shifts. It should be clear that finite-state based approaches (see e.g. (McTear, 1998)) are too rigid for these requirements; they could only model this amount of user-initiative if every state (representing a topic) not only had transitions leading to all alternative sub-topics, but also to *all* other topics as well—resulting in a number of states that is hardly practical.

The flexibility afforded by information-state-update (ISU) approaches (Traum and Larsson, 2003), on the other hand, seems better suited.[5] In

[5]Note that we are talking here about the power of the general approach of ISU-based dialogue management. The extant systems following this approach, such as GoDiS for example (Traum and Larsson, 2003), put some additional constraints on the dialogue management, by being relatively closely oriented towards template-filling and relying on system-initiative to do this, and hence are not capable of han-

this approach, dialogue is modelled as a sequence of *updates* of an (arbitrarily complex) *information state* recording discourse history as well as beliefs and plans, governed by *update-rules* that are triggered by *dialogue acts* and that produce such acts, which abstract over different linguistic realisations. Indeed, the notion of *proposals* as used above fits in nicely with what (Larsson, 2002) calls *issues under negotiation*, which are part of his Information State. In this approach, issues are represented as questions, and proposals are alternative answers to these questions. This mechanism, however, does not say anything about where the required semantic relations between questions (whether two questions are independent—what we would call a topic shift—or whether one further specifies another) is stored.

Our claim is that an underlying hierarchical organisation of topics (much like an ontology modelling the domain) is needed in any case, and that, combined with *weights* representing discourse history, this is indeed the *only* structure that is needed, obviating the need for storing explicit plans. This idea will be explored in the next section, but first we give the full list of dialogue-acts that we derived from our corpus and use in our system (Figure 2). Together with the examples given, the classes should be self-explanatory.[6] Note that we do not claim any general use for this set of acts besides describing this specific dialogue-genre, and that we have devised this set with our approach to dialogue-modelling in mind. Nevertheless, we have tested the coverage and reliability of this mark-up scheme, by getting two naive annotators to code up our dialogues. The achieved coverage of around 98% of all utterances (i.e., only 2% were marked as `other`) and the resulting κ value of .81 indicates the usefulness of the schema.

dling this kind of dialogue. That the *general* approach of ISU should be flexible enough to model it, is perhaps not surprising: it is meant to be a general framework for implementing and comparing dialogue management strategies, after all (Traum and Larsson, 2003).

[6]The acts help and garbage are only used in the implementation (they mark requests for producing a system message and recognition failure, respectively) and not for marking-up the dialogue examples.

3 The Wanderer: Dialogue Management with Topic Structures

3.1 Overview of the approach

The approach to dialogue management proposed here inherits traits from very different traditions: chatbots[7] and ISU-approaches. From the former we import the robustness and the locality of pattern-matching-based dialogue management, while the latter give us a model for abstracting from specific inputs by using dialogue-acts. In the system, we explore the chatbot-like strategy of letting local control decisions drive the dialogue forward—local decisions which, however, need access to an over-arching discourse model. This discourse model in our approach is rather different from earlier notions in that it is very closely related to the *content* model of the system. More specifically, we use a declarative, ontology-like model of domain knowledge as the central repository of information in the system; a repository that holds not only the conceptual knowledge and the associated linguistic forms, but also the information about what has been talked about already and what can or should still be put onto the agenda, represented as numerical *weights* on the topic nodes. Consequently, it is *the content* (together with, or rather, also representing the dialogue history) that is in charge of controlling dialogue flow—as opposed to other dialogue genres, where intentions and goals are in the driver's seat.

The 'information state' used in our system thus is quite simple: it does not contain *explicit* goals or beliefs or partitions; instead it contains an instantiated domain model, dynamically enriched with numerical information representing preferences for discussion.[8] As described in the previous section, we do use dialogue moves as an abstract layer between possible inputs and possi-

[7]Web-based systems for typed dialogue, using just pattern matching, but with quite sophisticated implementations. See, e.g., `http://www.alicebot.org`.

[8]We stress "explicit" here, because the weights can be seen as implicitly storing past intentions, and guiding the overall intention of "staying on topic" and "exploring the topic", and there is the implicit plan for doing the latter, namely by offering information about children nodes. Moreover, the dialogue acts of course represent communicative intentions; the point is that no explicit *planning* beyond the local decision on how to react to the last utterance is needed.

Dialogue Act	Example
ask-more	
ask-more-general	*"Can you tell me more about Potsdam's parks?"*
ask-more-specific	*"Is there a park in Potsdam?"*
ask-more-attribute	*"What are the opening times of the park?"*
reply	
reply-pos	*"OK."*
reply-neg	*"Not really, thanks."*
tell	
tell-topic-general	*"Potsdam has four large public parks."*
tell-spec-attribute	*" 'Sans, Souci.' was built in 1745."*
rule-out-topic	*"I am not that interested in parks."*
switch-topic	*"OK. What about museums in Potsdam? What's on offer there?"*
noncommittal	*"Oh well, I don't know."*
digression	*"Parks are good. I really like a good barbecue in the park."*
bye	*"Thanks. That's enough."*
opening / closing	*"Welcome. This is ..." / "Thank you and goodbye."*
help	*"Help!"*
garbage	

Figure 2: Dialogue Acts used to Describe Information-Seeking Chats

ble updates and also between update effects and possible outputs; our notion of 'update', however, is one of adjusting weights, which reflect the dialogue history as well as user's statements regarding her dis-/interest in particular branches of the domain model.

3.2 The Implementation

We realised The Wanderer using a *description logic* (DL, see e.g. (Baader and Nutt, 2003)), which has several useful properties: Being a 'structured fragment' of first-order logic, it organises knowledge in taxonomies and offers inference mechanisms dedicated to and optimised for taxonomical reasoning—in particular, *subsumption checking*. The knowledge base is split into a terminological part (concepts and relations between them) and an assertional part (specific instances of the concepts/relations in the terminology)—much like the class/instance distinction in object-oriented programming. Subsumption is computed among concept descriptions, and between instance descriptions and concepts. Hence, DL-systems offer sophisticated instance-retrieval: given an arbitrarily complex concept description, the DL finds the set of instances satisfying the description.

In our approach, the terminology part has two components: a model of the domain, and a model of linguistic utterance types. More specifically, the domain model serves as a taxonomy of *topics* that can be addressed in a conversation. In our example application, the root concept `city-topic` is partitioned into entities such as `people`, `buildings`, `parks`, `lakes` and the like, which in turn are decomposed into more specific categories; e.g. `buildings` can be `palaces`, `temples`, `museums` and so forth. Under `people` we assemble personalities of historic interest (the former kings, architects, gardeners, etc.) and those of importance for present-day life. Concepts then can be related. E.g., the domain model offers relations that link buildings to architects, to construction years, to the kings who commissioned the buildings, etc. Besides the historic perspective, buildings can also be related to things like street addresses and entrance fees (relevant for museums or movie theatres). Finally, the most prominent relations spanning different sub-domains are subsumed by a generic relation `sim-topic`, which will be explored when the system initiates a topic shift.

Note that the terminology does not include any Potsdam-specific entity—these all belong into the assertional part of the knowledge base. The terminology thus should be transferable to other cities without too much effort. Instantiations of the concepts then constitute the description of the city in question: *Sanssouci*, *Carlottenhof*, *Marmorpalais* and so on are the `palaces` in Potsdam, linked to their architects *Knobelsdorff*, *Schinkel*, etc. Similarly, specific movie theatres are linked to their respective concrete entrance fees—and so forth for

the entire model of Potsdam-related topics.

So far, we have described just static facts, with no relationship to actual linguistic utterances. In our implemented demonstrator (which aims at investigating dialogue strategy rather than linguistic processing), the system's utterances are largely pre-fabricated. Instead of writing them entirely by hand, however, we perform a semi-automatic mapping from the domain model to utterances that verbalise the facts of the model. Thus our terminology also includes a taxonomy of utterance types (essentially templates for different types of information to be transmitted), and a mapping process traverses the domain model and uses the utterance templates to "compile" the set of system utterances. Importantly, these utterances are again *instances* of the DL; their types are both the domain topic and the kind of utterance. Domain topics need not be leaves of the tree: there are for example utterances about `architects` *(The most profligate classicist Potsdam architects were Schinkel, Persius and Hesse)* as well as about the specific architect `schinkel` *(Schinkel was born in 1781 in Neuruppin)*; the former will be produced when jumping to a new general topic and the latter only when it is explored in more depth.

Turning now to the dialogue manager, the key idea is, as indicated earlier, to associate numerical weights with the topics. The weight of an instance characterises its relative salience for the next step of the conversation.[9] When initialising a dialogue session, weights are by default distributed evenly, unless the content designer has already marked some topics as more prominent than others (e.g., one of the palaces gets higher weight so it will be the starting point when the discussion turns to palaces). Similarly, the "kickoff topic" for beginning the conversation (if the system rather than the user sets the first topic) can be set by the content designer by assigning it the highest overall weight.

After initialisation, the system then moves through the following cycle: (1) get user input (simplified, as described above) — (2) classify this input into a dialogue-act and parameters — (3) choose an output utterance — (4) update the weights. The process stops when either the user ends the conversation (system identifies a BYE-act), or the system has nothing more to say, i.e. when the entire topic range has been exhausted.

The DL we use (LOOM, (MacGregor and Bates, 1987)) offers techniques from object-oriented programming, which we use for realising the response strategy of *The Wanderer*: it is a set of independent *methods* that fire when a particular combination of dialogue-act and parameter is identified in the user's input. This corresponds to the idea of local decision-making (see above) but is realised more flexibly than in chatbots, as more computation can be performed (exploiting the DL's services) to determine the optimal output utterance. The first step consists of constructing a concept description (consisting of domain topic and, if applicable, linguistic types) that is handed to Loom, whose query facility will find the set of candidate utterances. Among these, the selection is made on the basis of the weights, i.e., the highest-ranked utterance is chosen (with random choice in case of a tie). Hence it is the weight update mechanism evoked in step (4) that is responsible for steering utterance selection in order to ensure coherence.

The following types of weight-update-functions are implemented; they are used in the dialogue-act-rules as described in Figure 3.[10]

U1 Increase weight of utterance set by n%;
U2 Reduce weight of utterance set by n%;
U3 Increase weight of utterance set by the amount necessary to just outweigh all others;
U4 Reduce weight of utterance set to almost-zero.

An example shall illustrate this mechanism. In the following, we show a (constructed) user utterance, its dialogue-act-type, the effect on the weights, and the system reply:

[9] To keep the process of weight adaptation more transparent, we chose to ensure that the sum of all weights is kept constant (so it can in fact be interpreted as a probability distribution).

[10] Where 'same node / sister' in column 'SYS UTT FROM NODE' means: if there is a non-exhausted utterance left at same node, chose it; otherwise choose one from the highest-ranked sister node. In row 4: the answer to a specific attribute-question has to be looked up explicitly, and hence the weight-based selection doesn't apply. Rows 5,6: currently, the only system question to which the user REPLIES is "Do you want to learn more about TOPIC?" (the "probe question"). In 5-8, 'sim-topic' refers to the `sim-topic`-relation (mentioned earlier) that abstracts over some specific relations.

	USER DIAL.ACT	SYS UTT FROM NODE	WEIGHT UPDATE
1	ask-more	same node / sister	node and daughters: (U1 10)
2	ask-more-general	mother	mother and all her daughters: (U1 5)
3	ask-more-specific	highest-ranked daughter	all daughters: (U1 5)
4	ask-more-attribute A	response: info on A	none
5	reply-pos	same node / sister	node and daughters, sim-topics: (U1 5)
6	reply-neg	sister / any other	node and daughters: U4 sim-topics: (U2 5)
7	rule-out-topic T	highest-ranked node outside T prefer a probe question	node and daughters: U4 all probe questions: U3 sim-topics: (U2 5)
8	switch-topic to T	new topic T don't choose probe question	(old) node and daughters: U4 nodes below T: U3 intro-q's at T: U4 sim-topics-of-T: (U1 5)
9	noncommittal	same node / sister	none
10	digression	mother	all probe questions U3 node and daughters: (U2 5)
11	bye	bye	n/a
12	help	print info	none
13	garbage	same node / sister	none

Figure 3: The weight-update rules

"What about cinemas?"—(switch-topic(cinema))—(increase weight of cinema-topic and daughters, and related entertainment topics)—"*There are four cinemas in Potsdam, the Thalia, Melodie, UCI, and one in the Filmmuseum.*". | *"Where is the Melodie?"*—(ask-more-attribute(location))—(no effect on weights; retrieve attribute)—"It's on Ebert Street.". | *"Well."*—(noncommittal)—(no effect on weights, stay on topic)—"The entrance is only 4 Euros." (...)

It is clear that this mechanism has its limits. It currently cannot handle conversations where two topics are explored in parallel, and it also cannot handle complex queries like "which of the parks is closer to the central station?" (which however are also beyond the scope of most dialogue systems). While not being implemented at the moment, the use of variables (for example for recording the time of the visit) is possible, and sub-dialogues could be launched to fill them.

4 Related Work

We have already related our approach to finite-state-based and ISU-based approaches above; here we focus on the idea of reducing dialogue control to following a representation of domain knowledge. The Swedish WAXHOLM system (Carlson and Hunnicutt, 1996) goes some way in this direction. It makes use of hierarchical topic-maps as well, which however are only used for computing probabilities during topic spotting (see below for our approach to this). The "construct algebra" approach of Abella and Gorin (Abella and Gorin, 1999), used in the *HMIHY*-system, also explores this reduction-strategy. It uses so-called *constructs* that represent knowledge about the tasks that the system can handle (e.g., call forwarding, or giving billing information), which are organised in an inheritance hierarchy. Dialogue management then consists in creating such constructs and applying *dialogue motivators* on them (for filling in missing information, for example), until the constructs are satisfied (and hence the task is done). In our approach, however, it is not *know-how* that is represented but rather topical knowledge.

Finally, the German SmartKom system has recently promoted the use of ontologies in dialogue systems (Gurevych et al., 2003), mostly for coherence scoring. As we will sketch below, we also use our ontology for this, but in addition, as described above, we use it for the dialogue management as well.

5 Sketch of the other modules

Although the focus of this paper has been on the dialogue manager component of our system, we

shall now briefly describe the context in which this module works.

The matching of input to dialogue-acts is performed by pattern-matching combined with keyword spotting (the keywords being recorded on the topic-nodes). E.g., we specify patterns like "Tell me more about *TOPIC*", where *TOPIC* is a placeholder for a keyword. Together with knowledge about the current topic, a match resolves to either one of the acts from the dialogue-act family ask-more, or to a topic-shift. In case more than one keyword matches, the hypotheses are ranked using the weights, thus making double use of this device.[11] This has certain similarities with the chatbot approach of template matching; we plan to upgrade this in the future to either a linguistic analysis or a statistical model or a combination of both.

6 Summary and Further Work

We have presented a brief study of the genre "information-seeking chat", and have suggested that it has certain features distinguishing it from the kinds of information-seeking dialogues (e.g. travel information) predominantly modelled in dialogue systems, the main difference being that it is less driven by specific task-level goals ("ask about intended departure times", for example) than by the topical structure of the domain. We have proposed a taxonomy of speech acts that can describe the moves in such dialogues; and we have sketched a strategy to model such dialogues, together with an experimental implementation of that strategy. Our prototype implementation relies on a strict division between declarative domain model and dialogue management, so that moving the system to a new domain is a matter of replacing the domain model, not one of re-programming.

Besides further developing the modules of the system, we are also planning a more thorough evaluation of this component, through a Wizard of Oz-study: a three-party dialogue where the wizard classifies the user utterances into dialogue-act plus parameters, and the system produces the replies that are sent to the user. A questionnaire is used to assess the user's (dis-)satisfaction, as well as objective measures such as dialogue length, number of misunderstandings, etc. The base line will be given by a defective version of the dialogue manager (that for example makes random topic shifts), and a gold standard by evaluating human performance on the same task.

[11] While our system is based on written input at the moment, it should be possible to make the move to spoken input, by using these patterns to compile out speech recognition (SR) grammars and using the techniques of topic spotting developed in the spoken language community (see *inter alia* (Myers et al., 2000)). Of course, using automated SR would mean that some sort of error-clarification mechanism would have to be integrated, as for example described in (Schlangen, 2004), complicating the dialogue control mechanism.

References

Alicia Abella and Allen L. Gorin. 1999. Construc algebra: Analytical dialog management. In *Proccedings of the 37th Meeting of the ACL*, Maryland, USA. ACL.

Franz Baader and Wolfgang Nutt. 2003. Basic description logics. In Franz Baader et al, editor, *The Description Logic Handbook*. Cambridge UP, Cambridge.

Rolf Carlson and Sheri Hunnicutt. 1996. Generic and domain-specific aspects of the waxholm NLP and dialog modules. In *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Iryna Gurevych, Robert Porzel, Elena Slinko, Norbert Pfleger, Jan Alexandersson, and Stefan Merten. 2003. Less is more: Using a single knowledge representation in dialogue systems. In *Proceedings of the HLT-NAACL Workshop on Text Meaning*, Edmonton, Canada, May.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Goteborg University, Goteborg, Sweden.

Robert MacGregor and Robert Bates. 1987. The LOOM knowledge representation language. TechReport ISI-87-188, USC/ISI, Berkley, California.

Michael McTear. 1998. Modelling spoken dialogues with state transition diagrams: Experiences with the CSLU toolkit. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, Sydney, Australia.

Kary Myers, Michael Kerns, Satinder Singh, and Marilyn A. Walker. 2000. A boosting approach to topic spotting on subdialogues. In *Proceedings of the 17th International Conference on Machine Learning (ICML-2000)*, Stanford, USA.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th Workshop of the ACL SIG on Discourse and Dialogue*, Boston, USA, April.

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

Listener reaction to referential form

Gunnvald B. Svendsen[1], Bente Evjemo[1], Jan A. K. Johnsen[2] & Svein Bergvik[1]
[1]Telenor Research & Development
{gunnvald-bendix.svendsen/bente.evjemo/svein.bergvik}
@telenor.com
[2]Norwegian Center for Telemedicine
jan.are.kolset.johnsen@telemed.no

Abstract

In a recent paper it has been shown that observers use referential form as an indication of how well acquainted interlocutors are. In the present study it is investigated if the referential form used by the speaker influences the listeners evaluation of the speaker. An experiment with eighty subjects was conducted. Subjects were told to imagine themselves being spoken to by a stranger and to rate how agreeable they would perceive the stranger depending upon the utterances he or she made. Sentences that referred both implicitly and explicitly to a shared experience. were employed in the experiment. The results indicates that listeners are rating speakers as more agreeable when the speaker is using an explicit rather than an implicit form of reference . Two explanations are suggested and the results are discussed in relation to relation formation in text based computer mediated communication. It is suggested that referential form could function as a cue in this context.

1 Introduction

Referring is a central aspect of communication, and studies of referential form, have generally[1] shown that interlocutors follow Grice's cooperative principle of communication: "Make your conversational contributions such as is required, at the stage at which it occurs, by the accepted purpose of the talk exchanges in which you are engaged" (Grice, 1975, p. 45). In deciding what form of reference to use in a conversation, the speaker may take the listener's perspective. For instance speakers shape their way of referring based on the assumed knowledge of the listener (Clark & Wilkes-Gibbes, 1986, Fussell & Krauss, 1992, Isaacs &. Clark , 1987) and the cognitive load experienced (Horton & Keysar, 1996, Rossnagel, 2000). It is also well known that the referential expression is more effective when the topic is introduced at a later stage in a dialogue (Krauss & Fussell, 1990). The first time the topic is introduced, the referring could be rather complex:

> "I'm looking for an invoice from Doe et co, it's **pink and letter-sized**".

When the topic is introduced anew at a later stage it's shortened:

[1] But see Bard et al 2000, Keysar, 1994, Keysar, 1997.

"Could you help me find **the invoice?**"

Referring by using a pronoun, as an implicit reference, is an example of an extreme simplification:

"I have found **it**!"

Implicit referring has two salient features, it is efficient (for the speaker) and it is easily misunderstood. Take the following example. Two persons, A and B, wait at a street corner for the light to turn green. A young boy runs into the street and is nearly overrun by a car. The boy makes it over the street, the car continues, the light turns green and A and B walk away. By coincidence A stands behind B in line to by a newspaper later that day and A says as they see each other:

1) "If he continues to be so careless an accident is bound to happened sooner or later".

or

2) "If that boy who ran out into the street continues to be so careless an accident is bound to happened sooner or later".

In example 1) A refers implicitly to the event in his or her utterance. The utterance is efficient but could also be misunderstood, for instance B might wonder if A refers to the boy or to the driver. According to Grice's cooperative principle of communication, A must phrase the utterance so that it is quite evident for both A and B what A refers to. Thus in 1) A seems to imply something like "You know what I am thinking about, and I know that you know". The utterance in example 2) does not carry the same implication.

Recently it has been suggested that the referential form implies a relation between the interlocutors. Svendsen and Evjemo (2003) showed that observers perceive interlocutors that use implicit referring as having a closer relation than interlocutors that make explicit references. Svendsen and Evjemo explained the phenomenon by arguing that an implicit reference implies more than just a shared experience. Following Grice's cooperative principle, the speaker must be quite sure that the listener remembers the event and that the listener understands what he or she is referring to. Thus they argued that an implicit reference to a shared experience implies a higher degree of familiarity between the interlocutors than an explicit reference. Svendsen and Evjemo further showed that implicit referring occurs more often between family members than between colleagues, while it is less frequent in phone conversations than in face-to-face conversations. They suggested that the latter partly could account for the alienation felt in phone conversations compared to face-to-face conversations.

In the above-mentioned study, subjects rated interlocutors that used implicit referring as having a closer relation than interlocutors that used explicit referring. The finding is hardly surprising since an implicit reference to some topic tends to be meaningless if the interlocutors have no shared experience, while an explicit reference to the same topic is meaningful. Thus, a person that listens in on a conversation and assumes that the interlocutors try to make sense to each other, also must assume that the interlocutors that use implicit referring are better acquainted than those who do not refer in this way - other things being equal.

While it is no surprise that a person listening in on a conversation assesses the relation between the interlocutors in this way, it is harder to predict how the person spoken to would react to being spoken to in an implicit versus explicit manner. First, it is quite possible that the listener will not react differently to utterances referring implicitly or explicitly to an earlier shared experience. However, it could also be argued that a listener would prefer being spoken to in an explicit manner. In the above example B would react positively to being spoken to in an explicit manner because this requires less mental effort on his or her part. It could also be argued that B, under certain circumstances, would prefer the implicit utterance if an implicit utterance indeed imply familiarity between speaker and listener. The present study tries to shed light on this issue

by investigating how subjects react to implicit and explicit utterances.

2 Method

An experiment was conducted to assess whether the way a speaker refers to a shared experience influences how the listener perceive the speaker.

2.1 Subjects

Eighty subjects (Ss) aged 18 to 51 were recruited from a broad population including university students and teachers, pre-school personnel, undergraduate teachers, researchers and administrative personnel. The subjects were not given any compensation for their participation.

2.2 Procedure

Ss were given a short text describing of an imagined situation where they were standing in line at the cashier in a supermarket. A stranger was standing next in line behind them. When Ss came to the cashier they couldn't find their wallet, and after some searching, stepped out of the line and asked the stranger to pass them. The stranger passed, paid and left. After a moment they found their wallet, paid and left. Ss were further instructed to imagine that they met the stranger at a later time and that the stranger in this meeting uttered one of four sentences.

Sentences in explicit form:
A) Did you find the wallet when you wanted to pay yesterday?
B) Did you get to pay when you were at the store yesterday
C) It has happened to me a lot of times that I have been standing
looking for my wallet like you did when you wanted to pay yesterday
D) Did you have to leave empty handed or were you able to pay at
the cashier yesterday ?
Sentences in implicit form:
A) Did you find it?
B) Did you get to pay?
C) It has happened to me a lot of times.
D) Did you have to leave empty handed?

Table 1. The four sentences (A to D) used in the experiment in explicit and implicit form

For each sentence Ss were asked to rate how agreeable or sympathetic[2] they would have perceived the stranger to be if he or she had uttered that sentence. The rating was done on a 7-point Likert scale, with the anchors "very little" and "very much". Ss were given a sheet of paper with a description of the situation, the four different sentences, and the seven point rating scale beneath each of the sentences.

2.3 Design and analysis

The sentences had either an implicit or an explicit form as seen in Table 1. Subjects were randomly divided into two groups, an "*explicit – implicit"* group and an "*implicit-explicit*" group, with forty ss in each group. In the *explicit-implicit* group the subjects were presented with sentence A in explicit form, sentence B in implicit form, sentence C in explicit form and sentence D in implicit form. In the *implicit – explicit* group the order was reversed, so that sentence A had implicit form, sentence B had explicit form and so on.

Thus the independent variables were *presentation order* (*order*) with the levels "implicit – explicit" and "explicit – implicit", and *sentence* with four different levels corresponding to sentence A to D. *Order* is a between groups factor, while *sentence* is a repeated measure, within subjects, factor. Thus the design is a 2 way mixed model ANOVA with 2 X 4 levels. The dependent variable was the Ss score on the Likert scale.

A significant *sentence* effect would indicate that the sentences used make different impressions on the Ss, which is neither surprising nor interesting. A significant *order* effect would indicate that presentation order as such plays a part in the results. That would be a spurious effect. The interesting effect is the *order*sentence* interaction. A significant interaction would indicate that the referring used in the sentences influences

[2] The exact wording in Norwegian was: "Hvor sympatisk opplever du denne personen". The Norwegian concept "sympatisk" is roughly equivalent to the English concept agreeable.

how agreeable Ss think the speaker is, thus supporting that hypothesis.

3 Results

The analysis reveals that both the factor *sentence* and the *order*sentence* interaction are highly significant (see Table 2). As stated earlier, a significant sentence factor merits no interest. The significant interaction shows, however, that referential form influences the listener's evaluation

Effect	**SS**	**df**	**F**	**p**
Order	4.3	1	1,3	>0.2
Sentence	79,8	3	26,6	<0,000
Sentence X Order	43,9	3	14,6	<0.000

Table 2: Results of ANOVA

of the speaker. The interaction is easily seen in figure 1. The dotted line represents the likeability scores when the sentence A is uttered in it is explicit form, B in it is implicit, C in it is explicit and D in it is implicit. The solid line represents

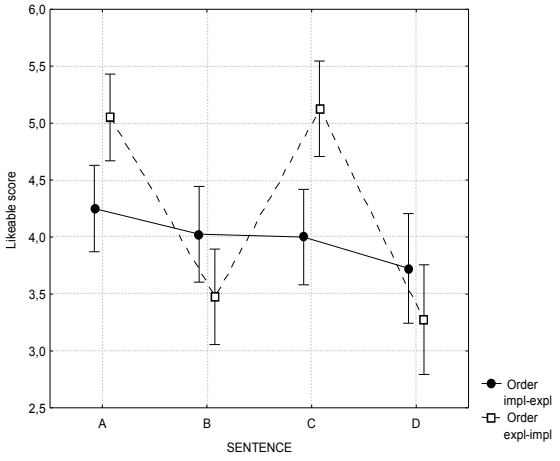


Figure 1. Ss assessment of how agreable a speaker is perceived depending upon the sentence spoken and whether it refers explicitly or implicitly to the theme

the scores when the order is reversed, that is: implicit-explicit-implicit-explicit. In all four instances, uttering the sentences in explicit form is scored as more agreeable or sympathetic than uttering the sentences in implicit form. Thus, the results show two things: First, the referential form of an utterance influences how the listener reacts towards the speaker. Secondly, given the circumstances of this experiment, the listener prefers an utterance that is framed in an explicit manner over one that is framed implicitly.

.

4 Discussion

The results show that people react to referential form and that an implicit way of framing a sentence makes a less agreeable impression than framing the sentence explicitly. The result can be explained by what may be termed the "effort-hypothesis" eluded to earlier; that a sentence in implicit form required more mental effort of the listener. Thus, a speaker that uses implicit referring makes it easy on him or herself by imposing work on the listener. It is hardly surprising that subjects find this the least likeable course of action.

The results could also be explained in another way, which may be coined the "relation-hypothesis". In the experiment the subjects were told that a stranger spoke to them. A speaker who wishes to be understood and uses an implicit reference to a shared experience must assume that the listener remembers the instance referred to quite well, or else the speaker is in breach of the Griceian cooperative principle. Thus a speaker using an implicit reference meta-communicates to the listener that "I remember our previous engagement quite well, and I assume that you remember it too." The listener presumably interprets this meta-communication contingent upon his or her relation to the speaker. If the relation is one that the listener wants or regards as positive, he or she would probably appreciate that the speaker remembers their earlier engagement. If the listener does not want any relation to the speaker, does not like the relation they have, or does not know the speaker, the implication that the listener remembers their shared history could be interpreted as imposing and rude. It might be argued that this is what happened in the present experiment.

The results do not indicate which of these explanations to favor, but the explanations give dif-

ferent predictions. The “effort-hypothesis” predicts that the listener would prefer the speaker to use explicit reference regardless of the relationship the listener would like to have to the speaker. The “relation-hypothesis” predicts that the listener would prefer that the speaker use implicit referring when the listener wants or regard a relation to the speaker as positive. Thus, further research should make it possible to choose between the hypotheses.

The results might shed light on relation formation and impression formation in text based computer mediated communication (CMC). Classic theories of media choice and media effects, with the so called 'cues filtered out' perspective (Culnan and Markus, 1987) predict that computer mediated communication would lead to task oriented communication and little or no relation forming because the media lacks the ability to convey non-verbal cues (Daft and Lengel, 1984, Rutter 1987, Short et al 1976, Sproull and Kiesler, 1986) However, it is well documented that relationships are formed through CMC (Kummervold et al 2002, Lea and Spears, 1995, Park and Floyd, 1996, Utz, 2000). This has led to theories that explain relation formation in spite of a largely textual communication channel (Jacobson, 1999, Lea and Spears, 1995, Walther, 1992). These theories assume that interlocutors are motivated to develop impressions of others in spite of limitations in media, and that they utilize the cues they have at their disposal both to give and gain information. The relation forming process takes longer however, since few cues are available [Walter, 1992, Walter et al 2001).

Research aimed at uncovering the cues used to convey impressions and build relations in CMC points to verbal and textual cues, like self disclosure, language intensity, participants’ screen names, form of address, and the discourse in which they engage as central (Jacobson, ibid, Walter and Burgoon, 1992). Further, participants’ linguistic style seems to play a role (Lea and Spears, 1992). Apart from the verbal messages themselves, chronemic cues, ie information about when messages have been sent, and emoticons or smilies have been shown to play a role in assessment of messages (Walther and D’Addario, 2001, Walther and Tidwell, 1995).

Granted that referential form influences how interlocutors perceive each other, as the present results indicate, referential form must be considered a new candidate as a cue users may employ in assessing each other on-line. This is especially the case since referential form can be manipulated just as well in text as in speech.

The present research will be continued with two foci. First it will be investigated which of the two hypotheses set forth earlier fit the facts best. Secondly, the use of referential form in CMC will be investigated..

References

Bard, E G, Anderson, A H, Sotillo, C, Aylett, M, Doherty-Sneddon, G, & Newlands, A Controlling the intelligibility of referring expressions in dialogue. Journal of Memory and Language, 42,1(2000), 1-22.

Clark, H H, & Wilkes-Gibbes, D. , “Referring as a collaborative process”, *Cognition*, 22 (1986),1-39.

Culnan, M J.; & Markus, ML. Information technologies. In F. M. Jablin, et al, Eds, *Handbook of organizational communication: An interdisciplinary perespective*. (420-443). Newbury Park, CA: Sage1, 1987.

Daft R, Lengel R H. Information richness: A new approach to managerial behavior and organizational design. In Staw B, Cummings L L, (eds) *Research in Organizational Behavior*. Greenwich, Conn. JAI Press, 191-233, 1984.

Fussell, S R. & Krauss, R M. Coordination of knowledge in communication: Effects of speakers' assumptions about others' knowledge. *Journal of Personality and Social Psychology,* 62 (1992), 378-391.

Grice, H P. Logic and conversation, in Cole, P. and Morgan, J. (eds), *Speech Acts,* New York, Academic Press, (1975), 41-58.

Horton, W S & Keysar, B When do speakers take into account common ground? *Cognition*, 5, (1996),91-117

Isaacs, E A, &. Clark H H. References in conversations between experts and novices. J*ournal of Experimental Psychology: General*, 116, (1987), 26-37

Jacobson, D. Impression Formation in Cyberspace: Online Expectations and Offline Experiences in Text-based Virtual Communities, *Journal of Computer-Mediated Communication,* 5,1,(1999)

Keysar, B. The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology* 26, (1994)165-208

Keysar, B. Unconfounding common ground. *Discourse Processes*, 24, (1997), 253-270.

Krauss, R M & Fussell, S R. Mutual knowledge and Communicative effectiveness In J. Galegher, R. Kraut, C. Edigo (eds) *Intellectual Teamwork*. Hilsdale: LAE 111-145, 1990.

Kummervold P E, Gammon D, Bergvik S, Johnsen J A, Hasvold T, Rosenvinge J H. Social support in a wired world: Use of mental health discussion forums in Norway. *Nordic Journal of Psychiatry*; 56 2002), 59-65

Lea, M., & Spears, R. Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing.* 2, (1992) 321-341.

Lea, M., & Spears, R. Love at first byte? Building personal relationships over computer networks. In J. T. Wood & S. Duck (Eds.), *Understudied relationships: Off the beaten track,* 197-233. Newbury Park, CA: Sage 1995

Parks, M. R.; & Floyd, K. Making friends in cyberspace. Journal of Communication, 46, 80-97 also *Journal of Computer Mediated Communication*,46,1(1996).

Rossnagel, C. Cognitive load and perspective taking: applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology*, 30 (2000) 429-445

Rutter D R. *Communicating by Telephone*. Oxford, Pergamon Press, 1987

Short J, Williams E, Christie B. The social psychology of telecommunication. London, Wiley, 1976

Sproull, L., & Kiesler, S. . Reducing social context cues: Electronic mail in organizational communication. Management Science, 32,11 (1986)1492-1512.

Svendsen, G.B. & Evjemo, B. Implicit referring as an indication of familiarity in face-to-face and phone conversations. *Proc. Interact* 2003, 920-923.

Walther, J B. Interpersonal effects in computer mediated interaction: A relational perspective. Communication Research, 19,1(1992) 52-90.

Walther, J. B., & D'Addario, K. P. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review, 19*,(2001) 323-345.

Walther, J. B., & Burgoon, J. K. Relational communication in computer-mediated interaction. *Human Communication Research,* 19,1(1992) 50-88

Walther, J. B., Slovacek, C., & Tidwell, L. C. . Is a picture worth a thousand words? Photographic images in long term and short term virtual teams. *Communication Research*, 28, (2001)105-134.

Walther, J. B., & Tidwell, L. C. Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing,* 5, (1995) , 355-378.

Utz, S. Social information processing in MUDs: The development of friendships in virtual worlds. *Journal of Online Behavior*, 1,1(2000).

Semantics, Dialogue, and Reference Resolution

Joel Tetreault
Department of Computer Science
University of Rochester
Rochester, NY, 14627, USA
tetreaul@cs.rochester.edu

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

Abstract

Most pronoun resolution research has focused on written corpora while using syntactical and surface cues. Though big gains have been made in this domain with those methods, it is difficult to do better than the 80% coverage in these domains without some world or semantic knowledge. We investigate this issue by incorporating rich semantic information into a proven reference resolution model over a very difficult domain of human-human task-oriented dialogues. Our results show that semantic information greatly improves performance and can even be viewed as a substitution for the usual syntactic filters.

1 Introduction

In this paper we present an automated corpus-based analysis of pronoun coreference resolution using semantics in a spoken dialogue domain. Most work in pronoun resolution has focused on using syntactic and surface features such as word distance or number of mentions to help improve accuracy rates. While many of these methods perform quite well on large corpora, (for example, (Tetreault, 2001), (Mitkov, 2000)), it seems that these methods can't do much better than 80% accuracy. Error analyses from these studies suggest that other information such as discourse structure, semantic information, reasoning, etc. is required to resolve these hard cases, which typically elude most pronoun resolution algorithms.

In addition, while most empirical work in the field has used large corpora of written text as the basis for evaluation, very little work has been conducted on spoken dialog domain, which are very important for use in natural language understanding systems. These domains are much more difficult than their written counterparts because speech repairs, interruptions, and other disfluencies make it hard to get reliable parses, and also very hard to track the focus (Byron and Stent, 1998). For example, (Byron, 2002) showed that syntax and salience metrics that would perform at 80% on Wall Street Journal articles could only perform at 37% over a large task-oriented spoken-dialog domain. Clearly, something other than syntax and surface methods are necessary for successful reference resolution. Furthermore, what work has been done in reference in spoken dialogs has focused on distinguishing between coreferential and demonstrative pronouns, and then the different types of demonstratives, and then trying to resolve each type (Eckert and Strube, 2000), (Byron, 2002). These metrics typically use semantic information of the verb and tracking of acknowledgments to determine type.

In our study we assume knowledge of the type of each pronoun and focus our work on coreferential pronouns specifically. This research is novel in two ways - first, we use semantic knowledge generated from a deep parser, along with surface constraints to aid in resolution; second, we evaluate our algorithm over a large spoken dialog do-

main. The results show that including semantics improves reference resolution. In the following section we discuss our spoken dialogue corpus. Next we discuss the algorithm and close with results and discussion.

2 Corpus Description

Our corpus consists of transcribed task-oriented dialogs between two humans called the Monroe domain (Stent, 2001). In these domains, one participant was given the task of resolving several medical and weather emergencies in a city by allocating resources to resolve all of them in a timely fashion. The other participant acted as a system to aid the first in planning.

Corpus construction (Swift et al., 2004) and (Tetreault et al., 2004) consisted of three phrases: disfluency annotation, parsing, and reference annotation. We annotated our corpus for disfluencies by marking all repeated phrases, repaired phrases, and also marking incomplete and ungrammatical sentences. Examples of incomplete and ungrammatical utterances are: *Actually it's right a ab* and *So ambulance sends generator*.

After removing the disfluencies, each sentence is parsed by a broad-coverage deep parser. The parser works by using a bottom-up algorithm and an augmented context-free grammar with hierarchical features. The parser uses a domain independent ontology combined with a domain model for added selectional restrictions and to help prune unlikely parses. The output is a syntactic and semantic representation of a sentence.

The semantic representation is a flat unscoped logical form with events and labeled semantic roles. Each term has associated with it an identifying variable, semantic relationships to other terms, and a semantic vector describing the term. The vector is a typed feature list meaning that there is a main type associated with the term (in our case, one of: physical object, abstract object, situation, and proposition) which licenses certain secondary features. For example, a physical object type would license features such as form, origin, mobility, intentional, etc. Likewise, a situation feature type would license features such as aspect, time-span, cause, etc. Each feature has a list of possible values. Some are binary such as the container feature which means an entity can either hold something, or it can't. And some have a wide range such as mobility: fixed, self-moving, non-self-moving. Examples of a term and the semantic vector (see the :SEM field) for the entity (an ambulance) are shown in Figure 1.

The parser was run over the entire corpus of 1756 utterances and its syntactic and semantic output was handchecked by trained annotators and marked for acceptability. The parser was able to correctly parse 1334 (85%) of the utterances. Common problems with bad utterances were incorrect word-senses, wrong attachment in the parse tree, or incorrect semantic features. For our purposes, this meant that there were many pronouns that had underconstrained semantics or no semantics at all. Underconstrained pronouns also can be found in utterances that did parse correctly, since sometimes there is simply not enough information from the rest of the sentence to determine a semantics for the pronoun. This becomes problematic in reference resolution because an underconstrained semantics would tend to match everything. We decided not to manually parse the utterances that did not parse correctly because we felt a reference resolution model operating in a spoken dialogue domain will have to deal with bad parses and one wants their results to reflect the "real world" situation. Sentences deemed ungrammatical or incomplete were omitted from the parsing and hand-checking phase. We felt that since there were pronouns and possible antecedents in these utterances, it is necessary to maintain some representation of the utterance. So each term in these sentences were generated manually.

The third phase involved annotating the reference relationships between terms. We annotated coreference relationships between noun phrases and also annotated all pronouns. Our annotation scheme is based on the GNOME project scheme (Poesio, 2000) which annotates referential links between entities as well as their respective discourse and salience information. The main difference in our approach is that we do not annotate discourse units and certain semantic features, since most of the basic syntactic and semantic features are produced automatically for us in the parsing phase. We labeled each pronoun with one of the

```
(TERM :VAR V213818
        :LF (A V213818 (:* LF::LAND-VEHICLE W::AMBULANCE)
        :INPUT (AN AMBULANCE))
        :SEM ($ F::PHYS-OBJ
                (SPATIAL-ABSTRACTION SPATIAL-POINT) (GROUP -)
                (MOBILITY LAND-MOVABLE) (FORM ENCLOSURE)
                (ORIGIN ARTIFACT) (OBJECT-FUNCTION VEHICLE)
                (INTENTIONAL -) (INFORMATION -)
                (CONTAINER (OR + -)) (TRAJECTORY -))
                )
```

Figure 1: Excerpt semantic features for "an ambulance"

following relations: coreference (pronoun is in an identity relation with another explicitly mentioned entity), speaker (one of the discourse participants), action (pronoun refers to an event), demonstrative (pronoun refers to an utterance or discourse segment), and functional (pronoun is related to an entity by an indirect relationship). We had a team of annotators work on the files and agree on how to tag each pronoun.

After the annotation phase, a post-processing phase identifies all the noun phrases that refer to the same entity, and generates a unique chain-id for this entity. This is similar to the $ante$ field in the GNOME scheme. The advantage of doing this processing is that it is possible for a referring expression to refer to a past instantiation that was not the last mentioned instantiation, which is usually what is annotated. As a result, it is necessary to mark all coreferential instantiations with the same identification tag.

So the final parsed corpus consists of lists of entities for each sentence. These entities are verbs, noun phrases, etc, and each has a semantic vector associated with it, though at varying degrees of acceptability depending on the parser success. Noun phrases and pronouns entities are annotated for reference.

3 Algorithm

We use a modified version of the Left-Right Centering algorithm (LRC) (Tetreault, 2001) to determine how much of an effect using semantics has in pronoun resolution in a spoken dialogue. We selected this algorithm because it is easy to use and has performed well in other large domains. It works as follows: while processing a sentence, put each noun phrase encountered on a temporary list, and once the sentence has been completely processed, place the temporary list on a history stack. When a pronoun is encountered, we first search the temporary list's elements from left to right taking the first entity (noun phrase) that fits constraints imposed by the pronoun and the context. If a suitable antecedent is not found, we search through the history stack, searching each sentence from left to right.

3.1 Additional Syntactic Constraints

Normally in LRC, the temporary list is sorted by grammatical function (subject, direct object, etc.) before being placed on the history stack. In our domain, syntax is not very helpful in ranking entities within a sentence since the sentences are so short, so we simply rank the list by word order.

We found that gender constraints, though common in written text evaluations, were more of a drawback then an aid. It was not uncommon in our corpus for people to refer to a person with a medical condition with *that*, or to refer to a digging truck with *he*. Number constraints were encoded in the :LF of the term as SET-OF, so it is easy to tell if an entity is a set or not (see Figure 2). Noun phrases such as road crew which have a singular representation but implicitly represent a group of people do not have the SET-OF notation in the :LF but have in their semantics the GROUP feature. When semantics is used, we leverage this information to allow these types of noun phrases to be referred to by plural and singular pronouns.

In addition to the number constraints, we also implemented three other syntax based constraints: binding, predicate-NP linking, and location ranking. Binding is a standard linguistic constraint

```
(TERM :VAR V3337536
      :LF (PRO V3337536 (SET-OF (:* LF::REFERENTIAL-SEM W::THEM))
      :INPUT (THEM)
      :SEM ($ F::PHYS-OBJ
            (F::MOBILITY F::MOVABLE))

            )
```

Figure 2: Excerpt semantic features for "them"

which prevents from noun phrases within the same verb phrase from co-referring. So in the sentence: *They will move that* the two pronouns would be prevented from referring to each other. The obvious exception is reflexives, though none exist in the corpus. This constraint only works well if the utterance parsed properly. There are some instances where a sentence was parsed into fragments so the binding constraint fails.

Predicate-NP linking is the process of replacing an underspecifed pronoun's semantics in a *be* verb phrase with that of its predicate. So in the sentence *it is the digging truck at Avon it* is underspecified but is in a identity relation with the co-theme of the verb phrase so, we replace the pronoun semantics with that of the truck's.

The final constraint, location ranking, is based on research (Tetreault, 2002) on implicit roles which showed that putting a preference order on verb location roles (ie. TO-LOC - where an entity is being taken, FROM-LOC -where an entity is coming from, and AT-LOC, where an entity is situated) improves resolution of implicit roles in a spoken dialogue. Since the dialogues are basically plan-based narratives, where an entity is taken to is more likely to be referred to by a subsequent pronoun than where it was taken from. So when searching for the antecedents for pronouns *there* and *here*, one looks back through each utterance in the discourse history, first re-ranking the possible location candidates, with entities in a TO-LOC role preferred over those in a FROM-LOC role, preferred over those in AT-LOC role or no role at all. For example, in the utterance *Send the digging truck from Elmwood to Mt. Hope* the preferred candidate would be *Mt. Hope* whereas in the original LRC formulation, *Elmwood* would be selected.

3.2 Semantic Filter

A semantic match occurs when the main type between the pronoun and antecedent are the same, and there is no conflict between the features (for example, a match would not occur if the pronoun were mobile but it's candidate was non-moving, but that feature would match if the candidate were self-moving). For pronouns with an underspecified semantics, we simply select the first entity in our search path that meets the remaining constraints. In our study, we only investigate pronouns marked for coreference. Pronouns with other relations, such as functional or demonstrative, were not considered.

4 Evaluation

For our evaluation we selected two baselines (both knowledge-poor versions of LRC): the first uses no semantic knowledge at all and simply selects the first noun phrase in the search regardless of constraints. The second incorporates number and binding constraints. This represents the canonical pronoun resolution constraints used in most systems. The results of both baselines are in Figure 3. The second column indicates what percentage of the 278 pronouns that each algorithm resolved correctly. The fourth column shows how many of the 83 underconstrained pronouns were resolved correctly, and the final column shows how many pronouns with acceptable semantics (out of 195) were resolved) correctly.

The additional rows in the table represent the cumulative effects of adding a constraint onto the constraints in the preceding rows. So adding the location constraint on top of the binding and predicate-NP constraints (and the basic baseline constraints) produces an improvement of 3.2% over not using the constraint. The final row rep-

resents only adding semantics to the baseline constraints.

The main result from this evaluation is that including semantics significantly improves pronoun resolution accuracy. The three syntactic constraints improve performance over the second baseline by 6.5%, or an error reduction of 20.8%. The biggest increase comes from adding semantics (5.4%), or a cumulative error reduction of 31.9%. Another positive outcome from this study is how much only using semantics improves things over the baseline. So from the standpoint of building a natural language system where response time is important, only using the semantic filter is a reasonable alternative to employing a battery of filters on top of semantics.

Another boost can be seen in resolving pronouns with semantics, as it resolves 26 more. This also reflects how useful it is to have a well-parsed corpus to get acceptable semantics for each entity.

We conducted a detailed analysis on the 92 pronouns resolved incorrectly to identify the main categories for error:

Wrong semantics (22) Cases where a bad parse leads to incorrect semantics for either the pronoun or its antecedent so there would be no way for a match to occur. The most common error was plural pronouns having a top-level semantic feature of situation when it should have been physical object. So these pronouns would incorrectly match with events in the discourse as opposed to a set of people, road crews, vehicles, etc.

Underconstrained pronoun - (15) Here there is either not enough information from the rest of the sentence for the parser to give a rich semantics for the pronoun. This means that the pronoun will match more entities than it should.

Difficult (13) There were ten cases in the corpus that required a combination of information and reasoning to resolve the pronoun correctly. Most of the time, the pronoun fit several of the error categories.

Three of the errors were related to discourse structure where some notion of common ground or embedded structure could be helpful in eliminating candidates during search. Usually this happens when pronouns have a long distance antecedent but the intervening utterances are an aside and not related to the topic of the pronoun's sentence. For example, utterances 10 and 11 in Figure 4 are an aside and if removed would prevent *it* from resolving to the *disability*.

UTT8	U	i can't fi nd the rochester airport
UTT9	S	it's
UTT10	U	i think i have a disability with maps
UTT11	U	have i ever told you that before
UTT12	S	it's located on brooks avenue

Figure 4: Excerpt from dialog s2

Bad Parse with intervening candidates (9) Unlike the first case, the semantics for the pronoun and entity are acceptable but intervening entities have incorrect semantics that coincidentally match with the pronoun's semantics. Because the algorithm works by selected the first candidate that meets all constraints, this intervening candidate is selected before the real antecedent is considered.

Pred-NP Binding (8) These cases involved pronouns in utterances that did not parse and thus binding constraints were not able to function. So the pronoun would refer to an entity intrasententially when it really should be blocked.

Locatives (8) The locative ranking method does improve performance for *there* and *here* but there are some cases where that ranking fails. For example, *Strong Hospital* in *the ambulance from Strong* should not be highly ranked because it is in an embedded phrase. And in Figure 5, our algorithm selects *east main* as the most salient entity, but the pronoun at the end refers to *rochester general*.

Set (6) We currently don't handle plurals with multiple antecedents, so the 6 cases of set membership are automatically wrong.

Algorithm	% Right	Right	USP Right	ACC Right
baseline 1	44.6%	124	43	81
baseline 2	55.0%	143	51	102
+binding	57.9%	161	54	107
+pred-np	58.3%	162	54	108
+location	61.5%	171	54	117
+semantics	66.9%	186	54	132
b2+semantics	65.5%	182	54	128

Figure 3: Pronoun Resolution Algorithm Performance

UTT198 S so i'm just gonna take the ambulance from rochester general to east main back to rochester general so that we have one ambulance there

Figure 5: Locatives Example

Intervening Candidate (6) In this case, all parses in the local context are good but there is a candidate that matches the pronoun but is not the correct antecedent.

Functional Semantics (2) There were two cases of pronouns in a functional relation being referred to by a co-indexing pronoun. These errors are due to metonymy.

The error analysis shows the effect of erroneous parses on performance. 39 of the errors (wrong semantics, bad parse with intervening candidates, and pred-NP binding) are due to bad parses producing incorrect semantics for the entities. This shows the difficulty to NLP systems that spoken dialoguess impose. Difficult sentences lead to incorrect parses which then can severely effect reference performance. On the other hand, the error distribution shows the great gains that can be made by getting better parses or by compensating with other metrics. Despite the underspecified semantics for some pronouns, or incorrect semantics, using semantics really improves accuracy instead of harming it.

5 Conclusion

In short, we performed an automated empirical evaluation of pronoun coreference resolution in a large spoken dialog domain using rich semantic information from a deep-parser. The results show that semantic information improves performance over recency-based heuristics, and despite the complications imposed by spoken dialogue.

Future work will include researching ways of dealing with underspecified pronouns and also using discourse cues, grounding, and thematic roles of verbs to further aid resolution.

6 Acknowledgments

Partial support for this project was provided by ONR grant no. N00014-01-1-1015, "Portable Dialog Interfaces" and NSF grant 0328810 "Continuous Understanding".

References

Donna K. Byron and Amanda Stent. 1998. A preliminary model of centering in dialog. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, student session*.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 2002 annual meeting of the Association for Computational Linguistics (ACL '02)*, pages 80–87, Philadelphia, USA, July.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

R. Mitkov. 2000. Towards a more consistent and comprehensive evaluation of anaphora resolution algo-

rithms and systems. In *2nd Discourse Anaphora and Anaphora Resolution Colloquium*, pages 96–107.

M. Poesio. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *LREC '00*, Athens.

Amanda J. Stent. 2001. *Dialogue Systems as Conversational Partners: Applying Conversation Acts Theory to Natural Language Generation for Task-Oriented Mixed-Initiative Spoken Dialogue*. Ph.D. thesis, University of Rochester.

M. Swift, M. Dzikovska, J. Tetreault, and James F. Allen. 2004. Semi-automatic syntactic and semantic corpus annotation with a deep parser. In *LREC'04*, Lisbon.

Joel Tetreault, Mary Swift, Preethum Prithviraj, Myroslava Dzikovska, and James Allen. 2004. Discourse annotation in the monroe corpus. In *ACL '04 Workshop on Discourse Annotation*, Barcelona, Spain, July 25-26.

Joel R. Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Joel R. Tetreault. 2002. Implicit role reference. In *2002 International Symposium on Reference Resolution for Natural Language Processing*, pages 109–115.

A free-format dialogue protocol for multi-party inquiry

Gerard Vreeswijk and **Joris Hulstijn**
ICS, Utrecht University,
P.O.Box 80.089, 3508 TB Utrecht
The Netherlands
gv@cs.uu.nl, jorish@vs.uu.nl

Abstract

In this paper we propose a formal account of multi-party inquiry. Inquiry is the dialogue game type in which participants try to get a common understanding of some open problem concerning an external state of affairs. We discuss some important issues for multi-party dialogue in general, and extend a simple account of inquiry, such that it accounts for the multi-party case.

1 Introduction

Formal dialogue is rapidly gaining status as a new paradigm for automated forms of information exchange. In this paper we consider the dialogue game of *inquiry*. According to Walton and Krabbe's typology of dialogue game types, inquiry is "a type of dialogue which strives to establish or 'prove' propositions in order to answer a question (solve a problem) in such a way that a stable and general agreement on the matter at issue results" (Walton and Krabbe, 1995, p. 72). Inquiry differs from persuasion dialogue in that it does not start from a conflict, but from an open problem. It differs from information exchange, in that parties have a common goal to reach agreement. For information exchange, the goal is mere dissemination of information.

In this paper we focus on multi-party inquiry: inquiry performed by a group of more than two cooperative agents. Dignum and Vreeswijk (2003) describe a simple protocol for multi-party inquiry. The protocol has a fixed turn-taking mechanism. Here we extend the protocol, looking in particular at different coordination and turn-taking mechanisms. Moreover, we believe that once we have solved the simpler case of inquiry, it becomes easier to extend the results to multi-party negotiation and persuasion dialogues.

The paper is structured as follows. First we give definitions for single party inquiry. Then we look at multi-party issues in general, such as open versus closed systems, roles and coordination. Section 3 contains an inquiry protocol adapted for the multi-party case.

2 Issues in multi-party dialogue

Multi-party dialogue can be conducted in various ways. To sketch the possibilities we briefly sketch the landscape. The following issues arise when considering dialogue games for more than two participants (Dignum and Vreeswijk, 2003).

Open vs. closed systems An issue that comes up right away, is who the participants are. In a closed system, all parties are present during the whole dialogue. Entry and exit to the dialogue is controlled, and therefore we can assume that each participant satisfies a basic set of assumptions. In an open system, any agent can join later or leave before the end of the dialogue. No assumptions can be made on for instance, common ground or use of vocabulary.

Roles A following issue is the role of each of the parties in the dialogue (Hulstijn, 2003). This can be looked upon from different perspectives. First, there are roles related to addressing. In a two-

party dialogue there is always a speaker and an addressee. In a multi-party dialogue we can distinguish: speaker, addressee, auditor, overhearer and eavesdropper (see e.g. (Bell, 1984)). Second, there are roles constituted by the particular dialogue game type. Such roles define the expectations, preferences and dialogue game rules associated with a participant. For example, in a typical two-party persuasive dialogue there is a proponent and an opponent. However, for dialogues of inquiry or deliberation, the distinctions already get blurred. Third, there are roles that depend on the social organization of the interaction situation. A good example is that of a chairperson. Such roles determine turn taking, termination or entry and exit to the dialogue.

For each of the perspectives on roles one can choose whether roles are fixed once or can change during the dialogue. Again, we may need specific communicative acts or rituals to signal such changes.

Channel One may distinguish between synchronous and asynchronous communication channels. This choice has repercussions for addressing too. For example, in an asynchronous channel that stores messages for a long time, such as a newsgroup, we may expect many overhearers. By contrast, a synchronous medium, such as speech, is less suitable for one-to-many communication.

Coordination On a synchronous channel, only one party can speak at a time. Therefore one needs a turn taking mechanism. We could use a round-robin protocol, which is a generalization of the strict turn-taking for two parties. Otherwise one could have a chairman explicitly assign turns. On an asynchronous channel, in principle everybody may speak at the same time.

Termination Participants engage in a dialogue for some particular purpose. This purpose differs for each dialogue game.

3 Inquiry

This section describes a simple dialogue game of inquiry. The game has two participants: *expert* and *nature*. The expert has knowledge about some particular aspect of the world, called a topic. For example, an agent may be an expert on the topic of financial information, or on real estate. The expert may do observations to extend its knowledge and combine bits of knowledge to reason with it. Inquiry can be seen as an information exchange with nature. When the expert carries out an experiment, this corresponds to a query; the observation provides the response. Although we believe the world itself is consistent, observations may be conflicting.

The knowledge of an agent is modeled by a knowledge base KB. We say that the agent knows φ whenever KB $\vdash_L \varphi$ for some suitable base logic L, with consequence $Cn()$. In this paper we will use propositional logic as the base logic, but obviously this can be extended with more expressive logics or knowledge representation formalisms, such as description logic.

Definition 1 (Dialogue state) *A dialogue state is a tuple $DS = \langle KB, Q, H, S\rangle$, where KB is the knowledge base, Q is a prioritized queue with queries that the agent is interested in, H is a sequence of moves representing the dialogue history, and S is a set with queries that have been made, but remain unresolved.*

We represent the fact that some expert e makes a query to nature whether φ holds by an expression query(e, φ). As a response, nature either allows the observation observation(e, φ) or observation$(e, \neg\varphi)$. In case the query must remain unresolved, there is no observation.

The dialogue state of an agent changes if it poses a query or if it does an observation. Thus, the meaning of a move, such as a query or observation, is the change it makes to the dialogue states that are kept by each agent. Accordingly, we define two transition functions that map dialogue states to new dialogue states.

Definition 2 (Inquiry) *If $DS_e = \langle KB, Q, H, S\rangle$ is a dialogue state, we define a query and observation as in Fig. 3.*

Thus, if e decides to pose a query, what effectively happens is this. First, e pops the next query from the repository of queries where it is interested in, namely the queue Q. According to the priority mechanism in Q, the formula φ may be considered as the most urgent query for e. To register that e has posed φ, the expert e appends "query(e, φ)" to its personal dialogue history H.

$$\text{query}(e,\varphi)(DS_e) = \langle\ \text{KB},\ \text{pop}(\text{query}(e,\varphi), Q),\ H + \text{query}(e,\varphi)), S \cup \{\varphi\}\ \rangle$$

$$\text{observation}(e,\varphi)(DS_e) =$$
if $\text{KB} \cup \{\varphi\} \nvdash \bot$ *then*
$\langle\ Cn(\text{KB} \cup \{\varphi\}), Q,\ H + \text{observation}(e,\varphi), S \setminus \{\varphi\}\ \rangle$
else
$\langle\ \text{KB}, Q, H + \text{observation}(e,\varphi), S\ \rangle$
end

Figure 1: Definition of query and observation.

Finally, to remember that φ is asked but not answered (yet), φ is added to the list of posed but unanswered queries S.

We now explain the second equation. If e observes φ, there are two possibilities: the observation φ is consistent with what e knows, or it is not. If φ is consistent with KB, then φ is "epistemically adopted,", i.e., φ is added to KB. To register that φ has been observed, e appends "observation(e,φ)" to its personal dialogue history H. Finally, φ is crossed off as an unresolved query. If φ is inconsistent with KB, i.e. if φ contradicts e's knowledge, then e should ideally revise its knowledge, for example according to the AGM paradigm on belief revision (Gärdenfors, 1988). Since belief revision is another issue that falls beyond the scope of this paper, we have chosen for the semantically crude (but we believe technically adequate) solution that in case of inconsistencies the observation is ignored and the query remains unresolved.

The initial state of an agent DS^0 is $\langle KB^0, Q^0, \langle\rangle, \emptyset\rangle$, such that which $KB^0 \nvdash_L Q$, and $\langle\rangle$ is the empty sequence. The desired end state of the dialogue DS' is $\langle KB', \langle\rangle, H', \emptyset\rangle$. This means that all queries in Q^0 have been resolved.

The two dialogue actions can be uttered at any dialogue state, in any order. This results in a protocol that is extremely simple and rather liberal compared to other mechanized dialogue games. It is even an issue if we might speak of a true protocol here. The idea is that, in a running dialogue, experts pose questions at will, and "fish" for answers when and where appropriate, for example if time allows. In particular, there is no turn taking and observations may "come in" at any time. If the latter is put in agent-oriented terminology we may say that nature is not obliged to respond.

4 Multi-party inquiry

In addition to the protocol proposed in section 3, we make the following assumptions.

(i) A fixed number of equivalent participants engage in an inquiry dialogue.

(ii) There are no specific roles for the agents, although they may be in productive mode, or consumptive mode (see below).

(iii) Agents communicate through a central medium, called the forum, the function of which may be compared to the function of an internet newsgroup. Messages are public. They are not addressed to specific agents.

(iv) Agents act (listen, reason, and speak) in turn, for a fixed number of rounds.

(v) There is no criterion for termination., compare point(iv).

The following properties are not typical multi-party issues, but also determine the course of a dialogue.

(a) Participants are cooperative. This means that they are sincere, i.e. do not lie about their beliefs. All agents acknowledge and process all applicable messages. Moreover we assume that all agents have ample time to reason, and all agents have the opportunity to post all the messages desired.

(b) Agents have reasoning capacities. In particular, they do not ask what they already know or can infer. Before asking, an agent tries to infer the desired item itself.

(c) The facilitation of information is dialectic: claims are justified with other claims or denied with reasons that support a contradiction. Agents accept claims if and only if they can be resolved to information that they believe to be true, either on the basis of observation, or derived from acquired information.

(d) Regression to previous messages is always possible. Agents are allowed to question or justify prior claims. Thus, an immediate response is not required.

(e) For simplicities sake the agents have a shared ontology. One consequence of this assumption is that propositions (internal representations of claims) conveyed through messages do not have to be renamed.

5 Architecture

In this section we will describe the architecture that lies at the basis of our implementation.

We suppose that agents belong to a discussion group $G = \langle A, F \rangle$, where A is a (finite) set of agents A, and F is a newsgroup-like data structure called a *forum*. A forum is a sequence of entries $F = \langle m_1, ..., m_n \rangle$, where m_n is the last entry published. An entry m_i is a pair consisting of a query and a sequence of observations that count as responses:

$$m_i = \langle\ \text{query}(i, \varphi), \langle \text{observation}(r_1, \psi_1), \ldots, \text{observation}(r_1, \psi_1) \rangle\ \rangle$$

Thus, entries behave like topics or threads as found in newsgroups or mailing lists.

The internal structure of a participant $k \in A$ contains (at least) the dialogue state of section 3, along with a bookmark i to remember the first unread entry, and bookmarks (i, j) per entry for the first unread response to that query: $DS_k = \langle \text{KB}_k, Q_k, H_k, S_k, i, \langle (1, j), \ldots, (n, j) \rangle \rangle$.

Agents run concurrently, and have access to a forum that is shared by all agents. The forum is a passive asynchronous channel, but is responsible for the administration of messages. The idea is that instead of making observations, the participants will now first query the forum.

Each agent may be in *consumptive mode* or in *productive mode*. In the consumptive mode an agent takes actions that are supposed to deal with the accumulation of new knowledge: reading from the forum or posting new queries to it. This can be expressed by $\text{read}(k, F, \text{obs}(e, \varphi))$ and $\text{post}(k, F, \text{query}(k, \varphi))$. In the productive mode an agent disseminates knowledge. In our case, answering questions of other agents: $\text{post}(k, F, \text{query}(k, \varphi), \text{obs}(k, \psi))$.

6 Experiments

The multi-party inquiry set up discussed above is rather simplified. With respect to all the multi-party issues discussed in section 2, it always takes a simple solution. In order to allow experiments with different set-ups, to test if the resulting dialogues that are generated make any sense, we have made an implementation of the dialogue architecture in Ruby. This allows us to run dialogue generation experiments. The purpose of the implementation is to test different dialogue game parameter settings.

We opted for an implementation in Ruby because it is a pure object-oriented scripting language with an intuitive syntax, suited for prototyping. Fig. 2 shows the data structures of three agents, viz. Mr. Priestley, a prominent English chemist and a strong proponent of the phlogiston theory of combustion, Mr. Lavoisier, the founding father of the oxygen theory of combustion, and you, the reader, who supposedly wants to know more of combustion theory and queries the experts Lavoisier and Priestley. Other queries can be posed as well, mostly with the same effects. The resulting dialogue is displayed in Table. 1.

During our experiments, we noticed that all discussion terminate. This can be understood as follows. As a finite number of queries may be linked to a finite number of answers. Moreover, agents keep an account of which queries they have answered, so that eventually termination is ensured. We also observed that agents will reach a conclusion on accessible facts within a reasonable amount of turns. This can be explained by the fact that explanations (i.e., explanatory rules) cannot be chained infinitely. A a consequence each justification has a stopping place, so that agents will either accept facts or abandon search on explained statements with a bounded number of dialogue moves.

7 Related Research

Although it is arguably one of the simpler types of dialogue, inquiry has received less attention than negotiation or persuasion. An exception is the work by McBurney and Parsons (2001) on scientific investigation. Our purpose is very similar to

1.	Reader:	Gentlemen, how is it to be explained that in combustion, heat and light are given off?
2.	Lavoisier:	Dear Reader. You asked why in combustion, heat and light are given off. Well, that is because pure air contains oxygen, pure air contains matter of fire and heat, and in combustion, oxygen from the air combines with the burning body.
3.	Priestley:	Dear Reader. You asked why in combustion, heat and light are given off. Well, that is because combustible bodies contain phlogiston, combustible bodies contain matter of heat, and in combustion, phlogiston is given off.
4.	Reader:	Let me think. Do I know that pure air contains oxygen? ..
5.	Reader:	.. no.
6.	Reader:	Sorry, I it is not clear to me why pure air contains oxygen. Can you explain this a bit more?
7.	Reader:	Let me think. Do I know that combustible bodies contain phlogiston? ..
8.	Reader:	.. no.
9.	Reader:	Sorry, I it is not clear to me why combustible bodies contain phlogiston. Can you explain this a bit more?
10.	Reader:	Gentlemen, how is it to be explained that pure air contains oxygen?
11.	Reader:	Gentlemen, how is it to be explained that combustible bodies contain phlogiston?
12.	Lavoisier:	Well, one of the hypotheses of my theory is that pure air contains oxygen.
13.	Priestley:	Well, one of the hypotheses of my theory is that combustible bodies contain phlogiston.
14.	Reader:	Ok, thanks Lavoisier.
15.	Reader:	Ok, thanks Priestley.

Table 1: Resulting dialogue.

theirs. They describe a risk agora, as they call it, that allows the storage of multiple arguments for and against some claim. However, they do not treat multi-party issues explicitly. The agora is an asynchronous channel; no coordination rules are given.

There is also a correspondence to the Newscast protocol (Voulgaris et al., 2003). This is a kind of 'gossiping' protocol that can be used to disseminate information in distributed systems. A difference is that the newscast protocol can only pass on information. No mechanism exists to specify queries. The Newscast protocol is complementary to our work, in the sense that it may provide an implementation of the forum in distributed systems.

8 Conclusion

In this paper we proposed a simple protocol of inquiry among experts and between experts and nature. We discussed several issues that are relevant to multi-party dialogue in general: open versus closed systems, roles, type of channel, coordination and termination. We then make some choices regarding these issues, for the game of multi-party inquiry. Under some assumptions, we can show such games will terminate. However, many assumptions remain unwarranted. Therefore we hope this first attempt will stimulate more research into multi-party issues.

References

A. Bell. 1984. Language style as audience design. *Language in Society*, 13:145–204.

F. Dignum and G. Vreeswijk. 2003. Towards a testbed for multi-party dialogues. In *International workshop on Agent Communication*.

D. Traum et al. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proc. of the First Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, pages 766–773. ACM Press, July.

P. Gärdenfors. 1988. *Knowledge in Flux: Modelling the Dynamics of Epistemic states*. Bradford Books.

J. Hulstijn. 2003. Roles in dialogue. In I. Kruijff-Korbayova, editor, *Proceedings of the 7th Workshop on Formal Semantics and Pragmatics of Dialogue (Diabrck)*, pages 43–50. Universitt des Saarlandes.

P. McBurney and S. Parsons. 2001. Representing epistemic uncertainty by means of dialectical argumentation. *Annals of Mathematics and Artificial Intelligence*, 32(1):125–169.

S. Voulgaris, M. Jelasity, and M. van Steen. 2003. A robust and scalable peer-to-peer gossiping protocol. In *Proceedings 2nd International Workshop on Agents and Peer-to-Peer Computing (AP2PC 2003)*.

D. N. Walton and E.C.W. Krabbe. 1995. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY.

```
#!/sw/bin/ruby

translation_table = {

    'E1' => 'in combustion, heat and light are given off',
    'E2' => 'inflammability is transmittable from one body to another',
    'E3' => 'combustion only occurs in the presence of pure air',
    'E4' => 'increase in weight of a [.. snip ..] weight of air absorbed',
    'E5' => 'metals undergo calcination',
    'E6' => 'in calcination, bodies increase weight',
    'E7' => 'in calcination, volume of air diminishes',
    'E8' => 'in reduction, effervescence appears',
   'OH1' => 'pure air contains oxygen',
   'OH2' => 'pure air contains matter of fire and heat',
   'OH3' => 'in combustion, oxygen from the air combines with the burning body',
   'OH4' => 'oxygen has weight',
   'OH5' => 'in calcination, metals add oxygen to become calxes',
   'OH6' => 'in reduction, oxygen is given off',
   'PH1' => 'combustible bodies contain phlogiston',
   'PH2' => 'combustible bodies contain matter of heat',
   'PH3' => 'in combustion, phlogiston is given off',
   'PH4' => 'phlogiston can pass from one body to another',
   'PH5' => 'metals contain phlogiston',
   'PH6' => 'in calcination, phlogiston is given off'

}


Agent.new(
   'name'      => 'Priestley',
   'questions' => {},
   'knowledge' => {
      'E1' => [ %w(PH1 PH2 PH3) ],
      'E2' => [ %w(PH1 PH3 PH4) ],
      'E5' => [ %w(PH5 PH6) ],
      'PH1' => TRUE,  'PH2' => TRUE,
      'PH3' => TRUE,  'PH4' => TRUE,
      'PH5' => TRUE,  'PH6' => TRUE
   }
)

Agent.new(
   'name'      => 'Lavoisier',
   'questions' => {},
   'knowledge' => {
      'E1' => [ %w(OH1 OH2 OH3) ],
      'E3' => [ %w(OH1 OH3) ],
      'E4' => [ %w(OH1 OH3 OH4) ],
      'E5' => [ %w(OH1 OH5) ],
      'E6' => [ %w(OH1 OH4 OH5) ],
      'E7' => [ %w(OH1 OH5) ],
      'E8' => [ %w(OH1 OH6) ],
      'OH1' => TRUE,  'OH2' => TRUE,
      'OH3' => TRUE,  'OH4' => TRUE,
      'OH5' => TRUE,  'OH6' => TRUE
   }
)

Agent.new(
   'name'      => 'Reader',
   'questions' => {
      'E1' => TRUE
   },
   'knowledge' => {}
)
```

Figure 2: Translation table, followed by data structures for three agents.

Presupposition and Belief in DRT: Towards A New Implementation

Yafa Al-Raheb
University of East Anglia
y.al-raheb@uea.ac.uk

1 Introduction

This poster is part of current research investigating presupposition and belief in human dialogues using Dynamic Interpretation Theory (DIT) to categorize dialogue utterances within the framework of Discourse Representation Theory (DRT). The work in progress aims at making dialogue representation within DRT more pragmatic, especially in relation to presupposition.

The developing implementation builds on Bos and Blackburn's Curt DRT program [1], and Bos's DORIS program [2] to include more examples of presupposition, augment DRT with DIT's dialogue acts, and to represent the beliefs of the participants to the dialogue [3].

2 Presupposition, Assertion, and Belief

As in dynamic semantics, presupposition is viewed here as anaphoric, lexically triggered and dependent on context (van der Sandt 1992). Examples of presupposition include:
(1) a. Speaker: **The** red book is interesting.
b. Speaker: Vincent likes **her** dress.
c. Speaker: **Mia** loves Vincent.

To make presupposition within DRT more pragmatic, presupposition is understood as being the property of the speaker. In this sense, the presupposition being 'taken for granted' means: the speaker believes the presupposition to be known or given information and not the focus or centre of her utterance. For example,
(2) Speaker: My car just broke down.

'my car' constitutes the given information that the speaker has a car, while 'just broke down' provides new information; the information that speaker is attempting to communicate is called assertion.

Presupposition is related to the beliefs of the speaker, regardless of whether the beliefs are part of the 'common ground' or not. Speaker belief leads to presupposition, which conveys the beliefs of the speaker to the hearer. This approach takes a stronger position to beliefs' relation to presupposition than Geurts (1999) by assuming that Grice's Cooperative principle is in place. Consider Stalnaker's example (2002):
(3) I have to pick up my sister at the airport.

If we were to assume that the participants in the dialogue are being cooperative, not lying, being relevant, etc, we can take the stronger position that the information introduced by the presupposition, here 'having a sister', is indeed a belief held by the speaker. This applies to whether 'having a sister' is known to the hearer or not.

The point to be made here is that the relationship between belief and presupposition, and belief and assertion helps clarify what is meant by presupposition. Additionally, introducing speaker and hearer perspectives contributes to the clarification of presupposition. Let us refer to presupposition by P, to assertion by A, to believe by bel, speaker by S, and hearer by H.

[1] www.comsem.org
[2] www.coli.uni-sb.de/ bos/doris/
[3] I would like to thank Dr Johan Bos for kindly sending me the code for DORIS and for his advice.

(4) Speaker: Vincent's wife likes chocolate.
Hearer: I thought she was allergic to it.

In the first part of examples (4), P is 'Vincent has a wife and Vincent is male', whereas A is 'she likes chocolate'. On the assumptions given above, the hearer can correctly come to the result that bel(S,P).

Belief places some constraints on assertion. 'Beliefs Constraint on Assertion I' is a constraint placed by beliefs on uttering A, that bel(S,¬bel(H,A)). Another constraint beliefs place on A, is called 'Beliefs Constraint on Assertion II': being cooperative, to utter A, S must bel that A, bel(S,A). Assuming the cooperative principle, belief also places a constraint on P, 'Beliefs Constraint on Presupposition': to utter P, bel(S,P).

The following is the representation of the requirements and consequences of the first part in a mini dialogue. From S's perspective, before uttering A and P: bel(S,P) ('Beliefs Constraint on Presupposition'), bel(S,A) ('Beliefs Constraint on Assertion II'), and bel(S,¬bel(H,A)) ('Beliefs Constraint on Assertion I'). If bel(S,bel(H,P)), S expects H to take P for granted. If bel(S,¬bel(H,P)), S expects H to accommodate P if P is unremarkable (Geurts 1999).

From H's perspective, for H to receive P, the new belief bel(H, bel(S,P)) is formed. If ¬bel(H,P) and P is unremarkable, accommodate(H,P). Accommodate can either mean accept(H,P), or bel(H,P). If bel(H,¬ P), reject(H,P). If bel(H,P), H takes P for granted. For H to receive A, the new beliefs bel(H, bel(S,A)) and bel(H,bel(S,¬ bel(H,A))) are formed. There are three options, accept(H,A), reject(H,A), or bel(H,A). Accept means put on hold, not yet believed, but not rejected. H has to provide feedback according to choice made.

3 Augmenting DRT with DIT's Dialogue Acts

DRT supports the idea that a description of dialogue has to represent mental states and their relation to the context. To represent beliefs, it is necessary to have a representation of dialogue acts in order to gain an insight into the cognitive states of both the speaker and the hearer (Asher 1986). Our implementation uses DIT's dialogue acts in order to shed light on the beliefs of the participants in a dialogue.

The use of dialogue acts in relation to belief, presupposition, and assertion is most relevant in the case of feedback. Generally speaking when S says something to H, H provides positive or negative feedback. 5.a represents weak positive feedback indicating A is received, accept(H,A), whereas 5.b represents strong positive feedback, where H indicates reception of A, and that bel(H,A). Rejecting A is a way of giving negative feedback, reject(H,A), 5.c.
(5) Speaker: Jody loves Butch.
a. Hearer: aha.
b. Hearer: I couldn't agree more!
c. Hearer: No, Jody is married to Vincent!

4 Towards a New Implementation of Belief and Presupposition

Current work on implementation involves incorporation of belief spaces, presupposition/ assertion distinction marked in Prolog, and relating presupposition to belief. Separating presupposition from assertion results in two stages with each new utterance, the presupposition stage and the assertion stage. The former represents the presupposition, relates it to beliefs and then applies it to context. The assertion part then gets represented, related to beliefs and then merged with the resulting context, context{P}. Future work will involve working on using the same method in representing more than one presupposition. An algorithm is currently under development for merging strategies that will represent both the speaker's and the hearer's beliefs in the main DRS.

References

Asher, N. 1986. 'Belief in Discourse Representation Theory'. *Journal of Philosophical Logic*, 15:127–189.

Geurts, B. 1999. 'Presuppositions and Pronouns: Current Research in the Semantics/ Pragmatics Interface'. Oxford: Elsevier

Stalnaker, R. 2002. 'Common ground'. *Linguistics and Philosophy*, 25(5-6):701-721.

van der Sandt, R.A. 1992. 'Presupposition Projection as Anaphora Resolution'. *Journal of Semantics*, 9:333-377.

Shared Scoreboards and Common Information

Anton Benz
Institut for Fogsprog, Kommunikation og Informationsvidenskab
Syddansk Universitet (Kolding)
benz@sitkom.sdu.dk

Abstract

Using a picture by Lewis, we call information structures that explicitly represent the dialogue participant's shared information '*scoreboards*'. The central claim is that we have to keep the aspect of common ground as a shared scoreboard that determines interpretations of dialogue contributions distinct from its aspect as mutually available information. We look at the following two conditions: If the shared scoreboard were a representation for common information, then (1) every dialogue participant would be licensed to add any proposition as soon as he/she has a proof that it is shared information. On the other side, (2) an interlocutor is not allowed to add a proposition as long as he/she does not know that it is also shared. We give arguments that both conditions can be violated.

1 Shared Information

David Lewis (1979) introduced the picture of a *scoreboard* to illustrate the role of the common ground in conversation. We can think of it as a shared board where all public activities and utterances are written down and thereby become *shared* facts about the dialogue. It provides the background against which to interpret new sentences, and forms the basis for expectations about behaviour of conversational partners.

Of course, a public scoreboard does not exist, and the common ground is generally identified with the proposition that represents the totality of information shared by dialogue participants. We distinguish two ways in which this information can be *shared*: As an explicit information structure or as the implicit common information. Depending on how we represent information, we may think of the common ground as a set of sentences or a set of possible worlds[1], where the first way is more in line with the scoreboard picture, and the second with the view of the common ground as implicit common information.

Both aspects play a different role in interaction: If you and I read a newspaper article about Nicole Kidman, and one of us refers to her with the description '*the actress*,' then we both have to *know* about each other what we have read in order to be sure that we both interpret the description in the same way. This is different from cases where coordination works without explicit representation of knowledge of others: The fact that all drivers drive on the right side of the street guarantees that no crashes occur. They have common information that they succeed in this aim but they don't need to have explicit *knowledge* about this, or about each other.

A need for an explicit representation of shared information is obvious for implemented dialogue systems. The standard definition for common knowledge reads as follows: a proposition φ is *common knowledge* for two interlocutors A and B

[1](Fagin et al., 1995; Meyer & v. d. Hoek, 1995; Hintikka, 1962)

iff φ is true, if A and B know that φ is true, if A and B know that A and B know that φ is true, if A and B know that A and B know that A and B know that φ is true, etc... If we read here *know that* as *has information that*, then we arrive at the intended definition of *common information*.

2 Coordination of Interpretation

The differences between explicit and implicit representations are mainly discussed in the literature with respect to *fine grainedness* and the problem of *logical omniscience*[2]. We concentrate here on their different roles in coordinating interpretation. One of the central tasks of dialogue participants is to make sure that they both interpret dialogue contributions in the same way, or else misunderstandings will arise. The aspect of coordination becomes especially prominent in a dialogue theory as that of H.H. Clark (1996). He analysed dialogue predominantly in terms of joint projects, i.e. every contribution of the speaker is seen as part of an activity where he and the addressee must work together towards a joint goal. If interlocutors want to be sure that they have success, then they need common information that they coordinate their activities in the right way; but they don't need necessarily explicit representations of this information.

If the interpretations of utterances and updates of the common ground depend in a non–trivial way on private information, then the interlocutors may end up with different interpretations, and hence fail in their coordination task. This motivates a strong restriction on interpretations:

(SP) The interpretation $[\varphi]$ of a phrase φ is totally determined by the phrase itself and the common scoreboards.

We call this the *scoreboard principle*. If we see scoreboards as representations *for* common information, then the following two principles should hold:

(1) If a interlocutor knows that it is common information that a sentence φ is true, then he has to add φ to his scoreboard.

(2) No dialogue participant is allowed to add a sentence φ at time t to his scoreboard unless there is common information that every participant knows that φ at time t.

We discuss a number of examples that show that the scoreboard principle (SP) can get into conflict with (1) and (2). The Muddy Children[3] example shows that there are situations where (1) together with (SP) leads to miscommunication; i.e. there is a context in this example where there exists a sentence φ such that even if a participant can prove that it is common information that φ is true, he is not allowed to add φ to his scoreboard. This shows that the participants scoreboards may be *less* informative than common information. We discuss the Time Imprecision Problem[4] in order to show that the participants scoreboards may also contain *more* information; i.e. there is a sentence φ that a dialogue participant has to add to his scoreboard at a time t although it is not common knowledge that φ at t. This violates principle (2).

Together these examples show that the content of shared scoreboards cannot be defined as a representation for common information. As all interlocutors have to maintain the same representations in order to coordinate their language use, the maintenance of identical scoreboards becomes itself a coordination task.

References

H.H. Clark. 1996. *Using Language;* Cambridge.

R. Fagin, J.Y. Halpern, Y. Moses, M.Y. Vardi. 1995. *Reasoning About Knowledge;* Cambridge, Massachusetts.

R. Fagin, J.Y. Halpern, Y. Moses, M.Y. Vardi. 1999. *Common Knowledge Revisited;* Annals of Pure and Applied Logic 96, pp. 89–105.

J. Hintikka. 1962. *Knowledge and Belief;* Ithaka, N.Y..

D. Lewis. 1979. *Scoreceeping in a Language Game;* In: R. Bäuerle, U. Egli, A. von Stechow (eds.): *Semantics from Different Points of View*, Springer, Berlin etc., pp. 172–187.

J.–J. CH. Meyer, W. van der Hoek. 1995. *Epistemic Logic for AI and Computer Science;* Cambridge University Press, Cambridge.

[2] See (Fagin et al., 1995, Ch. 7) and (Meyer & v. d. Hoek, 1995, Ch. 2) for a discussion of various notions of knowledge and belief in modal approaches and the problem of logical omniscience.

[3] See e.g. (Fagin et al., 1995, p. 4).

[4] See (Fagin et al., 1995), Chapter 11, Section 11.2, and (Fagin et al., 1999).

On some effects of lexical contrast in information-oriented dialogue

Francesca Carota
Department of Linguistics
University of Pisa
francesca.carota@ilc.cnr.it

1 Introduction

Although CMs are pervasive in information-oriented dialogues in Italian, their roles still need to be interpreted within a unified framework. A preliminary corpus study of the CMs of Italian *ma* and *invece* within the travel domain shows that their behaviour can be sensitive to the dialogue structure modelled in terms of topic and common ground units (CGU) (Traum, 1999) and can depend on information management and grounding. Such notions refer, respectively, to the negotiation of information -the dialogue content to be grounded as relevant to a current task- and to the coordination activity by which participants achieve *common ground* (CG), or a common mental state of *agreement*[1] about the negotiated information, which is the cognitive context where the negotiation takes place. The topic is intended as being the current *question under discussion* (QUD) (Ginzbourg, 1998) and hierarchically organized in global QUDs (GQUD), containing main topics and local QUDs (LQUD), containing subtopics. We assume that, in information-oriented dialogues, the hierarchically lowest-level subtopic can be the discourse entity of a LQUD discussed by participants as an alternative solution to an issue under negotiation (Larsson, 2001), and that topical entities, given or new, are linked to the cognitive context in terms of activation degrees, i.e. active-semiactive-inactive (Chafe, 1994), in the speakers' CG. The topical entities packaged in an utterance *theme* are seen as local cotextual instantiations keeping track of the status of the dialogue topic in the speakers' CG. If to be grounded, they can be brought to the interlocutor's attention through prosodic focus. Following these premises, CMs are associated to topic status in the speakers' CG and to grounding phases such as presentation, whereby a topic is introduced to the CG, and acceptance, whereby a feedback is provided to evidence whether the presented topic has accessed the CG. The positive feedback for established CG is the *acknowledgement* dialogue act (Traum, 1999).

[1] We intend the grounding process as consisting of the coordination and alignment of dialogue management at several levels, towards an quasi-shared mental state, rather than as implicating the philosophical notion of mutual understanding.

3 Corpus study and discussion

We analyzed occurrences of the above mentioned CMs through 150 dialogues, in which *ma* and *invece* respectively exhibit the discourse functions discussed here. In the representative example (1), the contrast cannot be modelled as concessive due to the lack of a *tertium comparationis*, or a claim for which a positive and a negative argument are provided in the context. Furthermore, the CM is neither a topic shifter nor a turn-taking device, used, according to Bazzanella (1991), to interrupt the interlocutor's turn.

(1) **GQUD:** *Il RITORNO, cosa voleva?*
The return, what did you need?
C.1: *Il ritorno, le avevo detto domenica, domenica pomeriggio, sul presto.*
The return, I said on Monday, Monday early in the afternoon.
LQUD: *Si*. ***Ma*** *come ORARIO*?
Yes. *But* as for the HOUR?

In the exemplified context, *ma* occurs after the acknowledgment *si*, which positively feedbacks the acceptance in the CG of the information distributed along with the topical hierarchically organized *continuum*GQUD1:"return">LQUD1:"Monday"> LQUD2:"afternoon"> LQUD3:"early".
Ma intervenes at the step LQUD3 and is positioned at the beginning of an elliptical open question about the entity in focus *orario*, to be co-activated in the CG because relevant for the current negotiation. The contextually given information in S.1 needs to be revised/clarified, because a missing informative parameter has to be supplied at the level of the LQUD3 related to the hour.

Consequently, the CM partly contradicts the previous *si* and blocks the acceptance phase by signalling that the CG has not been fully reached. By playing the reorientation function of redirecting the interlocutor's attention towards a more specific topic to ground, the CM cues the initial boundary of a sub-CGU, in which a new presentation to the CG explicitly requires the clarification of some inactive information about the missing informative parameter. Though still placed at a beginning of an elliptical open question representing a LQUD, the CM *invece* in (2) conveys a different kind of contrast. The host LQUD asks about new information as regards the given and supposedly semiactive topical entity in focus *Meridiana*, which has to be co-activated in the CG in alternative to the contextually given entity *Alpi Eagles*. A prototypical issue under negotiation is profiled, where the first alternative has been explicitly grounded so far by the acknowledgment *va bene*. The polar alternative is placed at the same level of the topical hierarchy within the current topic *flights*.

(2) ***GQUD****: Vediamo un po' quali sono i voli.*
Let's have a look at the flights.

O.1.: *Si`. Allora, prima di mezzogiorno, con l' Alpi Eagles, ci sarebbe un volo che parte da Roma alle dieci mattina e arriva a Verona alle undici.*
Yes. So, before midday with Alpi Eagles, there is a flight which leaves from Roma at 10 a.m. and reaches Verona at 11 a..m.

C.2 : *ah va bene.*
Ah, OK.

LQUD: ***Invece*** *con la MERIDIANA*?
Instead with the MERIDIANA?

The CM establishes an anaphoric link to the CGU in which the first alternative to the current issue has been discussed, by providing an instruction *instead of X->Y* for the interlocutor to update her information state by replacing the contextually given alternative X=*Alpi Eagles*with the entity Y=*Meridiana*. *Invece* contributes to a speaker's change-of-perspective local strategy of negotiating alternative solutions to a same issue. In (3), *invece* is involved in a more global topic-change strategy.

(3) **GQUD**: *Avrei bisogno di alcune informazioni sui treni Roma-Verona per partire venerdì 15 settembre* e *rientrare domenica 21.*
I need some information about the Roma-Verona trains, to leave on Friday the 19th of September at 8 and to come back on Sunday the 21th.

1 O : *Venerdì 19, c'è un treno alle 8.35.*
The 19th, there is a train at 8.35 a.m.

2 C : *Benissimo.* Perfect.

3 O : *OK. Per il RITORNO* ***invece****, c' e` un treno alle 16 da Roma. Puo` andar bene?*
OK. As for the return *instead*, there is a train from Roma at 4 p.m. Is it Ok?

4 C : *Alle 16, potrebbe andar bene.*
At 4 p.m., it should be OK.

After the acknowledgement *Ok*, which grounds the topic *departure* of the sub-CGU 1O-2C, the CM falls, significantly, in the left-detached contrastive theme *per il ritorno*: it instructs to substitute the grounded topical coordinate with the one indicated in the host theme (*instead of X->Y*). The CM opens a new CGU and inaugurates the negotiation of a new topic. It plays a meta-cognitive function at the ideational level of dialogue structure, while interacting with grounding process.

We have proposed to uniformly interpret some CMs in information-oriented dialogues according to different kinds of contextual polarities due to the interplay between information management and grounding. Further work intends to account for other CMs, by studying their interplay withother kinds of contrastiveness and providing statistical analysis of their distribution in corpora of different domains.

References

C. Bazzanella. 1991. I segnali discorsivi. In L. Renzi, G. Salvi, A. Cardinaletti (eds.), Grande Grammatica Italiana di Consultazione. Il Mulino, Bologna.

W. Chafe. 1994. Discourse, Consciousness, and Time. Chicago/London: University of Chicago Press.

J. Ginzburg. 1998. Clarifying Utterances. In Proceedings of the 2nd Workshop on the Formal Semantics and Pragmatics of Dialogue, Twente.

S. Larsson. 2002. . In Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue, Philadephia, pp. 103-112.

D. Traum. 1999. Computational Models of Grounding in Collaborative Systems. In *Working notes of AAAI Fall Symposium on Psychological Models of Communication*, pp. 124-131.

C. Soria, R. Cattoni, M. Danieli. 2000. ADAM: An architecture for XML-based Dialogue Annotation on Multiple Levels. In *Proceedings of the First SIGDIAL Workshop on Discourse and Dialogue*, Hong-Kong, pp. 9-18.

Developing a Typology of Dialogue Acts: Question – Answer Adjacency Pairs in Estonian Information Dialogues

Olga Gerassimenko	Tiit Hennoste	Mare Koit	Andriela Rääbis	Maret Valdisoo
Department of Linguistics University of Tartu	Department of Linguistics University of Tartu	Institute of Computer Science University of Tartu	Department of Linguistics University of Tartu	Institute of Computer Science University of Tartu
gerro@ut.ee	hennoste@ut.ee	koit@ut.ee	andriela@ut.ee	maret@ut.ee

Abstract

Estonian dialogue corpus includes 320 spoken dialogues. We have worked out a typology of dialogue acts and are using it for annotating of the corpus. In this paper, we give an overview of the typology. The second part of the paper is based on the analysis of information dialogues. Most frequent question and answer types and typical sequences of questions and answers are found out with the purpose to model questioning – answering strategies in a dialogue system.

1 Introduction

The Estonian Dialogue Corpus (EDiC) includes 320 spoken dialogues, among them 205 calls and 115 face-to-face conversations, with total length of 80 000 running words.

We have worked out a typology of dialogue acts and use it for annotating our corpus (Hennoste et al., 2003). Our goal is to develop a dialogue system that will be able to interact with a user in Estonian and provide him/her some information, following norms and rules of human-human communication. This is the reason why we are studying human-human spoken dialogues. For this paper, we have chosen 101 information dialogues (calls for information, to travel bureaus, shops and outpatients' departments), and analyze the question – answer adjacency pairs (APs), the most important dialogue acts in information dialogues.

2 The EDiC Typology of Dialogue Acts

Our typology departs from the point of view of conversation analysis (CA) that focuses on the techniques used by people when they are actually engaged in social interaction (Hutchby and Wooffitt, 1998). The main idea behind the analysis is that conversation is the collaboration of participants based on three mechanisms: turn taking, repair, and APs. An advantage of this approach is that CA departs from empirical data, i.e. it tries to find out explicit markers in the text that allow to determine utterance functions.

Based on the principles of CA we get the following main typology of dialogue acts.

1. Adjacency pair (AP) acts

1.1. Dialogue managing acts

Fluent conversation

1) Conventional (greeting, thanking, etc.)

2) Topic change

Solving communication problems

3) Other-initiated self-repair

4) Contact control

1.2. Information acts

5) Directives (request, proposal, giving information, etc.)

6) Questions and answers

7) Opinions

2. Non-AP acts, or single acts

2.1. Dialogue managing acts

Fluent conversation

8) Conventional (contact, recognition, etc.)

9) Responses (continuer, acknowledgement)
Solving communication problems
10) Self-repair
2.2. Information acts
11) Primary single acts (advance note, promise, etc.)
12) Additional information (specification, explication, etc.

The total number of dialogue acts is 126 in our typology. Act tokens are originally in Estonian.

3 Questions and Answers

There are three question types that depend on the expected reaction:
- questions that expect giving information: wh-question, open yes/no question
- questions that expect agreement/refusal: closed yes/no question, question that offers answer
- questions that expect the choice of an alternative: alternative question.

Open and closed yes/no question have similar form but they expect different reactions from the answerer (e.g. *Are you open in winter?* expects the answer *yes* or *no*, but by asking *Is there a bus that arrives in Tallinn after 8?* the questioner wants to know the departure times of buses). Open yes/no question is actually an indirect speech act – a request or wh-question that is expressed in form of yes/no question.

Our analyzed dialogues include 649 question tags: 233 wh-questions, 177 questions offering answer, 111 open and 81 closed yes/no questions, 27 alternative questions. The remaining 20 questions belong to the sub-type 'other'.

Different question types are used differently by participants. Most of the questions were asked by the client: 90% of open yes/no questions, 84% of closed yes/no questions, 77% of wh-questions, 66% of questions offering answer, and 52% of alternative questions. Wh-questions, open and closed yes/no questions are mostly used for topic initiation or continuation (74%, 92% and 73% of cases, respectively). Most of questions offering answer (60%) initiate repairs.

A typical information dialogue includes three parts: the conventional beginning, main information part, and conventional ending. The kernel of the information part is a question – answer adjacency pair: a question is asked and an answer is got. We have found three typical questioning – answering strategies in our dialogues.

Strategy 1. Client asks a question and gets a desirable answer. Two sub-types can be differentiated.

a) Client asks a wh-question or open yes/no question and gets the requested information (cf. Example), or (s)he asks an alternative question and gets one alternative as answer, or (s)he asks a closed yes/no question and gets answer *yes* or agreeing *no*.

Example (CA transcription used):

```
Client: (.) ei tea mis kellast
doktor Laane vastu võtab.  WH-
QUESTION
what is doctor Laane's reception
time
Officer: e kella neljateistkümnest
seitsmeteistkümneni.=  GIVING
INFORMATION
from two to five p.m.
```

b) Client asks a wh-question or open yes/no question and gets the asked information, like in the previous case, but after that (s)he initiates a repair. The typical repair initiation is repeating a phone number.

Strategy 2. Client gets an undesirable answer (information is missing). Such cases are seldom in our analyzed dialogues, it is difficult to find out a preferred strategy.

Strategy 3. The officer initiates an inserted sequence before answering (a repair or a question adjusting the conditions of the answer).

4 Further work

Our further work will concentrate on finding out of more communicative strategies and on formal definitions of dialogue acts that make it possible automatic recognition of user's goals in a co-operative dialogue system.

References

Tiit Hennoste, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo, and Evely Vutt. 2003. Directives in Estonian Information Dialogues. *Text, Speech and Dialogue. 6th International Conference TSD 2003.* Ed. V. Matousek, P. Mautner. Springer, 406-411.

Ian Hutchby and Wooffitt, Robin. 1998. *Conversation Analysis. Principles, Practices and Applications.* Polity Press.

Managing Uncertainty in Dialogue Information State for Real Time Understanding of Multi-Human Meeting Dialogue

Alexander Gruenstein, Lawrence Cavedon, John Niekrasz, Dominic Widdows, and Stanley Peters
Center for the Study of Language and Information
Stanford University, Stanford, CA 94305
{alexgru,lcavedon,niekrasz,dwiddows,peters}@csli.stanford.edu

1 Introduction

Human speech processing is riddled with ambiguity and uncertainty on a number of levels: *e.g.* uncertainty of speech-processing; lexical and structural ambiguity in parsing; dialogue–act classification; intention recognition and interpretation. Information-state approaches to dialogue management typically only maintain a single current state and utilize strategies for resolving ambiguities and uncertainty immediately they arise.[1]

We are concerned with tracking and understanding dialogue between multiple human participants—specifically, in meetings—in such a way that the dialogue system does not intervene. In this scenario, the system is not able to provide feedback on whether or not it has understood, and is unable to ask for clarification or ambiguity resolution. Our ultimate aim is to model human-human dialogue (to the extent that it is feasible) in real-time, providing useful services (*e.g.* relevant document retrieval) and answering queries about the dialogue state and history (*e.g.* "what action items do we have so far?"). Our approach has been to extend our existing dialogue system, based on the information-state update approach—which supports a rich semantic interpretation of multi-utterance constructions—to cope with the added uncertainty inherent in two-person meetings in which the participants speak, point, and draw on a whiteboard.

[1] Some previous work has considered the issue of dialogue management under uncertainty (e.g. (Levin et al., 2000; Roy et al., 2000)) but has not generally involved rich semantic dialogue states, linking speech directly to action.

1.1 Meeting artifacts and information state

We focus exclusively on meetings about *artifacts*: *i.e.* meetings that produce some constructed object as its end, such as a project plan with tasks and deadlines (*i.e.* a Gantt chart), or budget in some sort of spreadsheet format. This focus provides a concrete frame for interpretation of drawing and of spoken language.

Artifacts are represented in an ontology designed using Protégé,[2] including classes for the objects themselves (*e.g.* a plan and its components), relations among these entities, and the events which affect state-change in the entities or relations. The current artifact state, as represented by the ontology, is part of the information state of the dialogue and contributes to the interpretation of utterances. Indeed, most utterance sequences in our scenario can be viewed to have semantics defined in *operations* over the artifact under discussion.

A meeting history viewer graphically displays the relationships between changes to the artifacts in the information state and the utterances and actions which caused those changes. This provides a useful visual into the internals of the system, and comprises a tool by which a meeting can be indexed, allowing a user to skip to the dialogue segment associated with additions or changes to the artifact (*e.g.* revisiting the negotiation associated with the choice of a milestone date). Unlike standard meeting summarization systems, the history viewer is cross-indexed by both artifact and dialogue.

[2] protege.stanford.edu

2 Uncertainty management

There has been much work on dialogue management systems to detect and resolve ambiguity, such as by combining multiple sources of evidence—*e.g.* multimodal systems that combine speech and drawing/gesture (Oviatt, 2000), or systems that use prosodic features to help classify speech-acts (Venkataraman et al., 2003)—or by using corpus-based statistical techniques to identify most likely interpretation. However, little work has been done on *maintaining* the uncertainty that arises from such ambiguity over extended periods of time, rather than resolving it soon after its detection.

Previous applications of our dialogue management system—e.g. (Lemon et al., 2002))—have ignored uncertainty in interpretation and have resolved ambiguity immediately as it arose: *e.g.* only the top item of the speech-recognizer's n-best list was considered (regardless of probability), and clarification questions were used to resolve ambiguity. However, in the meeting-understanding application, uncertainty management becomes necessary as the system has only limited mechanisms for resolving detected ambiguities without intruding on the normal flow of the meeting.

2.1 Incorporating ASR uncertainty into dialogue state

An initial implementation of uncertainty in our dialogue state framework is to incorporate multiple results from an n-best list into the Dialogue Move Tree (DMT). As in previous work, each incoming utterance is classified as a type of dialogue move, and a corresponding node is attached to the DMT using an attachment algorithm (see (Lemon et al., 2002)). Here, however, all speech-rec results which can be interpreted in context are simultaneously attached to the dialogue move tree—these assignments are weighted depending on recognizer and dialogue-move classification confidences. As more evidence becomes available, either through subsequent utterances or through multimodal evidence,[3] nodes which represent unlikely interpretations are pruned from the DMT. The idea is that the tree may contain arbitrarily long threads representing competing interpretations of conversations which will be pruned as new evidence rules out unlikely threads.

[3]Multimodal integration is performed in collaboration with the Center for Human-Computer Communication at Oregon Graduate Institute; see (Kaiser et al., 2003)

2.2 Current and future directions

Many meetings have at least an outline of structure, such as a formal or pre-agreed agenda. Some agenda items may be directly related to the meeting artifact or component thereof (*e.g.* deciding the delivery date of a task). A direction we are currently exploring, one which does not seem to have been pursued in previous meeting-understanding projects, is to include some representation of meeting-state—as measured by progress against an agenda—to the dialogue information state. We are also investigating techniques for automatically detecting topic shifts. Such information can be added to dialogue-state and used to prime ASR language-models and disambiguate spoken utterances. Links from utterance to agenda-item or topic are themselves highly uncertain of course, and will require more sophisticated probabilistic models to be incorporated into the dialogue management process.

References

E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. 2003. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In *ICMI 2003*.

O. Lemon, A. Gruenstein, and S. Peters. 2002. Collaborative activities and multi-tasking in dialogue systems. *Traitment automatique des langues*, 43(2).

E. Levin, R. Pieraccini, and W. Eckert. 2000. A stochastic model of human-machine interaction for learning dialogue strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1).

S. L. Oviatt. 2000. Taming speech recognition errors within a multimodal interface. *CACM*, 43(9).

N. Roy, J. Pineau, and S. Thrun. 2000. Spoken dialog management for robots. In *ACL 2000*, Hong Kong.

A. Venkataraman, L. Ferrer, A. Stolcke, and E. Shriberg. 2003. Training a prosody-based dialog act tagger from unlabeled data. In *IEEE-ICASSP*, Hong Kong.

Using Discourse Structure in a Dialogue System to Search in Databases

Christian Hying
IMS, Universität Stuttgart
Azenbergstr. 12
70174 Stuttgart
Germany
christian.hying@ims.uni-stuttgart.de

Sunna Torge
ASL, Sony International (Europe) GmbH
Hedelfinger Strasse 61
70327 Stuttgart
Germany
torge@sony.de

Introduction We present the functionality of a *discourse processing component* (DPC) for dialogue systems that are applied to the task of browsing a database.[1] The DPC is implemented according to (Grosz and Sidner, 1986). It contains a *focus stack* which keeps information about the *intentions* and the linguistically relevant objects (*discourse objects*) which occur in the course of the dialogue. The intentions control the focus stack. They are computed by employing a simple semantics of utterances: utterances are mapped onto intentions to specify a database query. We identify the underlying *discourse purpose* with the goal of picking a single database item. This is subsumed by the database query specified. Thus, computing the relation between utterance intentions and discourse purposes boils down to comparing database queries. The focus stack is used to build a salience structure which contains *discourse objects*. These *discourse objects* serve as possible antecedents for anaphoric expressions. For each *discourse object* the salience structure holds information about salience, surface form and meaning in order to support an anaphora resolution component.

In order to show the applicability in user directed dialogue we have chosen an ill-structured task, cf. (Bernsen and Dybkjær, 2000), namely picking a song from a music database. There is no natural order in which the attributes *title*, *artist*, and *genre* have to be specified. We assume that the system has a small text display and the ability to produce spoken output. The user can provide input to the system only by way of spoken input.

Interaction The interaction between user and system is predetermined by the following interaction pattern: first, the user specifies a database query, and second, the system offers the user options to refine that database query. Note, that the latter also comprises the offering of single database items. Since our approach heavily relies on discourse processing, a closer look at possible user input shows, that generally speaking there are two possibilities: (i) The user specifies a new database query which does not relate to any previous material. Examples are shown in Figure 1 in utterances (1) and (5b). And (ii), the user can take up one of the options offered by the system by using an anaphoric expression such as a definite description, a name, an abbreviation of a name, or a pronoun. In Figure 1 utterance (3), the abbreviation "Folk" is anaphorical on the option named "Irish Folk". Similarly (5a) is anaphoric, too.

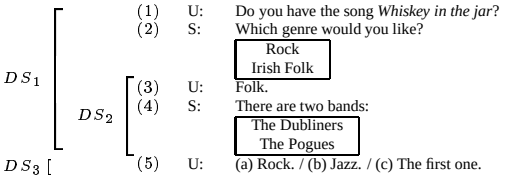


Figure 1: Example dialogue with discourse structure

Computation of Discourse Structure The setting of the task is such that we only need Grosz

[1] The work was done as a diploma thesis for the University of Stuttgart at Sony International Stuttgart. We would like to thank Ulrich Heid and Jan van Kuppevelt for their support.

and Sidner's *dominance* relation. We compute it by establishing a subsumption relation between database queries. A database query is represented by a set of attribute-value pairs where the attribute specifies a field of the database and the value specifies the value of the field.

(1) Let A and B be database queries. A *subsumes* B, iff $A \subset B$.

That means that A subsumes B, if and only if any attribute-value pair that is element of A is also element of B and B contains at least one pair that is not element of A.

The specification of database queries relates to the structure of discourse in the following way: each discourse segment is assigned exactly one database base query which characterises its discourse purpose. A discourse segment starts with the specification of a database query and comprises all successive utterances which do not specify another query that is *not* subsumed by it. A discourse segment embeds another discourse segment if the database query that is associated with it subsumes the query of the other segment. In the example DS_1 is associated with the database query ⟨title, "whisky in the jar"⟩ and DS_2 with {⟨title, "whiskey in the jar"⟩, ⟨genre, "irish folk"⟩}. If the user uttered (5a), DS_3 would be associated with ⟨genre, "jazz"⟩ and accommodated on top level. If he uttered (5b), DS_3 would be associated with the query {⟨title, "whiskey in the jar"⟩, ⟨genre, "rock"⟩} and embedded under DS_1. Finally, (5c) would yield DS_3 being embedded under DS_2.

Discourse Processing The DPC updates the focus stack with every utterance, so that the stack holds the information which is in the focus of attention at each point of the dialogue. The elements of the focus stack are *focus spaces*. In our implementation they are realized as feature structures of the type shown in Figure 2.

A focus space representation contains of three features: the feature PURPOSE holds the discourse purpose of the associated discourse segment in form of a database query. Each of the other two features, i.e. DISC and GRAPH, holds a list of representations of *discourse objects*. The first list (DISC) contains representations of *discourse objects* which have occurred in a natural language utterance, and the second one (GRAPH) contains representations of *discourse objects* which have been presented on the display. Object representations contain a database query as their denotation (SEM feature) and information about their surface realization (SURF feature).

$$\begin{bmatrix} \text{PURPOSE} & : \textit{database-query} \\ \text{DISC} & : \textit{List of} \begin{bmatrix} \text{SURF} & : \begin{bmatrix} \text{SYN} & : \textit{NumberGenderFS} \\ \text{TYPE} & : \textit{DatabaseFieldSymbol} \end{bmatrix} \\ \text{SEM} & : \textit{database-query} \end{bmatrix} \\ \text{GRAPH} & : \textit{List of} \begin{bmatrix} \text{SURF} & : \begin{bmatrix} \text{TYPE} & : \textit{DatabaseFieldSymbol} \end{bmatrix} \\ \text{SEM} & : \textit{database-query} \end{bmatrix} \end{bmatrix}$$

Figure 2: Type of *FocusSpaceFS*

Salience Structure After each update of the focus stack a copy of it is sent to the natural language understanding unit. We call this copy the *Salience Structure*. It provides a structured view on salient *discourse objects* which are possible candidates for antecedents of anaphoric expressions. We claim that it contains important information about *discourse objects* which serve as possible antecedents for anaphoric expressions: (i) salience, (ii) modality (DISC/GRAPH), (iii) the *order* of occurrence, (iv) syntactic properties, and (v) semantic denotation.

For example see the alternative options in Figure 1: (5a) and (5c) are treated anaphorical, because they can be uniquely matched by a *discourse object* in the salience structure. The expression "rock" is matched by the displayed option "rock" introduced in (2). And the expression "the first one" is matched by the option "the dubliners" introduced in (4).

References

N. O. Bernsen and L. Dybkjær. 2000. From single word to natural dialogue. In M. V. Zelkowitz, editor, *Advances in Computers*, volume 52, pages 267–327. Academic Press, London.

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

Christian Hying. 2002. A context manager exploiting discourse structure. Diploma Thesis, IMS, Universität Stuttgart.

Unifying contrast and denial

Emar Maier
Department of Philosophy
University of Nijmegen
The Netherlands
e.maier@phil.kun.nl

Jennifer Spenader
Center for Language and Cognition
University of Groningen
The Netherlands
j.k.spenader@let.rug.nl

1 Denial vs. contrast

The extensive literatures on contrast and on denial give the impression (despite terminological confusions) that the phenomena are quite far apart. We consider the following to be paradigmatic examples of denial and contrast respectively:

(1) A: Juan's English is OK.
B: No, his English is not *OK*; he's as fluent as a native speaker!

(2) I was hungry but the restaurants were all closed.

Some apparent differences: (i) Denials are essentially a dialogue phenomenon as is obvious from the fact that (ii) denials taken out of their dialogue context are often plain contradictions (Horn, 1989), and for this reason (iii) their analysis necessarily involves nonmonotonic operations. Contrast on the other hand is (i') a discourse relation frequently occuring in monologue, (ii') never involving overt contradictions (**I am hungry but I am not hungry*)[1] and therefore (iii') often treated as an essentially monotonic phenomenon: what licenses a contrastive conjunction is not overt contradiction, but rather a conflict between what's defeasibly implied by the first and second conjuncts.

However, despite these differences, some examples seem to fit both categories equally well:

(3) A: Juan speaks Spanish.
B: Well, he IS Argentinian, but he DOESN'T speak Spanish. He grew up in the States.

The second contrastive conjunct of B's first utterance echoes the statement made by A, and it seems to retract the erroneous information as a textbook case of denial, with B's first conjunct constituting a partial concession. However, B's first statement also fits neatly into an analysis as contrast, because indeed the first conjunct weakly or defeasibly implies that the second conjunct is not true (*Argentinian*$\rightsquigarrow$*speak Spanish*). This paper shows that the overlap in contrast and denial analyses' of this example is no coincidence and can be generalized to a unified account of both phenomena.

2 Denial

We propose the following general structure of denials in the form of a rhetorical relation, expressing a relation between discourse segments, each defined as expressing only one (easily formalizable) intention.[2]

issue: the common ground is incremented with the first speaker's utterance.

concession: optional concessions of 2nd speaker to part of the information conveyed by the first, are added to the representation as well.

correction: the actual denial, headed by some negative or concessive particle (*no, but*) and/or an echo, initiating a *downdate* with the correcting information, i.e. add new info

[1] Unless we interpret the 2 occurrences of 'hungry' as referring to different properties.

[2] We formalized this in Layered DRT (Maier&van der Sandt 2003, Geurts&Maier 2004), a semantic framework capable of representing different types of content at different layers, enabling us to treat the (weakly) implied contradictions of contrast and the overt ones of denial in a similar way. See a longer version of this abstract at www.kun.nl/phil/tfl/~emar

and revise current common ground revised by throwing out older material until consistency is restored.

In example (1), A's utterance sets up the issue, the concession slot is empty, and B's statement plays the role of a correction. Example (3) has all 3 parts: A's utterance is the issue; B's remark that, well, he is Argentinian constitutes a concession since it corroborates the issue; the second conjunct of that statement (*but he doesn't speak Spanish*) is the correction, conflicting with the issue and triggering a revision operation. Note that the correction here starts with a *but* whereas in concessionless denials the role of 'denial-marker' is often played by a negated echo of the previous speaker's utterance (as in (1)) and/or a negative particle like *No*, *No way!* or *Bullshit*.

The formal semantic treatment suggested by the above schema combines the reverse anaphora approach of Maier and van der Sandt (2003) with a non-monotonic update or revision operation as in (van Leusen, ms; Asher and Lascarides, 2003) based on belief revision (Gärdenfors, 1988). Crucial for this to work is the recognition of echoes and the representation of not only asserted but also implicated and presupposed material, as in (1) where *his English is not OK* is merely an echo and the only contradiction to be resolved involves the scalar implicature of 'OK'.

3 Contrast

Consider again the contrast example (2): it's considered contrastive because the first conjunct (*we were hungry*) defeasibly implies that we went and got something to eat, whereas the second conjunct implies the opposite, cancelling the first defeasible implication. We argue, as do e.g. Winter and Rimon (1994), that one often has to take into account the discourse context in order to find this *tertium comparationis* (Lagerwerf, 1998). Taking the dependence on an issue in the context seriously, we suggest that the first slot in a contrastive discourse relation should contain this issue. A second segment then suggests a partial answer to the issue (parallelling the monotonic information growth with a denial's concession), whereas the final third segment gives a conflicting answer necessitating a revision and correction.

As the example analysis of (2) below shows, this description parallels exactly the 3-part coherence relation of denial above. In line with the above remarks on contrast however, we need to give some context, in this case the example requires that the topic of conversation is the question whether the speaker ate, which constitutes the issue. *I was hungry* is analysed as a concession, together with the inference *speaker has eaten* from that assertion in the context of the issue *Have you eaten?* Assuming that inferences of this type enter the discourse representation, this leads to a cross-layer contradiction with the second conjunct (assuming *restaurants closed* again in this particular context implies *speaker didn't eat*): the correction, headed by a *but* (as was typical for standard denials with concessions too).

4 Rectification vs. contrast

The unified discourse schema analysis proposed above can easily account for some puzzling facts about rectification adversative particles and contrast-denial particles. Some languages such as German, have a dedicated adversative particle (*sondern*) for rectification uses, reserving *aber* for contrast-denial, while other languages have lexicalized both meanings with the same particle (English: *but*).

In (4) we see how *aber* and *sondern* fit into one correction segment.

(4) A: Habt Ihr gegessen?
B: Wir haben Hunger gehabt, {aber/*sondern} wir haben nicht gegessen, {*aber/sondern} nur Bier getrunken.

On our account we can give a general (descriptive) characterization of this distribution: *aber* is the correction marker and must occur correction-segment initially, while *sondern* occurs within a complex correction. Furthermore, the difference in position inside the correction segment readily accounts for the observation that speaker changes are not possible in clauses joined by rectification particles but are fine with a contrast-denial *but* (von Klopp, 1994) since speaker changes are only natural at discourse segment boundaries.

Conversational gameboard and discourse structure

Nicolas Maudet[⋆], **Philippe Muller**[†] **and Laurent Prévot**[♣]
[⋆] LAMSADE, Université Paris 9 Dauphine, Paris
[†] IRIT and Université Toulouse 3, Toulouse
[♣] LOA, ISTC-CNR, Trento

This paper tries to bring closer two theories of human communication:

Commitment stores and dialogue games Hamblin (Hamblin, 1970) introduced the notion of *commitments stores* whereby dialogue participants can keep track of (public) commitments that arise during the interaction. He also pointed out the rule-governed nature of dialogues, and tried to exhibit set of normative rules (*dialogue games*) which could prevent certain types of fallacies

Discourse semantics The primary aim of this approach was to extend Montagovian compositional semantics to account for phenomena observed at the discourse level. This motivated a shift from static truth-semantics to an update semantics (that is, sentences are regarded as update functions on possible worlds). In this perspective, the semantic/pragmatic interface becomes the focus of attention.

The case for a crossover. To begin with, one may ask why dialectical models are not enough to model human conversations. Very often, a turn is composed of several basic units. Under the assumption that the speaker obeys coherence principles (e.g. the so-called "right frontier" of discourse structure (Asher and Lascarides, 2003)), it is for instance possible to define those discourse referents that can be used later in the dialogue. This, of course, can prove to be crucial when facing the interpretation of follow-up utterances and dialogue turns. It is clear that current dialectical approaches fall short of being able to account for these aspects, as they are simply not equipped with notions allowing to deal with this level of analysis. On the other hand, the very same observations can be made at the level of dialogue turns, thus emphazing the need to take into account dialogue structure.

Semantic content, speech acts and rhetorical relations We take speech acts as conversational basic units, consisting of a propositional content and an illocutionary force. Following proposals in SDRT (Asher and Lascarides, 2003), semantic content is represented as Discourse Representation Structure (Kamp and Reyle, 1993) (K_π), and augmented by specifying the producer (L_π) and the mood (affirmative, interrogative, imperative) (M_π) of the utterance. In SDRT, assuming the coherence of a discourse means that each utterance has to be related to the context with a rhetorical relation (except the first one). Such relations are defined by their triggering conditions and their semantic effects. In discursive approaches, coherence is verified if an utterance can be successfuly attached to the context. Likewise, coherence in conventional approaches of dialogue corresponds to the successful integration of a dialogue act in an authorized dialogue game.

We use the discourse structure definition presented in (Asher and Lascarides, 2003)(section 4.4.1). $\mathcal{DS} = \langle A, \mathcal{F}, LAST \rangle$ where A is a set of labels, $\mathcal{F}$ is an assignation function from labels to well-formed SDRSs, and $LAST$ is the last discourse label introduced. A well-

formed SDRS is either (i) a logical form for atomic natural language clauses (like DRSs), (ii) a discourse relation between labels, (iii) the dynamic conjunction of two well-formed SDRSs or finally (iv) the negation of a SDRS.

A *Conversational scoreboard* consists of the discourse structure ($\mathcal{DS}$) and commitment stores ($\mathcal{E}_A$ and $\mathcal{E}_B$) of speakers A and B over certain elements of $\mathcal{DS}$: $\mathcal{CS} = \langle \mathcal{DS}, \mathcal{E}_A, \mathcal{E}_B \rangle$. Elements of $\mathcal{E}_X$ are SDRS contents, i.e. either simple DRSs, or complex constituants. Some of these contents received a negative polarity if speakers are commited to their falsity (linked to an expressed disagreement; this has nothing to do with private beliefs, but reflects public information).

We consider here how commitment evolves, and how this can be seen as an interpretation of coherence relations in a dialogue, whether we consider "monologic" relations[1] or properly dialogic relations. The following rules can thus be seen as update rules of the board for each recognized act for which a relation with the context can be inferred. The first case to consider is for monologic vericonditional relations, *i.e.* relations whose dynamic semantic is of the form ($\wedge_{dyn}$ is dynamic conjunction) as proposed in (Asher and Lascarides, 2003):
$(w,f)[\![R(\pi_1,\pi_2)]\!]_M(w',g)$ *ssi* $(w,f)[\![\mathcal{F}(\pi_1) \wedge_{dyn} \mathcal{F}(\pi_2) \wedge_{dyn} \phi_{R(\pi_1,\pi_2)}]\!]_M(w',g)$

Here ϕ stands for semantic effects due to each relation (e.g. $narration(\pi_1,\pi_2)$ implies a temporal succession of events in π_1 and π_2, and a common topic for the pair). In a dialogue context, the producer of π_1 and π_2 is then committed to the content of π_1 and π_2, *and the rhetorical link between them*, because semantic effects can be seen as conventionally implied.

We note $\Rightarrow_\pi$ the update of commitment stores by a constituant π. If both π_1 and π_2 are produced by the same speaker, her commitment store will evolve as follows (in two steps):

$\mathcal{E}_A \Rightarrow_{\pi_1} \mathcal{E}_A \cup \{\mathcal{F}(\pi_1)\} = \mathcal{E}_{A'} \Rightarrow_{\pi_2} \mathcal{E}_{A'} \cup \{\mathcal{F}(\pi_2), \phi_{narration(\pi_1,\pi_2)}\}$

The remaining of the board is left unchanged. If it is a "monologic" relation across speech turns (π_1 is said by A, and π_2 by B), then only the first update applies to A, while both updates apply to B's commitment store. Likewise, other relations can be interpreted as commitment updates, with "truth" replaced by the corresponding commitment of a speaker to a proposition.

Given that speaker B utters π_2, and that π_2 is to be attached to π_1 with the relation R we define commitments for dialogic relations in the following way :

- if $R = relation_q$ (i.e $relation_q(\pi_1,\pi_2)$ holds)[2] then B's commitments are not affected but *relation* defines the commitments concerning the answer to the question (see next case)

-if $R = QAP$ (Question Answer Pair) then B commits himself to the answer and to the link between the context and the question-answer pair. If $Relation_q(\pi_0,\pi_1)$ then $\mathcal{E}_B \Rightarrow \mathcal{E}_B \cup \{\mathcal{F}(\pi), \phi_{Relation(\pi_0,\pi)}\}$ where π corresponds to the resolved question-answer pair.

- if R = *acknowledgement*, then B commits herself to π_1 content.[3] $\mathcal{E}_B \Rightarrow \mathcal{E}_B \cup \{\mathcal{F}(\pi_1)\}$

- if R=*challenge* (as defined in (MacKenzie, 1979)) A cannot go further without either withdrawing or justifying the proposition. $\mathcal{E}_A \Rightarrow \mathcal{E}_A \cup \mathcal{F}(\pi_1)$ and $\mathcal{E}_B \Rightarrow \mathcal{E}_B \setminus \mathcal{F}(\pi_1)$

References

N. Asher and A. Lascarides. 2003. *Logics of conversation.* Cambridge University Press.

C.L. Hamblin. 1970. *Fallacies.* Methuen.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic.* Kluwer Academic Publishers.

J. MacKenzie. 1979. Question-begging in non-cumulative systems. *Journal of philosophical logic*, 8:117–133.

[1] "Monologic relations" have to be understood as relations that were already studied in the monologue case even if they hold also across speech turns.

[2] These relations are interrogative content relation (e.g *elaboration*, *narration*). The constituent associated with the question answer pair will be attached to the context through the corresponding "monologic" relation.

[3] Acknowledgement is more complex when communication is not taken as "perfect".

Small group discussion simulation for middle Level of Detail Crowds

Jigish Patel and **Robert Parker** and **David Traum**
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA
traum@ict.usc.edu

1 Introduction

In the Mission Rehearsal Exercise at University of Southern California, (Swartout et al., 2001), a leader is trained with a story-based immersive simulation including many characters, both teammates and others. While a number of these characters (especially the ones in the front and center) play lead roles in the story and interact heavily with the trainee (Rickel et al., 2002), there are also a number of "supporting" characters who play fairly minor roles, but are still important to the setting of the story. The original versions of these characters had all their motions painstakingly hand animated, and were set in loops when the interaction lasted longer than the amount of scripting. Such scripting has three problems: first it is labor intensive, second, it is not reactive to local circumstances, and third, the repetition can detract from the realism, even if well animated for short segments.

A solution to these problems is to use some automatic simulation rather than hand-scripting. As (O'Sullivan et al., 2002) point out, crowd and group simulations are becoming increasingly important for a number of applications, including movies, as well as games and simulations. Random or scripted behaviors are satisfactory for low levels of details (e.g., very distant crowds), and full animated conversational agents are adequate for the main characters, but these are overkill for middle-level group members who are seen at some distance and not directly interacted with. What we need for our middle-level characters is something good enough to look like characters involved in conversation without the overhead of fully intelligent agents. A very good starting point is provided by (Padilha and Carletta, 2002), who synthesize some of the best research on group dialogue behavior into a parameterizable, probabilistic algorithm for individual behavior as part of a group. We have re-implemented this simulation, with some enhancements, and used the results to animate the Bosnian crowd members in the Mission Rehearsal Exercise.

2 Crowd Simulation for Animation

While (Padilha and Carletta, 2002) have a simulation algorithm with results specifying outputs such as talking, gestures of a few sorts, and gaze, they did not actually link up the simulation to an animation system. Doing such, in this case to BDI's Peopleshop^{TM} characters, necessitated making individual choices of types of gestures to indicate speaking and other motions. Figure 1 shows a snapshot of the characters involved in conversation.

We have also made several extensions to the simulation of (Padilha and Carletta, 2002) to account for the use of this simulation as embedded in the virtual world. First some extensions to the gaze model, to account for change of addressee and audience gaze at multiple speakers. More importantly, though, we also allow attention to pass away from the group discussion to focus on external events such as the main conversation between the human trainee and main character virtual humans and other occurrences, such as explosions and people and vehicle movements.

Figure 1: Bosnian Group in conversation

The simulation runs by cyclically testing a set of parameter values against random numbers, with the results leading to decisions of whether to speak or listen or attend elsewhere and which gestures to make.

These parameters were defined in (Padilha and Carletta, 2002):

talkativeness: likelihood of wanting to talk.

transparency: likelihood of producing explicit positive and negative feedback, and turn-claiming signals.

confidence: likelihood of interrupting and continuing to speak during simultaneous talk.

interactivity: the mean length of turn segments between TRPs.

verbosity: likelihood of continuing the turn after a TRP at which no one is self-selected.

In addition, we added the following parameters:

responsiveness: likelihood of a participant reacting to interruptions from outside the group.

continuity: likelihood of selecting an addressee (for example, by asking a question to him/her specifically) at the end of the speaker's turn.

A loop (a modification of the algorithm in (Padilha and Carletta, 2002)) is executed every cycle (approximately 500 ms long) by each character. The main modifications involve allowing agent responsiveness to events and speech outside the group and the linking of abstract behaviors to specific animation calls for the characters.

3 Evaluation

Padilha and Carletta's evaluation plan involved comparing their simulation to transcripts of group conversation data, showing a better fit than simpler models. While this kind of evaluation would certainly be interesting, we propose a different kind of evaluation - whether the simulation "looks like a conversation" to a viewer. Two baselines for performance are whether the simulation looks more natural than random motion and whether the simulation looks more natural than the looping, hand-crafted animation.

We also want to evaluate the effects of the individual parameters. We have constructed experiments in which different characters are given different values for parameters (such as talkativeness and confidence), and then showed viewers recordings of different simulation runs with these parameters to judge features like apparent talkativeness of individual characters.

References

C. O'Sullivan, J. Cassell, H. Vilhjalmsson, J. Dingliana, S. Dobbyn, B. McNamee, C. Peters, and T. Giang. 2002. Levels of detail for crowds and groups. *Computer Graphics Forum*, 21(4).

E. G. Padilha and J. C. Carletta. 2002. A simulation of small group discussion. In *Proceedings of EDILOG 2002: Sixth Workshop on the Semantics and Pragmatics of Dialogue*, pages 117–124.

Jeff Rickel, Stacy Marsella, Jonathan Gratch, Randall Hill, David Traum, and William Swartout. 2002. Toward a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems*, 17.

W. Swartout, R. Hill, J. Gratch, W.L. Johnson, C. Kyriakakis, K. Labore, R. Lindheim, S. Marsella, D. Miraglia, B. Moore, J. Morie, J. Rickel, M. Thiebaux, L. Tuch, R. Whitney, and Jay Douglas. 2001. Toward the holodeck: Integrating graphics, sound, character and story. In *Proceedings of 5th International Conference on Autonomous Agents*.

Employing Context of Use in Dialogue Processing

Botond Pakucs
Centre for Speech Technology (CTT)
KTH, Royal Institute of Technology
Lindstedtsvägen 24, 100 44 Stockholm, Sweden
botte@speech.kth.se

Abstract

In this paper, a generic solution is presented for capturing, representing and employing the context of use in dialogue processing. The implementation of the solution within the framework of the SesaME dialogue manager and the Butler demonstrator is also described.

1 Introduction

In natural human-to-human communication, speakers are able to use implicit contextual information to increase the conversational bandwidth. The implicit contextual information is relevant knowledge about the actual situation. However, this knowledge is not necessarily part of the linguistic context (has not been uttered earlier).

The following real life example illustrates the use of implicit knowledge about the situation.

An employee leaving the office:
Q: When is my train leaving?
A: At 17:30.

This dialogue appears somewhat strange, incomplete and even incomprehensible to others than the participants. In spite of this, the dialogue exhibits a successful interaction. This short dialogue appears to be a repetition of similar interactions encountered previously. The answer contains prediction and beliefs about the dialogue partner's individual goals, intentions and preferences.

For achieving more natural and efficient communication, it is desirable to enable spoken dialogue systems to support similar interactions.

2 Employing context of use in SesaME

SesaME (Pakucs, 2003) is a generic dialogue manager specially developed to enable multi-domain dialogues in mobile environments. The central idea is to know as much as possible about the users. Each user is expected to use an individual and *highly personalized speech interface* to access a multitude of services and appliances. This is achieved through a personalized speech interface integrated into some personal and wearable appliance such as a mobile phone or a PDA. The application specific data, including the dialogue management capabilities, is locally stored at the service provider side and is dynamically plugged into the personalized speech interface and activated whenever the user enters a new environment.

The SesaME architecture is comparable to other agent-based architectures such as the TRIPS (Allen et al., 2000) architecture. A central information storage, a blackboard, stores the representation of the system's *information state* (Larsson and Traum, 2000). However, this representation is not formalized; the information state is merely a collection of all data available to the dialogue system. An event-based solution is used for updating the information state, where events can be dialogue moves, internal system events, or changes in the user's context of use.

2.1 Knowledge representation

After each interaction with a user, every utterance and the related contextual information is represented as a feature vector containing feature-value pairs of all relevant information related to the in-

teraction. The only common property of the features in the feature vector is the co-occurrence. The feature vectors are indexed and stored in individual user models implemented as vector-space models. For manipulating the user models, well-known information retrieval solutions are used.

Albeit, every feature vector is domain and task dependent, the individual user models are generic and they may contain feature vectors from several different domains and tasks. In this way, capturing and employing cross-domain user characteristics is also feasible.

2.2 Context based adaptation

In SesaME, a content-based solution (Zukerman and Albrecht, 2001) is employed for performing a context-based adaptation to individual users.

A context manager keeps track of the user's current context. During every new interaction, based on the available contextual information one or more feature vectors are built. These vectors are used for retrieving similar interactions from the user model.

The retrieved results are used to predict specific features of the ongoing interaction and to achieve adaptation to the current context. For example, based on earlier interactions with a voice controlled travel-planer it may be possible to detect that a commuter's most frequent choice of destination on weekday evenings is "Stockholm". Thus, it is possible to ask the user a more natural question: *"Would you like a ticket to Stockholm?"* instead of the impersonal default prompt: *"Where would you like to travel?"*.

However, if no similar interactions are present in the user model, or no obvious patterns are detected, the default prompt is used.

3 Application and Evaluation

The Butler (Pakucs, 2004) is a telephony-based multi-domain dialogue system developed for evaluating the employment of contextual information in SesaME. The services provided by Butler can be categorized in three main categories, *public services* such as accessing commuter and subway train timetables, menu information for the nearby restaurants, *accessing personal information* from calendars and *accessing workplace related information*, such as time and location of meetings and seminars.

All these services are based on information services available on the Internet. The relevant knowledge is automatically extracted from the available html-documents and transformed into dialogue descriptions at runtime. The users are identified through speaker verification or based on the used mobile-phone numbers (A-numbers).

Currently a long-term evaluation of the Butler is conducted. However, preliminary data indicates that erroneous system predictions are considered as natural and non-disturbing by the users.

4 Conclusions

In this paper, a generic solution for employing implicit contextual knowledge in dialogue management was introduced. By encoding contextual information in the user model it becomes feasible to predict specific features of an ongoing interaction. Thus, a simultaneous adaptation to an individual user and the user's current situation is supported. This solution appears to be a promising contribution to providing a more flexible and natural interaction in spoken dialogue systems.

Acknowledgments

This research was carried out at the CTT, a competence center at KTH and was also sponsored by the European Union's IST Programme under contract IST-2000-29452, DUMAS.

References

James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6(3-4):213–228, December.

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

Botond Pakucs. 2003. Towards Dynamic Multi-Domain Dialogue Processing. In *Proceedings of the Eurospeech'03*, volume 1, pages 741–744, Geneva, Switzerland, September.

Botond Pakucs. 2004. Butler: A Universal Speech Interface for Mobile Environments. In *Proceedings of the Mobile HCI 04*, Glasgow, Scotland, September.

Ingrid Zukerman and David W. Albrecht. 2001. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11:5–18.