How People Talk to Computers, Robots, and Other Artificial Communication Partners

Kerstin Fischer (Ed.)

SFB/TR 8 Report No. 010-09/2006
Report Series of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition
Universität Bremen / Universität Freiburg

Contact Address:

Dr. Thomas Barkowsky
SFB/TR 8
Universität Bremen
P.O.Box 330 440
28334 Bremen, Germany

Tel +49-421-218-64233
Fax +49-421-218-98-64233
barkowsky@sfbtr8.uni-bremen.de
www.sfbtr8.uni-bremen.de

How People Talk to Computers, Robots, and Other Artificial Communication Partners

Proceedings of the Workshop
Hansewissenschaftskolleg,
Delmenhorst
April 21-23, 2006

Kerstin Fischer (ed.)

Contents

Introduction to the Volume

Kerstin Fischer
University of Bremen
kerstinf@uni-bremen.de

There is a growing body of research on the design of artificial communication partners, such as dialogue systems, robots, ECAs and so on, and thus conversational interfaces are becoming more and more sophisticated. However, so far such systems do not meet the expectations of ordinary users. One reason that prevents systems being perceived as useful and fully functional may be that there is still very little known about the ways human users actually address such conversational interfaces. How naive speakers really interact with such systems and the language that they use to do so cannot be deduced by intuition; effective language of this kind is simply not available to introspection. Moreover, empirical linguistic and psychological studies of the ways people talk to artificial communication partners so far have yielded only very particular, corpus-, domain- or situation-specific results. What is needed, therefore, is to bring together results from various different scenarios in order to achieve a more general picture of the determining factors of different ways of talking to artificial agents, such as dialogue systems, ECAs, robots and the like, aiming at a model that promises both reusability of results achieved in different human-computer situations and predictability with respect to behaviours that may be expected of new human-computer interfaces. In this area, researchers have only just begun to explore the role of central pragmatic mechanisms, such as recipient design, alignment, and interactional strategies, such as feedback, in communication with artificial communication partners. Here, psychological and linguistic studies will certainly reveal dialogue strategies that support dialogue system design. Furthermore, system design may profit from the identification of different user groups. For instance, a compromise between fully speaker-independent systems (word-error rate too high) and fully speaker-dependent systems (low word-error rate but confined to one speaker) might be to establish different types of speakers according to their linguistic behaviour and to establish different recognizers especially tailored for these different groups. Finally, the fact that speakers align to their communication partners should be exploited by shaping the linguistic behaviour of speakers in a way which is

most useful for the system to understand. This involves issues of initiative, feedback, and dialogue act modelling. The contributions to this volume are thus highly relevant from theoretical and practical perspectives. The volume addresses one of the most urgent deadlocks in current dialogue system design and evokes an interdisciplinary perspective on the problem, providing theoretically interesting and practical ways out of current dilemmas, connecting scientists from different disciplines. The papers focus particularly on the following questions:

- Which different types of linguistic behaviours (phonetic, prosodic, syntactic, lexical, conversational) can be found in communication with artificial communication partners?
- Do these types of behaviours cluster in particular ways such that some behaviours tend to co-occur with others so that different types of users become apparent?
- Are there particular linguistic means to identify different types of users (unobtrusively and online)?
- Which aspects of the design condition which kinds of behaviours?
- Which roles do recipient design, alignment, and feedback play in the communication with artificial communication partners?
- Which kinds of problems in dialogue modelling and automatic speech processing can be prevented by modelling different kinds of linguistic behaviours and different types of users?

Three papers are concerned with the details of linguistic interation, how people react to particular linguistic features of linguistic output from robots. **Robert Porzel** looks at entrainment, the role of pauses, structuring cues, hesitation markers and discourse particles in human-to-human versus human-to-computer communication. **Britta Wrede, Stefan Buschkämper, Claudia Muhl** and **Katharina Rohlfing** are concerned with users' reactions to different kinds of feedback from the robot. **Thora Tenbrink** compares interaction with an autonomous wheelchair with and without linguistic feedback and shows how the robot's linguistic output can reduce the variability of linguistic structures and guide the speakers into producing what the robot understands best.

Three papers address the nature of language directed at systems. **Petra Gieselmann** and **Prisca Stenneken** investigate syntax and the lexicon

of language directed at a robot, providing further evidence for the register hypothesis [1]. Also **Anton Batliner, Christian Hacker** and **Elmar Nöth** investigate the properties of computer talk, focussing on the phonetic and prosodic delivery of utterances, comparing it with off-talk produced by the same speakers with how they address an automatic speech processing system. **Stefan Kopp** analyses the kinds of utterances speakers in unrestricted scenarios direct towards an embodied conversational agent. His investigation focuses on quantitative semantic and pragmatic analyses of such interactions with the result that many speakers apply communicative strategies from human-to-human communication in the communication with the embodied conversational agent.

Two papers deal with the users' mental models of artificial communication partner and their communicative consequences. **Kerstin Fischer** shows that only some users in human-computer and human-robot interaction attend to communicative strategies from conversations among humans, and that the different preconceptions, computer/robot as a tool versus as a social actor, have consequences for the users' linguistic behaviour on all linguistic levels, the so-called register features as well as their interactional behaviour, for example, with respect to alignment. **Elena Andonova** uses questionnaire data to establish mental models of robots before and after human-robot interaction. She identifies features that persist and thus constitute stable aspects of peconceptions of robots and features that may change during the course of the interaction.

Two papers address alignment in more deatil: **Holly Branigan** and **Jamie Pearson** discuss and compare findings on the relationship between alignment and recipient design in human-to-human versus in human-computer communication, argueing that speakers do not regard computers as social actors, contrary to claims by Clifford Nass, for instance [3, 2]. **John Bateman** provides a social/semiotic perspective both on register and alignment and discusses the problems for an implementation of alignment in dialogue systems.

Finally, **Robert Ross** discusses the usability of the information state update approach for a dialogue modeling that allows interactions with robots, not just on the level of tool-using, but as interactions with a social agent.

Acknowledgements The workshop was organised in cooperation with Anton Batliner (University of Erlangen) and took place April 21-23, 2006, at the Hanse Wissenschaftskolleg (HWK) in Delmenhorst in the vicinity of Bremen. We would like to express our gratitude to the HWK and in

particular to Wolfgang Stenzel for the excellent organisation, as well as to the SFB/TR8 *Spatial Cognition*, funded by the DFG, for the financial support.

References

[1] J. Krause and L. Hitzenberger, editors. *Computer Talk*. Hildesheim: Olms Verlag, 1992.

[2] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

[3] B. Reeves and C. Nass. *The Media Equation*. Stanford: CSLI and Cambridge: Cambridge University Press, 1996.

How Computers (Should) Talk to Humans

Robert Porzel
University of Bremen, Germany
porzel@informatik.uni-bremen.de

Abstract

End-to-end evaluations of more conversational dialogue systems with naive users have uncovered severe usability problems that, among other things, result in low task completion rates. First analyses suggest that these problems are related to the system's dialogue management and turn-taking behavior. This paper starts with a presentation of experimental results, which shed some light on the effects of that behavior. Based on these findings, some criteria which lie orthogonal to dialogue quality are spelled out. As such, they nevertheless constitute an integral part of a more comprehensive view on dialogue felicity as a function of dialogue quality and efficiency. Since the work on spoken and multimodal dialogue systems presented and discussed herein is aimed at more conversational and adaptive systems, we also show that - in certain dialogical situations - it is important for such systems to align linguistically towards the users. After describing the corresponding empirical experiments and their results, pragmatic alignment will be introduced as more general framework for these types of adaptation to users which are, in the light of the aforementioned studies critical to building more conversational dialog systems.

1 Introduction

Research on dialogue systems in the past has by and large focused on engineering the various processing stages involved in dialogical human-computer interaction (HCI) - e.g., robust automatic speech recognition, natural language understanding and generation or speech synthesis [3, 17, 6]. Alongside these efforts the characteristics of computer-directed language have also been examined as a general phenomenon [69, 67, 18]. The flip side, i.e., computer-human interaction, has received very little attention as a research question by itself. That is not to say that natural language generation and synthesis have not made vast improvements, but rather that the nature and design of

the computer as an interlocutor itself, i.e., the effects of *human-directed language*, have not been scrutinized to the same degree. Looking, for example at broad levels of distinctions for dialogue systems, e.g., between controlled and conversational dialogue systems [2], we note the singular employment of human-based differentiae, i.e., degrees of restrictedness in the linguistic behaviour for the *human* interaction. Differentiae stemming from the other communication partner, i.e., the computer, are not taken into account - neither on a practical nor on a theoretical level.

In the past controlled and restricted interactions between the user and the system increased recognition and understanding accuracies to a level that systems became reliable enough for deployment in various real world applications, e.g., transportation or cinema information systems [5, 30, 28]. Today's more conversational dialogue systems, e.g., SmartKom [61] or MATCH [37], have been engineered to be able to cope with less predictable user utterances. Despite the fact that in these systems recognition and processing have become extremely difficult, the reliability thereof has been pushed towards acceptable degrees by employing an array of highly sophisticated technological advances - such as:

- dynamic lexica for multi-domain speech recognition and flexible pronunciation models [55],
- robust mulit-modal fusion, understanding and discourse modeling techniques [36, 22, 1]
- and ontological and contextual reasoning capabilities [31, 51, 49].

However, the usability of such conversational dialogue systems is still unsatisfactory, as shown in usability experiments with real users [7] that employed the PROMISE evaluation framework [8], which offers some multimodal extentions over the uni-modal PARADISE framework [63].

The work described herein constitutes a starting point for a scientific examination of the whys and wherefores of the challenging results stemming from such end-to-end evaluations of more conversational dialogue systems. Following a brief description of the state of the art in examinations of computer-directed language, we shortly describe several prior experiments, which sought to lay the ground for a more systematic examination of the effects of the computer's linguistic behaviour in more conversational spoken dialogue systems. Based on these results, we will discuss the ensuing implications for the design of successful and felicitous conversational dialogue systems in which computers talk as they should followed by some comcluding remarks and future work.

2 Prior Work

The complete understanding of specific characteristics of dialogical interaction is still an unresolved task for (computational) linguistics. Linguistic adaptation, e.g., alignment, entrainment and the like, presents such a specific characteristic in dialogue, which has been explored by linguists [29] and recently came into focus of computational linguistics [16, 52]. Linguistic adaptation, in general, can be described as the process of tailoring any form of linguistic behavior or output towards the recipient of that output. We will firstly summarize prior art in human-human communication followed by a corresponding summary in human-computer communication.

2.1 Adaptation in Human-Human Communication

Speakers may not always be aware of the potential ambiguities inherent in their utterances. They leave it to the context to disambiguate and specify the message. Furthermore, they trust in the addressee's ability to extract that meaning from the utterance that they wanted to convey. In order to interpret the utterance correctly, the addressee must employ several recourses. Speakers in turn anticipate the employment of these interpretative recourses by the hearer and construct the utterance knowing that certain underspecifications are possible since the hearer can infer the missing information or that certain ambiguities are permissible, etc. The role of the communicative partner is of paramount importance in this process.

The general necessity of the inclusion of a partner model in the modeling of human-human communication seems undisputed at the moment, even though some of the views presented below have recently been challanged by some empirical findings [24]. Without a partner model several empirically observable phenomena cannot be explained. We will present some findings as they are relevant to the studies and work presented herein. A departure from prior modes of looking at human-human communication is summed up by social psychologists [42] who have pointed out that

> "'the traditional separation of the roles of participants in verbal communication into sender and receiver, speaker and addressee, is based on an illusion — namely that the message somehow 'belongs to' the speaker, that he or she is exclusively responsible for having generated it, and that the addressee is more-or-less a passive spectator to the event. (...) the addressee is a full participant in the formulation of the message — that is the vehicle by

which the message is conveyed — and, indeed, may be regarded in a very real sense as a cause of the message"' (ibid:96)

The listener has, therefore, come to be regarded as an essential part in the causation of speech production in a communicative setting; in part responsible for and shaping the speaker's behaviour through the following means:

- Back-channeling: Some results of back-channeling [68], — which is the phenomenon of verbal and non-verbal (or quasi-verbal) responses of the listener during the speaker's discourse, such as *yes*, *hmmm*, *I see*, *uh-huh*, facial expressions, nods, gestures, etc. — have been specified and experimentally displayed [43]. Therein, the effects of back-channeling on the development of the redundancy of words and phrases within a discourse are described. In general, the effect is, that exact repetitions of phrases and/or words are less likely when back-channeling occurs. In the event of back-channeling the usage of abbreviations and phrase-reductions increases. Back-channeling also has a significant bearing on the course of the discourse. It has also been shown that the availability of visual contact between speaker and listener greatly influences the efficiency of the discourse [44].

- Common ground: The influence of common ground, i.e., the shared knowledge, shared associations, shared sentiments, and shared defaults, between speaker and listener has been identified and described [39, 15]. Common ground has, therefore been shown to influence the lexicalization preferred by the speaker - for example, what kind of words to use - or whether to describe objects more figuratively or literally. Furthermore, it influences the type versus token ratio in the speakers' discourse as well as the length and specificallity of descriptions.

- Social factor(s): Further research has demonstrated that some verbalizations, e.g., non-egocentric localizations, demand more mental attention than, for example, egocentric ones [13], which speakers are more willing to invest when speaking to social superiors or based on some estimation of the recipient's cognitive competence, e.g. when talking to children [32].

In this light the notion of *lexical entrainment* [29, 9, 10] constitutes another crucial aspect of linguistic alignment. Research teams found that word choice within a dialogue is dependent on the dialogue history. In fact their results show that through hedging two interlocutors adopt each other's

terms and stay with it for the remainder of the dialogue. The variability in word choice is huge in any field. This phenomenon has been labeled as the *Vocabulary Problem* [27]. Although there are no real synonyms, i.e. two words that in all contexts would be used interchangeably, people still have individual preferences when referring to an object in a given context.[1] In some cases it further depends on the interlocutors' perspective whether they adapt to their conversational partner or whether they do not. For example, throughout a court trial in which a physician was charged with murder for performing an abortion, the prosecutor spoke of *the baby* while the defense lawyer spoke of *the fetus* [11]. If people wish to align within a conversation and adopt each others lexical choices, the interlocuter who introduces a term has been denoted as the *leader* and the one who adopts it as the *follower* [29].

However, entrainment represents the peak of a foregoing alignment, i.e. the cooperation process. First, the interlocutors need to establish a common ground for their conversation [9]. After that they hedge, i.e. they mark the term as provisional, pending evidence of acceptance from the other [10]. Only then do they agree on the same choice of words. As a last step, entrained terms are no longer *indefinite* and can be shortened, e.g. via anaphora, one-pronominalization, gapping or elision [45].

2.2 Adaptation in Human-Computer Communication

The first studies and descriptions of the particularities of dialogical human-computer interaction, then labeled as *computer talk* in analogy to *baby talk* [69], focused - much like subsequent ones - on:

- proving that a regular register for humans conversing with dialogue system exists [41, 26],
- describing the general characteristics of that register [40, 18].

The results of these studies clearly show that such registers exists and that their regularities can be replicated and observed again and again. In general, previous work focuses on the question: what changes happen to human verbal behavior when they talk to computers as opposed to fellow humans? The questions which are not asked as explicitely are:

- how does the computer's way of communicating affect the human interlocutor,

[1] For instance, in a user study conducted by Furnas *et al.* [27] subjects used several different words for *to delete*: *change*, *remove*, *spell* or *make into*.

- do the particulars of computer-human interaction help to explain why today's conversational dialogue systems are by and large unusable.

To the best of our knowledge, there has not been a single publication reporting a successful end-to-end evaluation of a conversational dialogue system with naive users. We claim that, given the state of the art of the adaptivity of today's conversational dialogue systems, evaluation trials with naive users will continue to uncover severe usability problems resulting in low task completion rates.[2] Surprisingly, this occurs despite acceptable partial evaluation results. By partial results, we understand evaluations of individual components such as concerning the word-error rate of automatic speech recognition or understanding rates [19, 33].

As one of the reasons for the problems thwarting task completion, researchers point at the problem of *turn overtaking* [7], which occurs when users rephrase questions or make a second remark to the system, while it is still processing the first one. After such occurrences a dialogue becomes asynchronous, meaning that the system responds to the second last user utterance while in the user's mind that response concerns the last. Given the current state of the art regarding the dialogue handling capabilities of HCI systems, this inevitably causes dialogues to fail completely.

We can already conclude from these informal findings that current state of the art conversational dialogue systems suffer from

- a lack of turn-taking strategies and dialogue handling capabilities and
- a lack of strategies for repairing dialogues once they become *out of sync*.

In human-human interaction turn-taking strategies and their effects have been studied for decades in unimodal settings [20, 57, 64] as well as more recently in multimodal settings [60]. Virtually no work exists concerning the turn-taking strategies that dialogue systems should pursue and how they effect human-computer interaction, except in special cases, e.g. in conversational computer-mediated communication aids for the speech and hearing impaired [66] or for turn negotiation in text-based dialogue systems [59]. Overviews of classical HCI experiments and their results also shows that problems, such as turn-overtaking, -handling and -repairs, have not been addressed by the research community [67].

[2]These problems can be diminished, however, if people have multiple sessions with the system and adapt to the respective system's behavior.

It has also been shown that entrainment is of major importance in tutorial systems [16]. Here the arguement goes that especially students do not always know specific terms and use common sense terms instead. Instead of treating those terms as completely incorrect students should however be given partial credit for expressing the right general idea. For this reason their system NUBEE, a parser within a tutorial system, looks up unknown words in the WORDNET database [23] and searches for synonyms that match.

Looking back at the notion of leader and follower in entrainment phenomena, it becomes clear that, especially in an expert-novice relationship, the expert should also function as follower and not only as leader. An open question, to be answered by means of one of the studies described below, is whether in shorter exchanges, e.g. in an assistance, help-desk or hotline setting, we find specific cases of entrainment or not among human interlocuters. Adaptation by computers to their users has been examined in various branches of natural language generation from epistemic factors such as prior knowledge or cognitive competence [34, 38, 47] via stereotypes [56] to multimodal preferences [21].

3 Studies on Computer-Human Interaction

In the following two sets of studies will be described, which sought to examine the effects of the computer's turn taking and entrainment behavious on human-computer dialogues.

3.1 Entrainment Studies

The notion of lexical entrainment was first established by Garrod and Anderson [29] and later explored by Brennan [9, 10].[3] It is, therefore, well known that in human-human dialogues the interlocutors converge on shared terms and phrases, e.g. if A talks to B and uses a term such as *pointer* to refer to an graphically displayed object, i.e. leads in the usage of the term - and B (from then on) also employs the term, i.e. follows lead of A, then we have a classic case of entrainment. A viable hypothesis, addressed in this research effort, is that dialogue efficiency and user-satisfaction could be increased considerably if spoken dialogue systems also adapted the user's choice of terms rather than staying with their own fixed vocabulary. In the

[3]We follow their understanding of the term *lexical entrainment*, i.e. that people adopt their interlocutor's terms in order to align with them over a certain period of time.

following we summarize two studies on entrainment - one in a human-human setting and one using a Wizard-of-Oz human-computer set-up.[4]

3.1.1 Entrainment in Asisstance Dialogues

As for implementing entrainment in a multimodal dialogue system that features spoken interaction as a modality, it is important to find out under which circumstances people entrain in human-human dialogues. Based on such findings decisions can be made whether it is viable and beneficial to train a classification system that can be used to compute in a specific dialogue situation that entrainment should be performed or not. Furthermore, there might be application scenarios in which entrainment is more necessary than in others.

In order to study entrainment in the domains of assistance systems, e.g., help-desks, hotline or call center systems, and to develop and test an annotation scheme we collected a corpus of human-human dialogues. The data collection was conducted by means of a Multiple Operator and Subject (MOS) study that is in essence akin to the benchmark impurity graph evaluation paradigm [46]. In the MOS study, new operators as well as new subjects were recruited after each session, resulting in new pairs for each session. By these means we were able to avoid long term adaptation through familiarity caused by prior interactions. During the trials, the operators were to act as a call-center agent who had to answer questions posed by the subjects regarding operating a very modern television set, that as an additional feature has Internet access. The subject's tasks included assigning channels to stations and changing Internet configurations. The purpose of setting up an assistance scenario was to gain an expert-novice relationship, in which ideally the operators would sometimes also act as the follower, i.e. we were hoping that they may adopt terms introduced by the subjects. The subjects were sitting on a couch in front of the TV set and talked via a hand-held phone to the operator and used a remote control for interacting with the TV set. Ten dialogues were recorded altogether. When the study was finished the dialogues were transcribed.

The first examinations of the transcriptions showed that indeed two basic levels of entrainment ocurred, namely phrasal entrainment and lexical entrainment. An annotation was conducted in order to measure the following aspects: Can entrainment be detected reliably? If yes, which kind of entrainment is it? And who was leader who was follower? For that purpose

[4]For the full description of these studies please see [53].

a manual was created that contained instructions on how to mark the aspects mentioned above. For the annotation, any two consecutive dialogue utterances were coupled. The coupled dialogue utterances were grouped as one *entrainment segment*, encompassing an utterance i and its successor $i + 1$. The next segment than repeats (uses) the successor $i + 1$ as i' with its successor $i' + 1$. Each entrainment segment was to be marked by the operator's role as follower or leader and which kind of entrainment could be detected.

In the first analysis, all entrainment segments were counted in both annotations. As was mentioned in the manual one dialogue entrainment segment in this case is defined as two succeeding operator-subject or subject-operator utterances. Also it was possible for one segment to hold more than one phenomenon that had been entrained, phrases and terms included. During this analysis phrases and terms were not distinguished from one another. Neither were different kinds of entrainment considered. The only thing that was important was if any entrainment phenomenon could be detected for each segment. Table 1 shows the distribution of assigned values (N/NE) in percent. The measured agreement was K = 0.76 using the Kappa coefficient [14], which showed a good reliability in terms of agreement between the annotators according to the interpretation by Altman [4]. As far as phrases are concerned, all occurrences of entrained phrases were counted. Additionally, one of the annotators counted all the phrases that might have been entrained but were not. A phrase was defined as a coherent word-chain that cannot be separated. For phrases percentages are given in Table 1 and the agreement was K = 0.92, which shows an excellent reliability. As for terms, all the terms were counted that had been assigned one of the kinds of lexical entrainment. In order to additionally gain the potentially entrainable terms, a program was written that returned the total number of tokens within the tagged dialogues. However, the different kinds of entrainment were at first not considered because we first aimed at a general result regarding lexical entrainment. The distribution is presented in Table 1. Again the reliability of agreement was excellent, since the Kappa result was K = 0.82.

Additionally to the agreement evaluation a statistical analysis of the dialogue data was calculated based on the annotation results of one of the annotators. The following section provides an overview of how many phrases and terms have been entrained. The sections after that present the evaluation results for different kinds of phrasal and lexical entrainment. The amount of entrained terms and phrases is called the entrainment rate. Additionally the results reveal if the operator was leader or follower when adopting terms.

Here we show the distribution of entrained phrases versus non-entrained

	Annotator 1	Annotator 2
Segment with E	33%	28%
Segment with NE	67%	72%
Phrases with E	7%	6%
Phrases with NE	93%	94%
Terms with E	18%	15%
Terms with NE	82%	85%

Table 1: Annotated Segments, Phrases and Terms

phrases, which could only be evaluated for a random of 50% of the dialogues. The reason for that is that the entire amount of phrases - entrained phrases as well as non-entrained phrases - could only be annotated in five of the dialogues. As Figure 1 shows, phrases were entrained in about 9% of all cases.

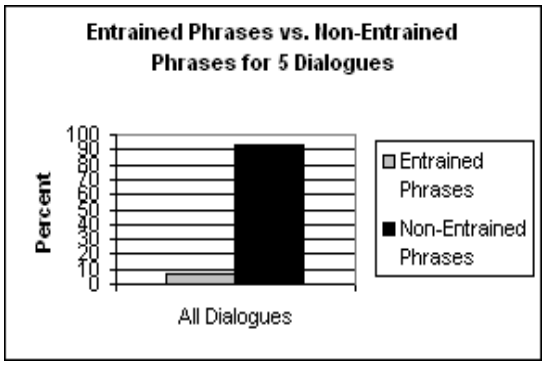

Figure 1: Entrained Phrases vs. Non-Entrained Phrases

On top of that, further comparison between entrained phrases and entrained terms, as presented in Figure 2, affirms this observation on another level: it shows clearly that entrainment occurs a lot more often on a lexical level than on the phrasal one. As for different kinds of entrainment, the statistical analysis showed that ad hoc entrainment occurred more often than later phrasal entrainment.
Figure 3 shows a first overview of how many terms were entrained and how many remained non-entrained. As for each individual dialogue, the results showed that there were some in which the interlocutors entrained very successfully.

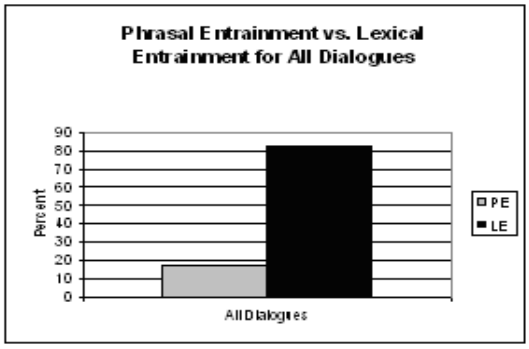

Figure 2: Phrasal Entrainment vs. Lexical Entrainment

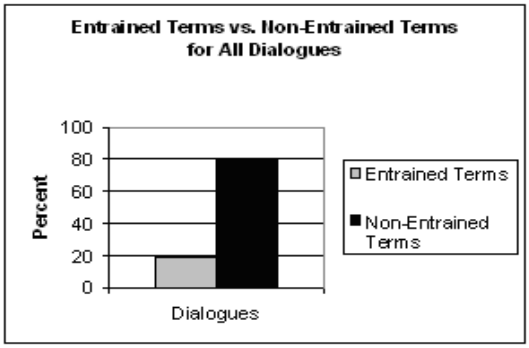

Figure 3: Entrained Terms vs. Non-Entrained Terms

Intuitively, the amount of entrainment within a dialogue can depend on several factors:

- Age of operator and subject
- Profession (i.e. Computer Expert / Novice)
- Psychological factors
  - Cooperative behavior
  - Security/Insecurity of one of the interlocutors
  - The sensibility to detect signs of insecurity
- Conversational flow
- Dialogue length

All of these aspects are closely intertwined with one another and thus influence the amount of entrained terms within a dialogue.

As far as the interlocutors' roles as follower and leader are concerned, Figure 4 shows that the operator was leader in most of the dialogues. In dialogue 7 both operator and subject introduced new terms as well as they adopted terms from their conversational partner at an equal distribution. Dialogue 9 is the only dialogue in which the operator functioned as follower more often than the subject. As always one has to keep in mind that both subjects and operators were in a situation that was imposed on the them - in that very moment the subjects neither had really bought a TV, nor had they really lost the manual. Considering these drawbacks, operators and subjects played their role very well. If one were to truly prove that people entrain in an expert-novice relationship in the same setting, one would have to collect dialogue data from a real call center agent-customer dialogue. Also, people react differently if they know that they are being recorded, since recording causes people either to act more timidly or overeagerly than in situations in which they are not being recorded [58].

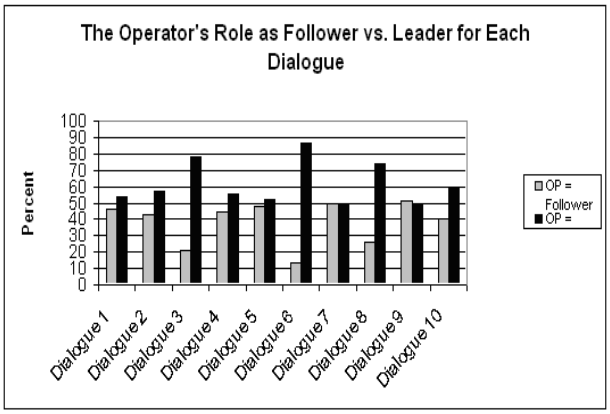

Figure 4: The Operator's Role as Follower vs. Leader

3.1.2 Wizard of Oz Experiment

Based on these and prior [48] empirical examinations of human-human interaction, we performed an entrainment experiment for multimodal human-computer interaction in an assistance setting. The aim of this study was to test the potential effects of entrainment performed by the system the is engaged in the multimodal interaction.

Order	Task
Task 1	Assigning Channels to Stations
Task 2	Accessing the Internet
Task 3	Changing Mouse Speed
Task 4	Changing Font Size in Browser

Table 2: Overview: Tasks in MOS and WoZ Study

Experimental Set-Up: In our experimental setup we created an entraining and a non-entraining *Wizard of Oz* system [25].

- The HILFIX-E system was piloted by a wizard who had to use a fixed set of replies.
- The HILFIX+E system was piloted by a wizard who could entrain towards the user by exchanging parts of the set of fixed replies.

We employed the two mock-up systems with a diverse set of users on the very same tasks, shown in Table 2 as in the MOS Study described in Section 3.1.1. Also the modalities of spoken and remote control interaction that were involved in the human-human study stayed the same. Only this time subjects thought they talked to an actual dialogue system. The system, however, was piloted by an operator, who - after hearing the subject's questions - selected which answer was to be synthesized.

The central task of the operator/wizard, therefore, was to deliver appropriate answers. Half of the subjects used HILFIX-E and the other half HILFIX+E. In the former the answers were derived from the TV manual and in the letter they heard answers, which - despite having the same propositional content as the ones in HILFIX-E - featured an alignment to the subject's lexical and phrasal choices, i.e. entrainment.

Since it was impossible to anticipate all possible particular lexical and phrasal choices of the subjects, the operator/wizard had to insert the appropriate linguistic surface structures on the fly, which called for a special one-way muting device, but did not affect response times, as in both systems identical latency times - corresponding to those of state of the art multimodal systems - were employed.

The results after five subjects using the entraining and another five using the non-entraining system indicate that there is a noticeable speed-up completion time. Looking at all subjects, this amounts to an improvement of

task-completion time by one minute. While this can already be regarded as a good finding, we noticed that the speed-up is even doubled when comparing the non-experts' performance with the experts' as shown in Figure 5. This means that non-experts gained on average two minutes. Experts, however, using the adaptive system were not helped at all, on average they needed even a little longer with the entraining system, even though in this case the sample is definitely too small to make any kind of significance judgment. Clearly not so in the case of the non-experts. Using a PARADISE-like general user-satisfaction questionnaire [63], the adaptive system - as one would expect - scored better in all respects.

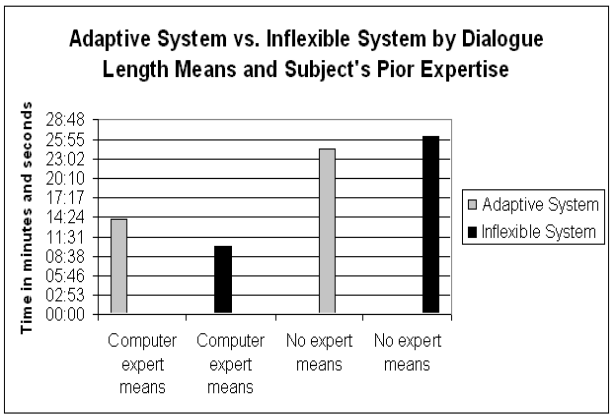

Figure 5: Task Completion Times

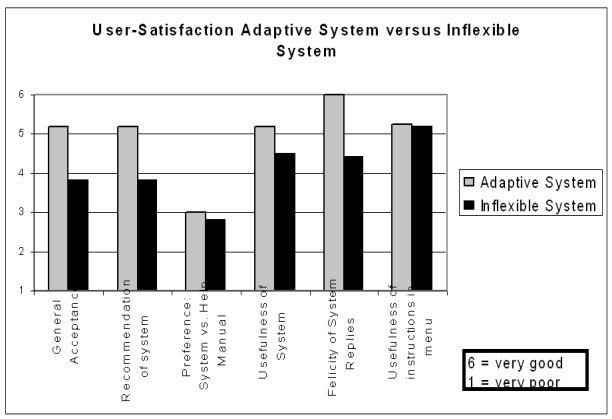

Figure 6: User Satisfaction

Figure 6 shows that, after calculating the means of all user replies, in nearly all cases the subjects preferred the adaptive system rather than the inflexible one. This is also true for the computer experts who solved the task more slowly using the adaptive system than those using the inflexible system. The only two categories that do not show a distinct result is whether people would prefer the help manual over the system and whether they needed the instructions in the help manuals rather than system replies. While the adaptive system shows slightly better results - also in these categories - the difference was slight. The result that stands out most is the felicity regarding system replies. All of the subjects testing the adaptive system rated felicity of system replies by marking down the top score. None of the subjects testing the inflexible system gave the same rating regarding this question.

In these studies we have shown that subjects and operators did entrain despite the fact that the they were put in a situation which was unfamiliar to them within laboratory conditions (where subjects were situated on a couch facing the TV in a usability lab and operators in an office environment). Furthermore, operators had to explain a process they had been taught themselves only minutes before the experiment started. Generally speaking, the results of the Multiple and Operator and Subject study showed - with respect to human-human interaction - that entrainment is not a matter of minor importance. In fact, if operator and subject show a great willingness to align, as was the case in one of the recorded dialogues, the entrainment rate is at 30%. Considering that two people do not constantly repeat each other in a dialogue this rate - as well as the overall average of 20% lexical and 9% phrasal - is rather high. Additionally, our Wizard-of-Oz experiment showed that, especially for domain novices, entrainment behaviour on the computer side increases both measured dialogical efficiency as well as questionnaire-based user satisfaction rates.

3.2 Feedback and Signal Studies

For conducting these experiments we developed a new paradigm for collecting telephone-based dialogue data, called *Wizard and Operator Test* (WOT), which contains elements of both Wizard-of-Oz (WoZ) experiments [25] as well as Hidden Operator Tests [54]. This procedure also represents a simplification of classical end-to-end experiments, as it is - much like WoZ experiments - conductible without the technically very complex use of a real conversational system. As post-experimental interviews showed, this did not limit the feeling of *authenticity* regarding the simulated conversational system by the human subjects (S). The WOT setup consists of two major

phases that begin after subjects have been given a set of tasks to be solved with the telephone-based dialogue system:

- in `Phase 1` the human assistant (A) is acting as a wizard who is simulating the dialogue system, much like in WoZ experiments, by operating a speech synthesis interface,
- in `Phase 2`, which starts immediately after a system breakdown has been simulated by means of beeping noises transmitted via the telephone, the human assistant is acting as a **human** operator asking the subject to continue with the tasks.

This setup enables to control for various factors. Most importantly the technical performance (e.g., latency times), the pragmatic performance (e.g., understanding vs. non-understanding of the user utterances) and the communicative behavior of the simulated systems can be adjusted to resemble that of state of the art dialogue systems. These factors can, of course, also be adjusted to simulate potential future capabilites of dialogue systems and test their effects. The main point of the experimental setup, however, is to enable precise analyses of the differences in the communicative behaviors of the various interlocutors, i.e., human-human, human-computer and computer-human interaction.

During the experiment S and A were in separate rooms. Communication between both was conducted via telephone, i.e., for the user only a telephone was visible next to a radio microphone for the recording of the subject's linguistic expressions. The assistant/operator room featured a telephone as well as two computers - one for the speech synthesis interface and one for collecting all audio streams; also present were loudspeakers for feeding the speech synthesis output into the telephone and a microphone for the recording of the synthesis and operator output. With the help of an audio mixer all linguistic data were recorded time synchronously and stored in one audio file. The assistant/operator acting as the computer system communicated by selecting fitting answers for the subject's request from a prefabricated list which were returned via speech synthesis through the telephone. Beyond that it was possible for the assistant/operator to communicate over telephone directly with the subjects when acting as the human operator.

The experiments were conducted with an English setup, subjects and assistants in the United States of America and with a German setup, subjects and assistants in Germany. Both experiments were otherwise identical and in each 22 sessions were recorded. At the beginning of the WOT, the test

manager told the subjects that they were testing a novel telephone-based dialogue system that supplies touristic information on the city of Heidelberg. In order to avoid the usual paraphrases of tasks worded too specifically, the manager gave the subjects an overall list of 20 very general touristic activities, such as *visit museum* or *eat out*, from which each subject had to pick six tasks which had to be solved in the experiment. The manager then removed the original list, dialed the system's number on the phone and exited from the room after handing over the telephone receiver. The subject was always greeted by the system's standard opening ply: *Welcome to the Heidelberg tourist information system. How I can help you?* After three tasks were finished (some successful some not) the assistant simulated the system's break down and entered the line by saying *Excuse me, something seems to have happened with our system, may I assist you from here on* and finishing the remaining three tasks with the subjects.

The PARADISE framework [62, 63] proposes distinct measurements for dialogue quality, dialogue efficiency and task success metrics. The remaining criterion, i.e., user satisfaction, is based on questionaries and interviews with subjects and cannot be extracted (sub)automatically from log-files. The measurements described herein mainly revolve around dialogue efficency metrics. As we will show below, our findings show that a felicitous dialogue is not only a function of dialogue quality, but critically hinges on a minimal threshold of efficiency and overall dialogue management as well. While these criteria lie orthogonal to the criteria for measuring dialogue quality such as recognition rates and the like [63], we regard them to constitute an integral part of an aggregate view on dialogue quality and efficiency, herein referred to as *dialogue felicity*. For examining dialogue felicity we will provide detailed analyses of efficiency metrics *per se* as well as additional metrics for examining the number and effect of pauses, the employment of feedback and turn-taking signals and the amount of overlaps.

The length of the collected dialogues was on average 5 minutes for the German and 6 minutes for the English sessions.[5] The subjects featured approximately proportional mixtures of gender (25m,18f), age (12< >71) and computer expertise. Table 3 shows the duration and turns per phase of the experiment.

First of all, we applied the classic metric for measuring dialogue efficiency [63], by calculating the number of turns over dialogue length. Figure 7 shows the discrepancy between the dialogue efficiency in `Phase 1` (HHI)

[5]The shortest dialogues were 3:18 (English) and 3:30 (German) and the longest 12:05 (English) and 10:08 (German).

Phase	HHI-G	HHI-E	HCI-G	HCI-E
Average length	1:52 min.	2:30 min.	2:59 min.	3:23 min.
Average turns	11.35	21.25	9.2	7.4

Table 3: Average length and turns in `Phase 1` and `2`

versus `Phase 2` (HCI) of the German experiment and Figure 8 shows that the same patterns can be observed for English.

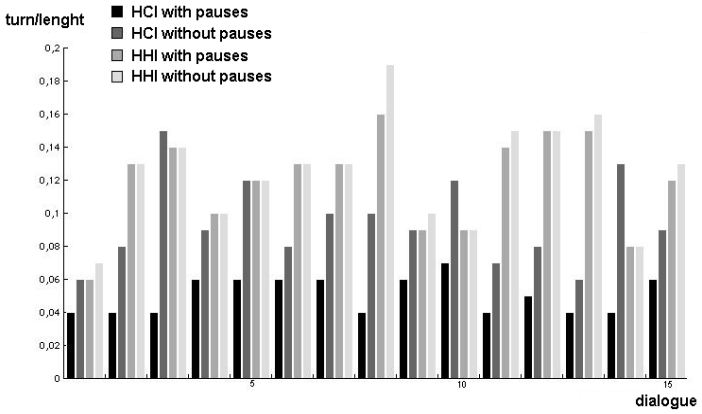

Figure 7: Dialogue efficiency (German data)

As this discrepancy might be accountable by latency times alone, we calculated the same metric with and without pauses. For these analyses, pauses are very conservatively defined as silences during the conversation that exceeded one second. The German results are shown in Figure 9 and, as shown in Figure 10, we find the same patterns hold cross-linguistically in the English experiments. The overall comparison, given in Table 4, shows that - as one would expect - latency times severely decrease dialogue efficiency, but also that they alone do not account for the difference in efficiency between human-human and human-computer interaction. This means that even if latency times were to vanish completely, yielding actual real-time performance, we would still observe less efficient dialogues in HCI.

While it is obvious that the existing latency times increase the number and length of pauses of the computer interactions as compared to the human operator's interactions, there are no such obvious reasons why the

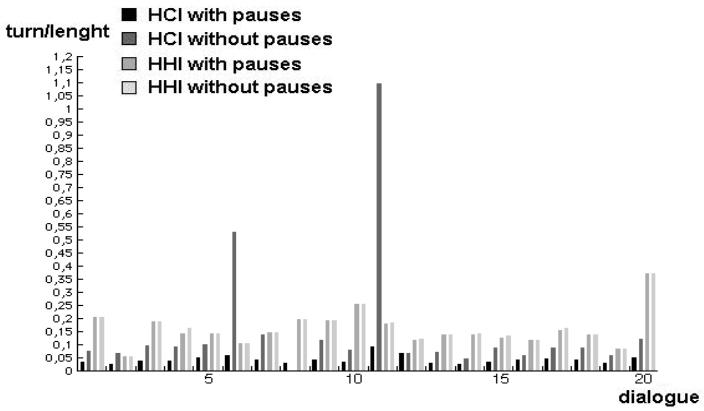

Figure 8: Dialogue efficiency (English data)

number and length of pauses in the human subjects' interactions should differ in `Phase 1` and `Phase 2`. However, as shown in Table 5, they do differ substantially.

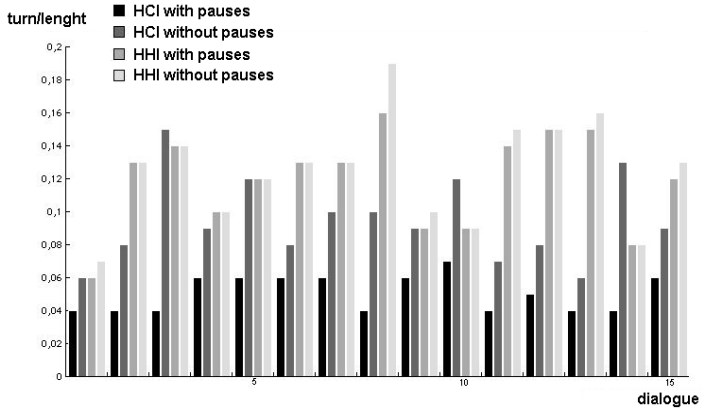

Figure 9: Efficiency w/out latency in German

Next to this *pause-effect*, which contributes greatly to dialogue efficiency metrics by increasing dialogue length, we have to take a closer look at the individual turns and their nature. While some turns carry propositional information and constitute utterances proper, a significant number solely consists of specific particles used to exchange signals between the communicative partners or combinations thereof. We differentiate between dialogue-structuring signals and feedback signals [68]. Dialogue-structuring signals -

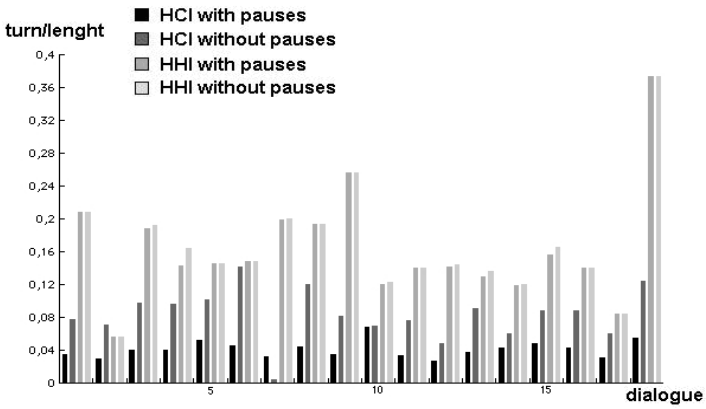

Figure 10: Efficiency w/out latency in English

Efficiency	HCI -p	HCI +p	HHI -p	HHI +p
Mean German	0.18	0.05	0.25	0.12
Standard-deviation	0,04	0,01	0.06	0.03
Mean English	0.16	0.05	0.17	0.17
Standard-deviation	0.25	0.02	0.07	0.07

Table 4: Overall dialogue efficiencies with pauses +p and without pauses -p

such as hesitations like *hmm* or *ah* as well as expressions like *well*, *yes*, *so* - mark the intent to begin or end an utterances, make corrections or insertions. Feedback signals- while sometimes phonetically alike - such as *right*, *yes* or *hmm* - do not express the intent to take over or give up the speaking role, but rather serve as a means to stay in contact with the speaker, which is why they are sometimes referred to as *contact signals*.

In order to be able to differentiate between the two, for example, between an agreeing feedback *yes* and a dialogue-structuring one, all dialogues were annotated manually. The resulting counts for the user utterances in `Phase 1` and `2` are shown in Table 6. Not shown in Table 6 are the number of particles employed by the computer, since it is zero, and those of the human operator in the HHI dialogues, as they are like those of his human interlocutor.

Pauses	HCI-G	HHI-G	HCI-E	HHI-E
Number total	79	10	94	21
Number per dialog	3.95	0.5	4.7	1.05
Number per turn	0.46	0.05	0.64	0.05
total length	336sec	19sec	467sec	48sec
% of phase	9.37	0.84	13.74	1.75
% of dialogue	5.75	0.3	7.46	0.766

Table 5: Overall pauses of human subjects: `Phase 1` and `2` German (HCI-G/HHI-G) and English (HCI-G/HCI-E)

Particles	structure HCI	particle HHI	feedback HCI	particle HHI
Number total	112 G 90 E	225 G 202 E	18 G 0 E	135 G 43 E
per dialogue	5.6 G 4.5 E	11.25 G 10.1 E	0.9 G 0 E	6.75 G 2.15 E
per turn	0.4 G 0.61 E	0.59 G 0.48 E	0.04 G 0 E	0.26 G 0.1 E

Table 6: Particles of human subjects: HCI vs. HHI

Again, the findings for both German and English are congruent. We find that feedback particles almost vanish from the human-computer dialogues - a finding that corresponds to those described in Section 2. This linguistic behavior, in turn, constitutes an adaptation to the employment of such particles by that of the respective interlocutor. Striking, however, is that the human subjects still attempted to send dialogue structuring signals to the computer, which - unfortunately - would have been ignored by today's "conversational" dialogue systems.[6]

[6]In the English data the subject's employment of dialogue structuring particles in HCI even slightly surpassed that of HHI.

Before turning towards an analysis of this data we will examine the overlaps that occurred throughout the dialogues. Most overlaps in human-human conversation occur during turn changes with the remainder being feedback signals that are uttered during the other interlocutor's turn [35]. The results on measuring the amount of overlap in our experiments are given in Table 7. Overall the HHI dialogues featured significantly more overlap than the HCI ones, which is partly due to the respective presence and absence of feedback signals as well as due to the fact that in HCI turn takes are accompanied by pauses rather than immediate - overlapping - hand overs.

Overlaps	HCI-G	HHI-G	HCI-E	HHI-E
Number total	7	49	4	88
per dialogue	0.35	3.06	0.2	4.4
per turn	0.03	0.18	0.01	0.1

Table 7: Overlaps in Phase 1 versus Phase 2

Lastly, our experiments yielded negative findings concerning the type-token ratio and syntax. This means that there was no statistically significant difference in the linguistic behavior with respect to these factors. We regard this finding to strengthen our conclusions, that to emulate human syntactic and semantic behavior does not suffice to guarantee effective and therefore felicitous human-computer interaction.

The results presented above enable a closer look at dialogue efficiency as one of the key factors influencing overall dialogue felicity. As our experiments show, the difference between the human-human efficiency and that of the human-computer dialogues is not solely due to the computer's response times. There is a significant amount of *white noise*, for example, as users wait after the computer has finished responding. We see these behaviors as a result of a mismanaged dialogue. In many cases users are simple unsure whether the system's turn has ended or not and consequently wait much longer than necessary.

The situation is equally bad at the other end of the turn taking spectrum, i.e., after a user has handed over the turn to the computer, there is no signal or acknowledgment that the computer has taken on the baton and is running along with it - regardless of whether the user's utterance is understood or not. Insecurities regarding the main question, i.e., *whose turn is it anyways*, become very notable when users try to establish contact, e.g., by saying

hello -pause- *hello*. This kind of behavior certainly does not happen in HHI, even when we find long silences.

Examining why silences in human-human interaction are unproblematic, we find that, these silences have been announced, e.g., by the human operator employing linguistic signals, such as *just a moment please* or *well, I'll have to have a look in our database* in order to communicate that he is holding on to the turn and finishing his round.

To push the relay analogy even further, we can look at the differences in overlap as another indication of crucial dialogue inefficiency. Since most overlaps occur at the turn boundaries and, thusly, ensure a smooth (and fast) hand over, their absence constitutes another indication why we are far from having winning systems.

As the primary effects of the human-directed language exhibited by today's conversational dialogue systems, our experiments show that:

- dialogue efficiency decreases significantly even beyond the effects caused by latency times,
- the human interlocutor ceases in the production of feedback signals, but still attempts to use his or her turn signals for marking turn boundaries - which, however, remain ignored by the system,
- the increases in the amount of pauses is caused by waiting- and uncertainty-effects, which are also manifested by missing overlaps at turn boundaries.

Generally, we can conclude that a felicitous dialogue needs some amount of extra-propositional exchange between the interlocutors. The complete absence of such dialogue controlling mechanisms - by the non-human interlocutors alone - literally causes the dialogical situation to get out of control, as observable in the turn-taking and -overtaking phenomena described in Section 2. As witnessable in recent evaluations, this way of behaving does not serve the intended end, i.e., efficient, intuitive and felicitous human-computer interaction.

4 Towards Pragmatic Alignment

We see the results of the aforementioned studies to contribute part of an emerging picture that shows how interlocutors employ a variety of linguistic or paralinguistic instruments to make dialogues efficient, align to their interlocutors and, thereby, guarantee their felicity. One way of looking at

this ensemble of instruments is to view them as means for *pragmatic alignment*. We motivate the choice of the term *pragmatic* by the fact that these instruments exhibit both a discourse functional dimension - beyond that of the morpho-syntactic and semantic levels as well as by the fact that they are, by their very nature, context-dependent.

Therefore, we have to take a look at two fundamental, but notoriously tricky, notions for human-computer interface systems, which frequently are regarded as one of the central problems facing both applications in artificial intelligence and natural language processing. These, often conflated, notions are those of *context* and *pragmatics*. Indeed, in many ways both notions are inseparable from each other if one defines pragmatics to be about the ways of encoding and decoding of meaning in discourse, which, as pointed out numerously [12, 65, 50], is always context-dependent. This, therefore, entails that pragmatic inferences (also called *pragmatic analyses* [12]) are impossible without recourse to contextual observations. In a sense suprisingly[7], the context-dependency of these features elevates their status from mere automatically produced garnishings of a given discourse to the level of flexibly employed workhorses thereof.

In order to address how computers should talk to humans we face two corresponding challenges:

- how to enable to encode the computer's internal processing and stance to their human interlocutors in order to avoid phenomena discussed above such as turn-overtaking, dialogical inefficiency and general dissatisfaction;

- how to decode these signals and adaptations provided by their human interlocutors in order to understand them better, manage natural turn-taking and react felicitously.

Last but not least, the distinction between pragmatic knowledge - which is learned/acquired - and contextual information - which is observed/inferred - is also of paramount importance in designing scalable context-adaptive systems, which seek to align to their human users and, thereby, to (inter)act felicitously with them.

[7]Surprising as these central and functionally critical features of discourse have been by and large overlooked in the design and development of dialogue systems.

References

[1] J. Alexandersson and T. Becker. Overlay as the basic operation for discourse processing. In *Proceedings of IJCAI*. Springer-Verlag, 2001.

[2] J. Allen, G. Ferguson, and A. Stent. An architecture for more realistic conversational system. In *Proceedings of Intelligent User Interfaces*, pages 1–8, Santa Fe, NM, 2001.

[3] J. F. Allen, B. Miller, E. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In *Proc. of ACL-96*, 1996.

[4] D. Altman. *Practical Statistics for Medical Research*. Oxford University Press, Oxford, 1990.

[5] H. Aust, M. Oerder, F. Seide, and V. Steinbiss. The Philips automatic train timetable information system. *Speech Communication*, 17:249–262, 1995.

[6] G. Bailly, N. Campbell, and B. Möbius. Isca special session: Hot topics in speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[7] N. Beringer. The SmartKom Multimodal Corpus - Data Collection and End-to-End Evaluation. In *Colloquium of the Department of Linguistics*, University of Nijmwegen, June 2003.

[8] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk. PROMISE: A Procedure for Multimodal Interactive System Evaluation. In *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, Spain, 2002.

[9] S. Brennan. Lexical entrainment in spontaneous dialogue. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 41–44, Philadelphia, USA, 1996.

[10] S. Brennan. Processes that shape conversation and their implications for computational linguistics. In *Proceedings of ACL*, Hong Kong, 2000.

[11] S. E. Brennan. Centering as a psychological resource for achieving joint reference in spontaneous discourse. In M. Walker, A. Joshi, and E. Prince, editors, *Centering in Discourse*, pages 227–249. Oxford University Press, Oxford, U.K., 1998.

[12] H. Bunt. Dialogue pragmatics and context specification. In *Computational Pragmatics, Abduction, Belief and Context*. John Benjamins, 2000.

[13] B. Bürkle. "von mir aus ..." Zur hörerbezogenen lokalen Referenz. Technical Report Bericht 10, Forschergruppe "Sprachen und Sprachverstehen im sozialen Kontext, 1986.

[14] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[15] H. H. Clark and C. Marshall. Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, editors, *Linguistic Structure and Discourse Setting*. Cambridge University Press, 1981.

[16] M. G. Core and J. D. Moore. Robustness versus fidelity in natural language understanding. In R. Porzel, editor, *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 1–8, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[17] R. Cox, C. Kamm, L. Rabiner, J. Schroeter, and J. Wilpon. Speech and language processing for next-millenium communications services. *Proceedings of the IEEE*, 88(8):1314–1334, 2000.

[18] C. Darves and S. Oviatt. Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, U.S.A., 2002.

[19] J. Diaz-Verdejo, R. Lopez-Cozar, A. Rubio, and A. D. la Torre. Evaluation of a dialogue system based on a generic model that combines robust speech understanding and mixed-initiative control. In *2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

[20] S. Duncan. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3, 1974.

[21] C. Elting, J. Zwickel, and R. Malaka. Device-dependant modality selection for user-interfaces - an empirical study. In *Proceedings of International Conference on Intelligent User Interfaces (IUI'02)*, San Francisco, CA, January 2002. Distinguished Paper Award.

[22] R. Engel. SPIN: Language understanding for spoken dialogue systems using a production system approach. In *Proceedings of ICSLP 2002*, 2002.

[23] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass., 1998.

[24] K. Fischer. *What Computer Talk Is and Isn't: Human-Computer Conversation as Intercultural Communication*. Saarbrücken: AQ, 2006.

[25] J.-M. Francony, E. Kuijpers, and Y. Polity. Towards a methodology for wizard of oz experiments. In *Third Conference on Applied Natural Language Processing*, Trento, Italy, March 1992.

[26] N. Fraser. Sublanguage, register and natural language interfaces. *Interacting with Computers*, 5, 1993.

[27] G. Furnas, T. Landauer, and G. Dumais:. The vocabulary problem in human-system-communication: an analysis and a solution. *Communications of the ACM*, 30(11):964–971, 1987.

[28] F. Gallwitz, M. Aretoulaki, M. Boros, J. Haas, S. Harbeck, R. Huber, H. Niemann, and E. Nöth. The Erlangen spoken dialogue system EVAR: A state-of-the-art information retrieval system. In *Proceedings of 1998 International Symposium on Spoken Dialogue (ISSD 98),* Sydney, Australia, 30. Nov., 1998, pages 19–26, 1998.

[29] S. Garrod and A. Anderson. Saying what you mean in dialog: A study in conceptual and semantic co-ordination. *Cognition*, 27, 1987.

[30] A. L. Gorin, G. Riccardi, and J. H. Wright. How may I help you? *Speech Communication*, 23:113–127, 1997.

[31] I. Gurevych, R. Porzel, and S. Merten. Less is more: Using a single knowledge representation in dialogue systems. In *Proceedings of the HLT/NAACL Text Meaning Workshop*, Edmonton, Canada, 2003.

[32] T. Herrmann and J. Grabowski. *Sprechen. Psychologie der Sprachproduktion*. Spektrum Akademischer Verlag, 1994.

[33] R. Higashinaka, N. Miyazaki, M. Nakano, and K. Aikawa. Evaluating discourse understanding in spoken dialogue systems. In *Proceedings of Eurospeech*, pages 1941–1944, Geneva, Switzerland, 2003.

[34] A. Jameson and W. Wahlster. User modelling in anaphora generation: Ellipsis and definite description. In *Proceedings of the European Conference on Artificial Intelligence (ECAI '82), 1982*, pages 222–227, 1982.

[35] G. Jefferson. Two explorations of the organistion if overlapping talk in coversation. *Tilburg Papers in Language and Literature*, 28, 1983.

[36] M. Johnston. Unification-based multimodal parsing. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, 1998.

[37] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. Match: An architecture for multimodal dialogue systems. In *Proceedings of ACL'02*, pages 376–383, 2002.

[38] R. Kass and T. Finin. Modeling the user in natural language systems. *Computational Linguistics*, 14(3):5–22, 1988.

[39] D. Kingsbury. *Unpublished Honor Thesis*. PhD thesis, Havard University, 1968.

[40] H. Kitzenberger. Unterschiede zwischen mensch-computer-interaktion und zwischenmenschlicher kommunikation aus der interpretativen analyse der dicos-protokolle. In J. Krause and L. Hitzenberger, editors, *Computer Talk*, pages 122–156. Olms, Hildesheim, 1992.

[41] J. Krause. Natürlichsprachliche mensch-computer-interaktion als technisierte kommunikation: Die computer talk-hypothese. In J. Krause and L. Hitzenberger, editors, *Computer Talk*. Olms, Hildesheim, 1992.

[42] R. Krauss. The role of the listener: Addressee influences on message formulation. *Journal of Language and Social Psychology*, 6:91–98, 1987.

[43] R. Krauss and S. Weinheimer. Changes in the lenght of reference phrases as a function of social interaction: A preliminary study. *Psychonomic Science*, (1):113–114, 1964.

[44] R. Krauss, S. Weinheimer, and S. More. The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, (9):523–529, 1977.

[45] S. Nariyama. Pragmatic information extraction from subject ellipsis in informal english. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 1–8, New York City, New York, June 2006. Association for Computational Linguistics.

[46] T. Paek. Empirical methods for evaluating dialog systems. In *Proceedings 2nd SIGdial Workshop on Discourse and Dialogue*, pages 100–107, Aalborg, Denmark, 2001.

[47] C. L. Paris. *User Modeling in Text Generation*. Pinter, London, 1993.

[48] R. Porzel and M. Baudis. The Tao of CHI: Towards effective human-computer interaction. In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 209–216, Boston, Massachusetts, USA, May 2 - May 7 2004.

[49] R. Porzel and I. Gurevych. Contextual coherence in natural language processing. In P. Blackburn, C. Ghidini, R. Turner, and F. Giunchiglia, editors, *Fourth International Conference on Modeling and Using Context*, Berlin, 2003. Springer (LNAI 2680).

[50] R. Porzel and I. Gurevych. *Contextual Coherence in Natural Language Processing*. LNAI 2680, Sprigner, Berlin, 2003.

[51] R. Porzel, N. Pfleger, S. Merten, M. Löckelt, R. Engel, I. Gurevych, and J. Alexandersson. More on less: Further applications of ontologies in multi-modal dialogue systems. In *Proceedings of the 3rd IJCAI 2003 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Acapulco, Mexico, 2003.

[52] R. Porzel, A. Schaffler, and R. Malaka. How entrainment increases dialogical efficiency. In *Workshop on on Effective Multimodal Dialogue Interfaces, Sydney, January, 29th,*, 2006.

[53] R. Porzel, A. Scheffler, and R. Malaka. How entrainment increases dialogical effectiveness. In *Proceedings of the IUI'06 Workshop on Effective Multimodal Dialogue Interaction*, Sydney, Australia, 2006.

[54] S. Rapp and M. Strube. An iterative data collection approach for multimodal dialogue systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 2002.

[55] S. Rapp, S. Torge, S. Goronzy, and R. Kompe. Dynamic speech interfaces. In *Proceedings of 14th ECAI WS-AIMS*, 2000.

[56] E. Rich. User modeling via stereotypes. *Cognitive Science*, 3:329–354, 1979.

[57] S. Sack, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 1974.

[58] J. Schu. Formen der Elizitation und das Problem der Natürlichkeit von Gesprächen. In K. Brinker, G. Antos, W. Heinemann, and S. Sagere, editors, *Text- und Gesprchslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*, pages 1013–1021. Springer, 2001.

[59] T. R. Shankar, M. VanKleek, A. Vicente, and B. K. Smith. A computer mediated conversational system that supports turn negotiation. In *Proceedings of the Hawai'i International Conference on System Sciences*, Maui, Hawaii, January 2000.

[60] E. Sweetser. Levels of meaning in speech and gesture: Real space mapped onto epistemic and speech-interactional mental spaces. In *Proceedings of the 8th International Conference on Cognitive Linguistics*, Logrono, Spain, July 2003.

[61] W. Wahlster, N. Reithinger, and A. Blocher. Smartkom: Multimodal communication with a life-like character. In *Procedings of the 7th Eurospeech.*, 2001.

[62] M. Walker, D. Litman, C. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, 1997.

[63] M. A. Walker, C. A. Kamm, and D. J. Litman. Towards developing general model of usability with PARADISE. *Natural Language Engeneering*, 6, 2000.

[64] K. Weinhammer and S. Rabold. Durational Aspects in Turn Taking. In *Proceedings of International Conference Phonetic Sciences*, Barcelona, Spain, 2003.

[65] D. Widdows. *A Mathematical Model of Context.* LNAI 2680, Sprigner, Berlin, 2003.

[66] R. Woodburn, R. Procter, J. Arnott, and A. Newell. A study of conversational turn-taking in a communication aid for the disabled. In

People and Computers, pages 359–371. Cambridge University Press, Cambridge, 1991.

[67] R. Wooffitt, N. Gilbert, N. Fraser, and S. McGlashan. *Humans, Computers and Wizards: Conversation Analysis and Human (Simulated) Computer Interaction.* Brunner-Routledge, London, 1997.

[68] V. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, Chicago, Illinois, April 1970.

[69] M. Zoeppritz. Computer talk? Technical report, IBM Scientific Center Heidelberg Technical Report 85.05, 1985.

Analyses of Feedback in HRI

Britta Wrede, Stephan Buschkaemper,
Claudia Muhl and Katharina J. Rohlfing
University of Bielefeld, Germany
bwrede,cmuhl,rohlfing@TechFak.Uni-Bielefeld.de
sbuschkaemper@Uni-Bielefeld.de

Abstract

Feedback is one of the crucial components of dialogue which allows the interlocutors to align their internal states and assessments of the ongoing communication. Yet, due to technical limitations, immediate and adequate feedback is still a challenge in artificial systems and, therefore, causes manifold problems in human-robot interactions (HRI). Our starting point is the assumption that the manner and content of the feedback, that robots currently are able to provide, often disturbs the flow of communication and that such disruptions may impact the affective evaluation of the users towards the robot. In our study we therefore analysed quantatively how different feedback behavior of the robot resulted in different affective evaluations. In a subsequent qualitative analysis we looked at how the different feedbacks actually affected the communicational flow in detail and produced hypotheses on how this might influence the interaction and thus the affective evaluation. Based on these analyses we conclude with hypotheses about the implications for the design of feedback.

1 Introduction

One central assumption in social robotics states that if users are to accept robots in their private lives, robots need to blend in the social situation and act according to social rules. This means that embedded in social situations, a robot is not only situated in an environment with humans and can interact with the other agents [7], but is also designed to respect the rules of dialogue. The first ability, to blend in the social situation, is known as "social embeddedness" [9], while the second ability, to respect the rules of dialogue, is also referred to as "interaction awareness" [7]). Yet, in a natural interaction, the two abilities are interweaved: If a robot respects the rules of

a dialogue, it will more likely be embedded in social situation; a socially embedded robot has to act according to "human interactional structures" [7]. A phenomenon that combines the two aspects in an natural interaction is feedback. Feedback is a response signaling an immediate result in the environment which in turn can be used as a basis for another more adapted behavior. This way, a basic pattern of interaction depending upon mutual monitoring can emerge and creates a social interaction (cf. [14]). In our approach, we pursued the question of which factors may be crucial for the two central abilities of a social robot, social embeddedness and interaction awareness, and how they can be used to design feedback undesirable for a successful communication.

To design a robot that is able to blend in a social situation, factors like anthropomorphism [8], [22] and perceived personality [26] have been discussed. In our study, we accessed the quantitative correlations between the robot's behavior and the user's reaction by asking how users perceive the personality of a robot they have been interacting with in a non-restricted situation. The underlying scenario for which our robot is designed mainly consists of showing and explaining locations and objects to a robot in a home-like environment. The goal is to teach the robot enough knowledge in order to enable it to autonomously navigate to perform fetch and carry jobs or basic object manipulation tasks such as laying out the table. In such a scenario, the initiative is mainly with the user, however the degree of initiative taking of the robot may vary and thus be used as a cue to convey different robot personalities or may otherwise affect the users' evaluation of the robot. In our study, we varied initiative taking behavior of the robot and analyzed the effects this had on the users' perception. In detail, we addressed the following three questions: (1) If asked to describe a robot's personality with traits established in personality psychology, how easy do users find this task and how sure are they about their judgment? (2) Does the robot's initiative taking behavior influence the perceived personality? (3) Which factors are relevant for the affective evaluation of the robot?

Based on these results, we attempted to explain how can a robot act according to social rules. In our approach, we applied these quantitative findings to qualitative analyses based on methods derived from sociology. We assessed the situative factors of the communication by applying ethnomethodological conversation analysis to each interaction and by characterizing the given feedback by evaluating users strategies and difficulties in keeping the communication in its flow. Pursuant to social constructionism, individuals actively participate in the creation of their perceived reality. Accordingly, social situations consist of mutual processes of attribution and the

ascription of meaning. That there is to be some kind of feedback, is part of the actors expectancy of communication settings. If there is not, the expected indicator is not given and hinders the follow up. We reveal strategies of users' elaborations that substantiate social expectation in communicative processes. These identified basic interaction patterns in HRI seem to be close to social communication practices in human-human settings.

Our results demonstrate the synergetic effects of the combination of quantitative and qualitative analysis: the combined analysis allows us to formulate hypotheses as to why users rate the robot and its interaction in a certain way. In detail, we present hypotheses on what situative and personal factors influence the interaction and what kind of feedback is necessary for a successful complex human-robot interaction.

2 The Robot System

The basis of our data collection is a user study carried out with our mobile robot BIRON [12] (Bielefeld Robot Companion), an interactive robot based on an ActiveMedia PeopleBot platform. This robot is able to carry out multi-modal interactions with a user fully autonmously. The main component is a person attention system [15] which enables the robot to focus its attention on a person. Based on this attention system the robot can physically follow the person of interest and engage in verbal interactions. A multi-modal object attention module allows the system to learn new objects shown by the user. The verbal interactions are carried out by a dialog component [17] that enables mixed-initiative interactions. For the purpose of this study we disabled BIRON's mobility so that it remained fixed on its place. Note that the complete software also runs on our anthropomrophic robot BARTHOC [13] where further interaction capabilities such as facial expressions are currently being integrated. However, at the time of the study the running system was only available for the mobile robot BIRON.

The different initiative behaviors of BIRON are realized by the dialog system. In contrast to many finite-state based dialog systems for robots ([20, 4, 1]) we implemented a powerful computational model of multi-modal grounding [16] which is based on the grounding idea of Clark [6] and extends the conventional computational grounding models of Traum et al. [24] and Cahn et al. [5].

For the personality study we implemented two dialog modes for BIRON (a high-initiative mode and a low-initiative mode) that differ from each other in whether or not the robot initiates the interaction with the user

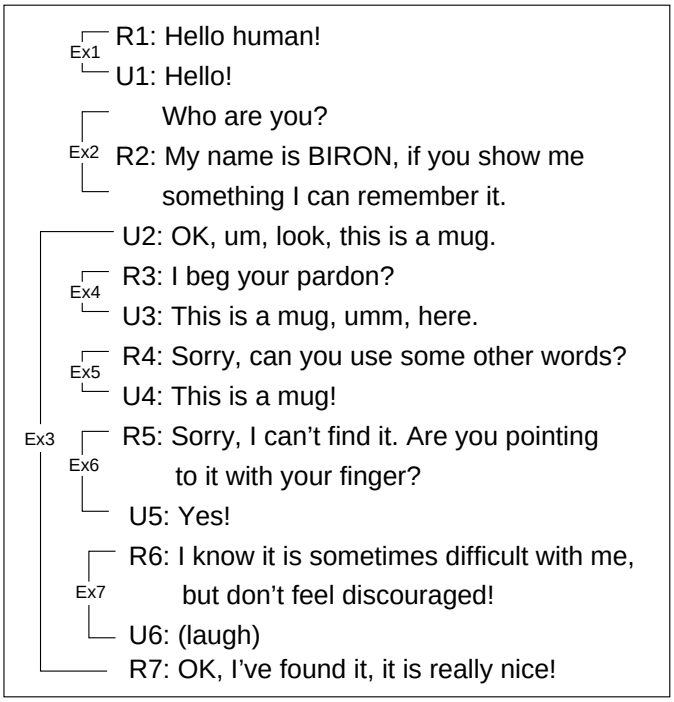

Figure 1: Taking initiatives (Ex: exchange, U: user, R: robot)

and makes comments on the ongoing interaction based on its evaluation of the communicative success as shown in Table 1. Figure 1 presents a dialog example from a user interaction with the high-initiative version of BIRON. In Ex1, BIRON actively greets a person once it detects her and in Ex6 it makes remarks on its own poor performance. The low-initiative BIRON does not have these two capabilities. The technical realization of them is described in detail in [17].

3 Data Collection

For the data collection we used a between-subject design with a total of 14 users aged between 25 and 37 years interacting with BIRON. Each subject had to go through two subsequent interaction sessions. In the first warm-up session the users were asked to familiarize themselves with the robot by asking questions about its capabilities upon which the robot would give a short explanation ("You can show me objects and locations") and the users would start showing objects. Before the second session the users were given

Feedback in case of..	High Init.	Low Init.
User command	+	+
User query	+	+
Error messages from system	+	+
Seeing human	+	-
Well going interaction	+	-
Badly going interaction	+	-

Table 1: Feedback behavior of system with high initiative ('High Init.') vs system with low initiative ('Low Init.')

more technical information about the details of the underlying functionality in order to minimize technical failures which can occur when users do not stand still or do not look into the robot's camera while speaking etc. These instructions were intended to help to reduce perception errors of the system and to make users feel more comfortable during the interaction. Then the subjects were given the instruction to show specific objects to the robot. The mean interaction time of each session was about 10 minutes, yielding an overall interaction time of about 20 minutes per subject. After the second session the users completed a set of questionnaires regarding their judgment of the interaction as well as ratings of the perceived personality of the robot, of their own personality and on how much they liked the robot. The personality of the robot and the user were each assessed by a time-economic questionnaire, the BFI-10 [23], which measures personality according to the widely accepted and cross-culturally [10] as well as more or less even cross-speciesly [25] applicable Big Five Model of personality [21]. Furthermore after rating the robot's personality users were asked how easy the task of judging BIRON's personality was and how sure they felt about their judgement. Each of these questions was answered by a 5-point verbal rating scale with 'very easy' / 'very sure' and 'very difficult' / 'not sure at all' as the extreme anchor points. As an affective evaluation of the interaction users were asked if they liked BIRON, this question was to be answered with a simple 'yes' or 'no'.

For the qualitative analysis, the interactions were video taped and later analyzed in detail.

In order to assess the influence of different initiative-taking behaviors of the robot on its perceived personality we used two different interaction types of the dialog system that were randomly distributed over the subjects. In the

low initiative interaction type the robot only gives feedback when addressed by the user. Only in case of errors the robot takes the initiative and reports them to the user. In contrast, the pro-active interaction type will actively engage in a conversation by issuing a greeting when it detects a person facing the robot. It will also give comments relating to the success of the communication at certain points during the interaction (e.g. "It's really fun doing interaction with you" or "I know it's sometimes difficult with me, but please don't feel discouraged"). Note that in contrast to other studies on the perception of artifical agent's personality we use an interactive cue that is not pre-programmed but depends on the actual interaction situation and thus takes the user in the loop as an active interaction partner into account.

4 Quantitative Study on Personality

In this section we report on some quantitative findings from the questionnaire study on the perceived personality of BIRON. In general the subjects reported to feel 'very sure' (71.4%) about their judgements concerning BIRON's personality. Also, most of them (57.1%) thought the task of answering the personality items was 'very easy' or 'rather easy'.

Users interacting with the pro-active interaction type of BIRON rated the robot significantly higher on extraversion than users interacting with the low initiative version (t-test for independent samples: $p < .05$, see Fig. 2). Interestingly, the pro-active version of the robot might also provoke more heterogenous personality judgments than the less initiative version. The standard deviations of the ratings of the robot's personality traits were larger by 1.11 to 3.02 times in the user group interacting with the more initiative version than in the user group interacting with the less initiative version of BIRON.

The third research question we addressed was, which factors might influence the affective evaluation of the users concerning BIRON. Overall 57.1% of the users answered that they liked BIRON. Most interestingly it turned out that in the group of users interacting with the pro-active version of BIRON 85.5% liked the robot, while this was only the case for 28.6% of the users interacting with the less initiative version. The correlation of $r = .577$ ($p < .05$) indicates that 33.3% of the variance in the users' answers concerning this question could be explained by the root's interaction behavior. In short, there was a significant and strong tendency of the pro-active version being preferred by the users over the less initiative version.

However, while this quantitative analysis provides us with a good basis

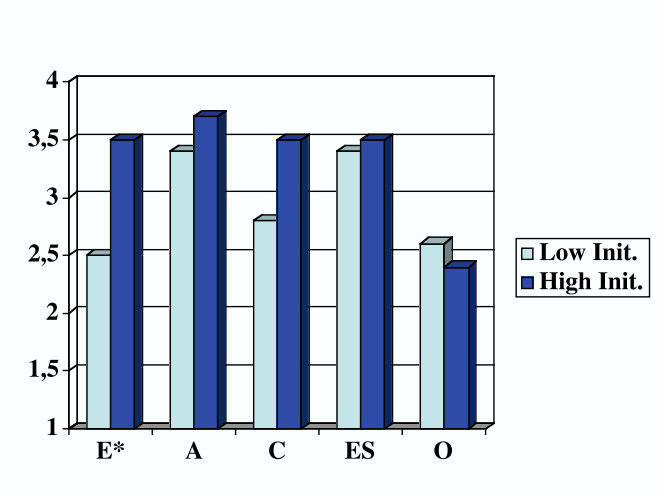

Figure 2: Personality ratings of users interacting with robot with high vs. low initiative interaction behavior. Star marks significant difference between the two settings. (E: Extraversion, A: Agreeableness, C: Conscientiousness, ES: Emotional Stability, O: Openness to Experience)

for statistical correlations it can not answer the question *why* users tend to prefer the extroverted behavior. Thus, in order to produce more concrete hypotheses about this question we performed a qualitative analysis of the interactions which is described in the following section.

5 Qualitative Analysis of Feedback in HRI

Our basic assumption is that users will prefer a robot when they perceive its behaviour as social. But what does it mean for a robot to act according to social rules? In order to concretize the social phenomena and special character of an interaction situation and to explicitly frame the constraints and context of information given, we analyzed interactions with BIRON from a sociological point of view. As methodological approach we employed ethnomethodological conversation analysis techniques. The empirical case study is presented in the following section where some findings and also interpretations are given.

We apply the social constructionism and Niklas Luhmanns systems theory as theoretical frame for HRI in a sociological perspective. There, com-

munication is seen as the central vehicle establishing social relations. But how do people act in the face of a non-human interaction partner? How do people adjust their interaction in these specific human-robot settings? Do they establish relevant patterns of behavior? The focused phenomenon chosen for our analysis upon HRI is the variability of *feedback*. Our case studies bring to light that the users interprete the context and syntonize their performances according to their interpretation of the situation.

5.1 Theoretical Framing - Constructionism and Systems Theory

The paradigm of social constructionism (a theory of knowledge) as developed in the 1960s (e.g. [2]), anticipates that there is not *one* single and true reality, but the world consists of subjective constructions of the perceived phenomena made by subjects. According to this, dealing with *reality* means that the individual always refers to its own perceptions (which evidently differ from each other). From this follows for interactions that the interpretations of the communicating partner's actions and his decisions that keep the conversation going are context-driven, situative and individual.

5.1.1 Constructionist Prerequisites

From the perspective of the social constructionism, a situation is built up in a human's mind from variables as context, knowledge and the ascription of meaning (e.g. the specific cultural background). Thus, social reality is a dynamical construction made and renewed by practical acting. As such, each action has to be understood as communication practice and, vice versa, communicating is a constructive action.

5.1.2 Systems of Communication

The sociologist Niklas Luhmann's systems theory describes the functional differenciation of society. In these terms, modern societies build up a web of distributed functionalities [19]. A social system's main function is to lead and organize interactions. Accordingly, the main operation is the attempt to *understand* the other's communicative distributions, and to assign some discourse elements as well. This operation is much more complex, than it might appear. Communication consists of the triad information, message and understanding [19]. Which means that it is not evident to access a simple transfer of facts but a communication consists of the longing for 'accessibility' in several dimensions.

5.1.3 Systems of Interaction

A social system according to Luhmann is built on coordinated actions of several persons [18]. Because social systems can be characterized mainly by their communicative procedure [19], social systems are systems of communication. While action is constituted by processes of attribution, cognition has a high impact on interaction proceedings.

5.2 Adapting Sociological Systems Theory to HRI

In this paper we discuss feedback as a problem of expectation. Drawing decisions about the next action is a kind of selection which refers to former action-decision settings. Concrete actions reduce the complexity of all possible actions by means of attribution and expectation. Generalizing the own intentions leads to expectations that lower the world's complexity: if to my thoughs, there is only one possibility to behave, I can await its appearance and in any other case, decline all not expected operations. The interaction itself is an operation of registering the operations of others and comparing them to one's own suggestions which leads to concrete decision taking and further actions. Systems of interaction are interrelating constructions driven by expectance and estimation. HRI deals with the overall problem of communication in a specific context. The reciprocal setting of interrelated expectations differs from a sheer humanoid interaction where both partners tend to interprete the other's actions flexibly. In case of interacting with a robot, we have to ask what is *social* about the situation and what is special in the human's behaviour? The general strategy in lowering the costs of interacting in HRI is to implement dialogue strategies that match human speech behavior as much as possible. Feedback plays an important role in the attempt of understanding as it serves as checkback signal for both counterparts. Since strategies of interaction are revealing social expectation in communicative processes the aim is thus to establish and reestablish step-by-step access and connectivity. Based on this considerations, we focussed our analysis on feedback in the HRI experiments.

5.3 Qualitative Evaluation

5.3.1 Ethnomethodological Conversation Analysis

In contrast to the experimental setting described in IV., qualitative analysis is based on fine grained observations on the behavior [3]. The methodology refers directly to the observed and video taped interaction. Ethnomethod-

ological conversation analysis is a technique acted out by the researcher him-/herself. Not the user's ratings about their own interactions were collected in a questionnaire, but their performances were reviewed and analyzed.

5.3.2 Hypotheses and Questions

Feedback serves as an interacting mechanism. There is a dynamic interplay between the user's feedback and BIRON's verbal behavior. By studying the experimental setting qualitatively, we analyzed how users react, if the assigned reaction is not the expected one, or if the robot shows no feedback at all.

5.4 Results of Qualitative Analysis of Feedback in HRI

Due to the qualitative researcher's intention in studying *concrete* interaction proceedings in daily life, we analyzed the users' communicative behavior in its situative context. We found a huge variability of human behavior in HRI with BIRON. Several fixed phenomena consist of verbal variations, mimic assertions or rhetorical comments. The human users are referring to both, sensorial perceived utterances and conceptual ones, like their own expectations.

Based on the user's reactions BIRON's feedback can be interpreted either as a matching or a mismatching reply. BIRON's denounced feedback did not fit in every situation with the user's expectance. But in each situation the user reacted *somehow*.

5.4.1 Matching and Mismatching Feedback

Users expect from an interaction situation an alternating interplay. This bilateral mechanism which includes the functionality of turn-taking ensures the flow of communication. So the distribution of any feedback from the robot mostly is registered positive and answered by the user. We differentiate the two cases of a "good" *matching* reply that is directing to the continuation of the interaction process, and a "bad" *mismatching* reply that leads to an interruption in the flow of communication.

In the following we define a matching reply as holding the line by answering somehow as expected, as the first example of an dialogue with BIRON underlines:

User 03: "Look here!"
BIRON: "Yes, I'm looking."
User 03: "This is a keyboard."
BIRON: "It's really interesting."

From the human user's point of view it is also expectable that BIRON would give some internal insights, so he reacts positive to a given problem report as shown in the following example:

User 13: "What can you do?"
BIRON: "I just understood you partially. What did you say?"
User 13: "What can you do?"
BIRON: "I can follow you to an other location and you can show me something."
User 13: "Ok."

Even if the robot does not give semantically correct feedback, users tend to reprise the dialogue if they get *any* feedback from the system as can be seen in the next example:

User 01: 'Hello!"'
BIRON: "Sorry, I can't understand you. Maybe you can use some other words?"
User 01: "What can you do?"

We define an episode in our experiment as mismatching if the user's expectation can not be linked to the robot's statement hence the user can not answer the robot.

The user's expectation is linked to an awaited reaction. A not expected comment from the robot can irritate the dialogue because the user does not wait for it and is not ready to interprete it. In mismatching cases in which BIRON did not produce feedback at all or was giving an unanticipated answer, the users showed individual reactions. They often expressed their open loop as verbal repeating of their previous assertions. This meant on the prosodic level, modifying their accentuation by emphasis of singular words or syllables, also slowed reechoing and even spelling have been shown. In some cases the user switched to a different vocal pitch. On the lexical level, some usage of synonyms or differing expressions took place. The users expressed on non-verbal level mimic variation, such as lifting the eyebrows

or augmenting the general frequency of gesture usage. There has also been discursive feedback just like encouraging the robot (e.g. User 03: “Oh, come on! Talk to me please!”).

Also, the users were shifting to thematic cues in form of naming and commenting the robot's mistakes as in the following dialogue example:

BIRON: ”I know it is sometimes difficult with me, but please don't feel discouraged!”
User 03: “What choice do I have?”

Some contributions are made (e.g. User 03: “Please don't tell me it's my fault.”) and even suppositions about the internal state of the robot are not rare (e.g. User 02: “I suppose that he wishes to end the conversation with me!”).

Interestingly users also tried out an other variance: they shifted to a meta reflexive level by addressing the experimentator. They interrupted the mismatching HRI and established an interaction with a human communication partner to whom they are familiar with and the flow of communication retained - in this case with a different partner.

5.4.2 Missing of Feedback

We can learn much more about the problem of communication by looking at the critical cases: As most critical moment within those interactions with a robot we found a given order by the user, not being reacted to at all. More specifically, if the robot does no show any reaction, there is no access for slightly and effortless continuing the interaction. After Garfinkel [11], those moments show fruitful efforts in applying repairing strategies. If a communicative lack occurs, the human will be trying to provoke any reset of the former dialogue to gain new access to the communication. In those situations the human users have to improve the interaction and they have manifold possibilities: they might be awaiting even longer for the robot to answer - and most of them in our study already did. Others tended to evoke a new and better accessible interactional element. This would be an assertion, provoking some feedback. Some non-verbal cues like snipping the fingers or waveing were acted out too. In each case, even the mismatching trials, the act of communicating *continues*, even if the interaction with the robot is cut off finally.

These general replying mechanisms are leading to typical behavior people acted out in the experiment setting: The users reactions tend to continue

the interaction and offer some renewal of accessibility. If some spoken instructions remain not-answered, the user is getting irritated. Irritation will be augmenting by its duration.

Feedback is a reciprocal mechanism of monitoring, interpreting and answering the interaction partners' verbal, mimic and embodied expressions as well as actions. The users tend to obtain and retain orientation towards the robotic system.

6 Conclusion

The quantitative results have shown that the likeability of the robot is significantly correlated to the robot's interaction behavior with the more extrovert system being preferred over the less initiative one. This result can be interpreted from a sociological point of view that by giving more feedback, the robot provides more access to the user to re-enter the communication after it has been interrupted by a system failure. Thus, by excusing for a fault, the robot gives the user an opportunity to make sense of the communication again and, thereby, to answer.

In contrast to this positive feedback, the robot's message "I've lost you" does not relate to the user's own experience and thus does not provide access for the user to re-enter the conversation since it does not make sense to her. This means that the understanding and correct interpretation of feedback is closely related to the context that the conversation is taking place in.

From these findings we can draw some conclusions about the design of feedback: A criterion for feedback that contributes to successful communication is that it needs to produce accessibility in order to motivate the user to continue the communication even when in trouble. In contrast, feedback that does not produce accessiblity will demotivate the user because it can not be related to the user's own world of experience and expectations in the concrete context.

Acknowledgment

This work is funded by the European European Commission Division FP6-IST Future and Emerging Technologies within the Integrated Project COGNIRON (The Cognitive Robot Companion) under Contract FP6-002020 and by a fellowship of the Sozialwerk Bielefelder Freimaurer e.V..

References

[1] K. Aoyama and H. Shimomura. Real world speech interaction with a humanoid robot on a layered robot behavior control architecture. In *Proc. Int. Conf. on Robotics and Automation*, 2005.

[2] P. L. Berger and T. Luckmann. *The Social Construction of Reality.* Doubleday, Inc., Garden City, New York, 1966.

[3] J. R. Bergmann. *Qualitative Sozialforschung. Ein Handbuch*, chapter Konversationsanalyse, pages 524–537. Reinbek: Rowohlt, 2000.

[4] R. Bischoff and V. Graefe. Dependable multimodal communication and interaction with robotic assistants. In *Proc. Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, 2002.

[5] J. E. Cahn and S. E. Brennan. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[6] H. H. Clark, editor. *Arenas of Language Use.* University of Chicago Press, 1992.

[7] K. Dautenhahn, B. Ogden, and T. Quick. From embodied to socially embedded agents - implications for interaction-aware robots. *Cognitive Systems Research*, 3(3):397–428, 2002.

[8] D. C. Dennett. *The intentional stance.* MIT Press, 1987.

[9] B. Edmonds and K. Dautenhahn. The contribution of society to the construction of individual intelligence. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors, *Proc. Workshop "Socially Situated Intelligence" at SAB98 conference, Zuerich, Technical Report of Centre for Policy Modelling*, pages CPM–98–42. Manchester Metropolitan University, 1998.

[10] A. T. et al. National character does not reflect mean personality trait levels in 49 cultures. *Science*, 310:96–99, 2005.

[11] H. Garfinkel. *Studies in Ethnomethodology.* Englewood Cliffs, N. J.: Prentice-Hall Inc., 4 edition, 1967.

[12] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. BIRON –

The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors, *Proc. Int. Workshop on Advances in Service Robotics*, pages 27–32, Stuttgart, Germany, May 2004. Fraunhofer IRB Verlag.

[13] M. Hackel, S. Schwope, J. Fritsch, B. Wrede, and G. Sagerer. A humanoid robot platform suitable for studying embodied interaction. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 56–61, Edmonton, Alberta, Canada, August 2005. IEEE.

[14] K. Kaye. *Studies in mother-infant interaction*, chapter Toward the origin of dialogue, pages 89–119. Academic Press, London, 1977.

[15] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*, 2003.

[16] S. Li, B. Wrede, and G. Sagerer. A computational model of multi-modal grounding. In *Proc. SIGdial workshop on discourse and dialog*. ACL, 2006.

[17] S. Li, B. Wrede, and G. Sagerer. A dialog system for comparative user studies on robot verbal behavior. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication*. IEEE, 2006.

[18] N. Luhmann. *Soziologische Aufklärung 2. Aufsätze zur Theorie der Gesellschaft*, chapter Interaktion, Organisation, Gesellschaft. Anwendungen der Systemtheorie, pages 9–24. VS Verlag für Sozialwissenschaften, 1975.

[19] N. Luhmann. *Soziologische Aufklärung 3. Soziales System, Gesellschaft, Organisation*, chapter Die Unwahrscheinlichkeit der Kommunikation, I. Allgemeine Theorie sozialer Systeme, pages 29–40. 4 edition, 1981.

[20] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services,. In *AAAI*, 1999.

[21] R. R. McCrae and O. John. Introduction to the five-factor model and its applications. *Journal of Personality*, 60:175–215, 1995.

[22] J. L. P. Persson and P. Lonnqvist. Anthropomorphism - a multi-layered phenomenon. In *Proc. Socially Intelligent Agents - the Human in the Loop, AAAI Fall Symposium, Technical Report FS-00-04*, pages 131–135. AAAI, August 2000.

[23] B. Rammsted and O. John. Measuring personality in one minute or less: A 10-items short version of the big five inventory in english and german. In *submitted*.

[24] D. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.

[25] A. Weiss, J. E. King, and L. Perkins. Personality and subjective well-being in orangutans. *Journal of Personality and Social Psychology*, 90(3):501–511, 2006.

[26] S. Woods, K. Dautenhahn, C. Kaouri, R. te Boekhorst, and K. L. Koay. Is this robot like me? Links between human and robot personality traits. In *AISB 2005*, 2005.

Teaching an autonomous wheelchair where things are

Thora Tenbrink
SFB/TR 8 Spatial Cognition, University of Bremen, Germany
tenbrink@informatik.uni-bremen.de

1 Introduction

How do users react when asked to inform an autonomous wheelchair about the locations of places, objects and about spatial relationships in an indoor scenario? This paper presents a qualitative analysis of speakers' spontaneous descriptions in such a task, analyzing how non-expert German and English users talk to a robot that is supposed to augment its internal map with the information the users provide. The analysis focuses on a range of aspects which reflect systematic features and variability in the linguistic descriptions: choice of strategy, granularity level, presupposition, underspecification, vagueness, and syntactic variations. A brief language comparison reveals systematic differences between German and English usage with respect to spatial descriptions. First (sketched) results of a follow-up study point to desirable effects of allowing the robot to react verbally to the users' input on speakers' spontaneous choices.

The results presented here are explorative and qualitative, reflecting work in progress within a larger research enterprise that comprises technological as well as linguistic endeavours. The linguistic work is part of project I1-[OntoSpace] of the DFG-funded major research program SFB/TR 8 Spatial Cognition situated in Bremen and Freiburg. Other projects within this program deal with implementations of the linguistic findings within a dialogue system (I3-[SharC]), and a broad range of robotics-related issues that concern the matching of perceptual and verbal input with the robot's prior spatial knowledge, for example, via computational models (e.g., R3-[Q-Shape], A2-[ThreeDSpace]).

Related work is carried out also in other projects dealing with human-robot interaction, for example, within the EU funded major project COSY, the recently completed SFB 360 in Bielefeld, and the SFB 378 in Saarbrücken. Also, relevant work on spatial language semantics and usage is carried out

at several places (e.g., the LIMSI group in Paris, and the research groups around K. Coventry and L. Carlson, among others), the results of which influence the interpretation and evaluation of our specific findings as detailed below. A thorough and systematic overview of relevant knowledge about spatial language is given in [14]. In this paper, I focus on the specific results of our empirical studies involving "Rolland".

2 Experimental Study I

2.1 Method

In our[1] scenario, the robot (the Bremen autonomous wheelchair "Rolland") [9], is situated inside a room that is equipped with a number of functionally interesting objects and furniture, intended to resemble a disabled person's flat. Our users (non-disabled university students) are seated in the wheelchair and given four tasks: first, they are asked to steer the wheelchair (whose automatic functions are not operating) around inside the room they are currently in, and teach it the positions of the 'most important' objects and places so that it can augment its internal map. Second, they are placed at one specific position inside the room and asked to describe the spatial relationships of the locations to each other from there. Third, they are asked to steer the wheelchair along the hallway and visit some predetermined places, explaining, again, the locations that they encounter along the way. Their final task then is to instruct the wheelchair to move autonomously to one of the places just encountered. In this baseline experiment, the wheelchair does not react in any way throughout the study. In a follow-up study described briefly below, the robot gives detailed verbal feedback; first results of this study complement the current analysis.

It is one of the most prominent aims in our project to identify speakers' spontaneous ideas on how to address robots in carefully controlled spatial tasks (cf. [2]). From a technological perspective, this approach enables the system designers to allow for the interpretation of an increasing range of utterances that are spontaneously produced in a given context, without having to provide the users with a predefined list of commands. A sophisticated dialogue system is currently under development (see e.g., [10, 15]); also, other modules of the robotic system are being developed toward increasing integration of perceptual and linguistic information (e.g., [6]).

[1] The study was carried out in cooperation with other researchers within the SFB/TR 8, most notably K. Fischer.

From a linguistic perspective, this approach allows for optimal flexibility in investigating generalizable features of human-robot interaction. The main idea here is to restrict the setting rather than the users' utterances. Given a clearly defined discourse context, the range of users' reactions remains within reasonable (and analyzable) limits, in spite of the fact that users are not trained or asked explicitly to restrict their utterances in any way. Thus, the contents of what users say are restricted by the setting and the discourse task, not by the prior expectations of an experimenter limiting the possible outcomes. In our scenarios, there is an emphasis on spatial language; therefore, the extralinguistic context is essential for speakers' choices and their interpretations.

2.2 Procedure

We collected utterances by 23 German and 7 English native speakers, which provides a useful basis for a qualitative language comparison. The approximate duration of the study was 30 minutes per participant.

The spoken language data were stored in video and audio files and subsequently transcribed into an xml format for annotation and analysis. About 12.600 German and 5.800 English words collected (2.100 & 800 speech units or 'utterances').

2.3 Results

Given the present scenario, it could be expected that users adhere to a number of principles that they consider adequate for an automatic system: they might be specifically precise and explicit, they might try to be especially consistent, they might try to identify those items that might be relevant for an autonomous wheelchair, and they might adopt a specifically formal or otherwise peculiar kind of language as "computer talk". In our data, it turns out that neither of these expectations is met in any consistent manner. Instead, we encounter a range of variability in the ways that speakers conceptualize their task, and therefore choose systematically different strategies to solve it. These are directly reflected in their linguistic choices. In the following, I present a qualitative linguistic analysis concerning the variability in the linguistic descriptions with respect to choice of strategy, granularity level, presuppositions, underspecification and vagueness, and spatial reference to locatum, relatum, and origin. Also, results of a qualitative language comparison are presented.

With respect to **strategy choice**, a basic distinction can be identified

that has already proved to be specifically stable across discourse tasks, also in previous work within our research group (e.g., [3, 7]): a part of the utterances describe entities (goals) and their positions, while others simply refer to the path or direction how to get there. In the first task of the present scenario, many speakers combine both strategies, as in:

(1) so I need to go right again (...) and I am going to go over to the computer

The fact that a substantial amount of utterances[2] is direction-based rather than goal-referring is astonishing, because our users were not asked to describe their own movements. It therefore reflects something else, for instance, a strategy towards achieving their aim: via knowledge about the spatial movements and directions, the robot is supposed to be able to infer the positions of objects and to establish spatial relationships. However, this is a particularly difficult task for a robot to achieve, since spatial directions are notoriously vague and involve a high number of complexities with respect to implementation. In the second task, users only employ the goal-based strategy, which is reasonable because they are not moving throughout this task. However, in the third task (describing the places in the hallway) users are even more inclined than before to simply describe their movements, such as:

(2) go slightly to my left, okay straight ahead, and then to my left

This may be due to the nature of this spatial task, which involves navigation within a hallway environment with a clear structure (unlike the previous tasks which took place inside a room without pre-defined paths). Also the route instructions in task 4 contain a high number of such "incremental" instructions. Interestingly, many of these utterances do not refer to entities at all, in spite of the fact that the general task scenario focused on information conveyance about the locations of entities. Clearly, the contents of the speakers' utterances depend very much on their conceptualizations of the task, especially with respect to what they think might be useful for the interaction partner (in this case, the robot). While this phenomenon might be specifically obvious where an unfamiliar interlocutor with unknown abilities is involved, we consider this result as reflecting a more general discourse

[2]Numbers or relative frequencies presuppose a valid general and objective measure against which a suitable comparison could be made, which is a non-trivial endeavour that we are currently pursuing. For the moment, the qualitative insight should suffice that speakers do use both strategies, and combined ones, rather frequently.

factor which certainly comes into play – albeit in much more subtle ways – in any kind of discourse context (e.g., [1]). In our framework, the main challenge in this respect is to (subtly and unobtrusively) influence users' conceptualizations in such a way as to trigger suitable linguistic representations (see below).

The second aspect of analysis concerns the levels of **granularity** reflected in the descriptions. Some users refer to objects by their basic level class name and leave it at that, such as:

(3) bin jetz' am Tisch, fahre jetz' zum Sessel (I'm at the table now, now driving to the armchair)

However, this coarse level of granularity was actually quite rare. Most speakers conveyed information on a much more specific level of detail. Here we can distinguish between two main foci: some users concentrate on perceptual or object-oriented, others on functional aspects. Functional utterances often contain information with respect to what to do with the objects:

(4) when you feel like taking a break, you can relax, and watch TV

(5) there's a plant on the table, and it's important not to forget to water it

(6) this is the dining table, this is also a very important part in a house because this is where people get together to eat

.

Perceptual (object-oriented) utterances, on the other hand, describe the objects in much detail. This might be the case either with respect to a single object, as in:

(7) auf dem Tisch liegt eine lila Tischdecke, kariert lila weiß mit Streifen, an den Seiten, mit Pflanzen blau und grün und pink
(on the table there's a purple tablecloth, checkered purple white with stripes at the sides, with plants blue and green and pink)

or with respect to descriptions of small items that happen to be present:

(8) there's a ruler here on the table which is to the right hand side, and there is a light and a staple, and some folders

Generally, and again surprisingly, functional utterances prevail, in spite of the fact that the wheelchair will probably not be able to utilize this kind of information. With regard to this level of analysis, our conclusion is that speakers attend to a high degree to the affordances and functions of the objects about which information is to be conveyed. These must therefore be taken into account in any model or system concerned with real world scenarios involving natural objects.

With respect to the analysis of **presuppositions**, one finding is fairly remarkable throughout the data. Speakers seldom introduce entities as new, independent of whether they have been encountered before. Instead, they switch freely between the true introduction of entities as in:

(9) there's an armchair in front of me

and referring to them by definite articles as though they were already known, as in:

(10) I want you to go to the remote control, which is lying on the table

which was uttered almost at the start of the study, without prior mention of a table or remote control. This may reflect a general speaker tendency to refer to present entities as known, or to make use of exophoric reference (presupposing the recognition of present entities) if possible. However, in our situation the actual task is to introduce a robot to the entities; even here, speakers do not consistently express this linguistically. This lack of linguistic signposting may be specifically problematic for a robot, as the robot's perceptual abilities and functionalities differ systematically from those of the human. Therefore, the identification of present objects cannot be presupposed in the same way as with humans.

Furthermore, utterances are both highly **underspecified** and **vague**. This concerns mainly the spatial descriptions involved, which were a major factor of the given discourse task. Many utterances do not contain a spatial term at all, but simply point to the existence of an object, as in:

(11) there's also a candle

The spatial terms that do occur are typically not precise, but simply give a vague spatial direction, as in:

(12) to my left here there's another computer

This finding is in accord with some of our own earlier results which point to the fact that speakers seldom modify spatial terms by precisifiers, as long as there are no competing objects nearby that would fit the same description [12]. However, since in the present task the users were expected to inform the wheelchair about spatial positions, it could have been expected that they provide more specific descriptions even in the absence of competing objects. This was overwhelmingly not the case, except for a number of utterances that contain metric information about estimated distances:

(13) separated by about twenty feet

and some attempts at providing more precise angles, for which there is of course no guarantee concerning correctness, yielding hesitations and self-corrections:

(14) sixty-five degrees is the coffee-table, no that's like more like eighty, eighty degrees seventy-five or eighty degrees to the right is the coffee-table

In addition to vagueness, there is underspecification: most spatial terms are relational and thus require a relatum, a different entity that serves as the basis for a spatial description (such as "my" in "to my left"). This relatum is often not provided on the linguistic surface, as in:

(15) the first computer on the left

Finally, we turn to the variability involved on the **language surface**, which is specifically important for linguistic text type analyses as well as the automatic processing of natural language utterances. Here, of course, variability is already predicted by the range of variation with respect to strategy and granularity levels as just described. In addition to that, even one single speaker may switch freely between illocutionary acts and syntactic constructions without any apparent reason, as exemplified in the following sequence:

(16) I want to go over to the sofa (...) so go right (. . .) I want you to go to the remote control (. . .) so I'm at the small table that's the coffee table (. . .) just to my left is the television (. . .) I need you to turn round (. . .) if I need to to sit in the sofa (. . .) and there's an armchair in front of me, and it is just to the right of the coffee table

Remarkably, this speaker switches between describing her own actions and desires, instructing the robot (which can't move autonomously), describing the scene, and describing hypothetical actions. All of these result in different surface forms. They may reflect the speakers' uncertainty as to how to address a robotic wheelchair, although the task itself seemed to be clear and was not often misunderstood by the participants.

A comparison between English and German language structures yields interesting results with respect to the occurrence and syntactic distribution of the three components of a projective spatial relationship; namely, locata, relata, and origins. A locatum is the object the location of which is being described, a relatum is another object in relation to which the locatum is described, and an origin is the point of view taken for the spatial description (see [14] for a systematic account). Although all projective terms (i.e., *left, right, front, back*, and so forth) presuppose the existence of these three elements, not all spatial descriptions contain all of them explicitly. Most often, the origin (or perspective) is omitted, as it is taken for granted by the speakers. This is not surprising in light of the fact that, in this scenario, there is essentially only one perspective available, as the speaker shares the view direction with the robot wheelchair they are sitting in. Altogether, the perspective is mentioned explicitly 17 times in the German data, but only twice throughout the English data. This result is consistent with results in other settings in which German speakers also tended to mention perspective more often than English speakers, and primarily so if there is a potential conflict [13].

Furthermore, there is a notable difference in information structure between the two languages whenever the relatum is mentioned (which is not always the case). In German, the relatum (which is assumed to be known in the context and now serves as a point of departure for the new object or location) is typically (in about 65% of cases) mentioned first, while the newly introduced object (the locatum) appears at the end of the utterance (as "news"). Examples for this structure are:

(17) sehe ich auf der linken Seite einen Kühlschrank, auf dem Kühlschrank liegt ein kleines Häkeldeckchen
(on the left side I see a fridge, on the fridge lies a small doily)

(18) daneben ist ein Kühlschrank, daneben ist ein Tisch, direkt daneben...
(beside it there's a fridge, beside it there's a table, directly beside it...)

This kind of structure has been suggested in the literature as a default

strategy for spatial descriptions [5, p119]. Some speakers also use themselves as locatum and the introduced objects as relatum:

(19) jetzt steh ich vor einem großen Tisch
(now I am in front of a big table)

In English, in contrast, the locatum is mentioned first in about 75% of cases. Thus, the focused (new) element comes first, followed by the description of its spatial relation – even if the relatum has just been mentioned. An example is:

(20) the plant is to the right of the cookies; the computer is to the right of the plant

From our data, we can therefore tentatively conclude that Herrmann & Grabowski's proposed default strategy may indeed be a prominent strategy for Germans (in scenarios like the present one in which the strategy is suitable), but not to the same degree for English speakers. It would be interesting to follow this hypothesis up with more controlled experimental studies or broader corpus investigations; to my knowledge, this has not been done.

In general, the results of our study point to a broad range of systematic variability in the language directed to a robot within the given scenario of a map augmentation task. In order to enable successful verbal human-robot interaction, the system needs to be designed to account for this variability. This is not in all cases easy to achieve, since speakers' utterances contain a high number of complexities and underspecifications that are difficult to handle for the system [11]. However, robotic output that is specifically tailored on the basis of these results may induce users to modify their linguistic choices in a way that could be better suited for their artificial interaction partner. The results of our recent follow-up study indicate how this may be achieved, which is briefly outlined next.

3 Experimental Study II

3.1 Method

The same four tasks as in Study I were carried out, this time in a "Wizard-of-Oz" scenario. In this by now well-established paradigm, a person hidden behind a screen triggers pre-recorded robot utterances suitable for the situation, while the experimental participants are induced to believe that the

robot responds autonomously. The idea behind this approach is that system requirements and planned functionalities can be tested even before the system is fully developed. Furthermore, speakers are influenced to a high degree by robotic output, and they can therefore be influenced towards using the kind of language that the robot will be able to understand. This process works, for example, on the basis of interactive alignment mechanisms as described by [8]. The specific aims of this follow-up study were therefore, on the one hand, to investigate in how far the features of speakers' spontaneous language productions are influenced by the robotic output, and on the other hand, to test the suitability of the Rolland's pre-determined utterances to influence the speakers' choices in a useful way. An important goal here is to reduce variability in speakers' utterances while still refraining from providing the user with a list of possible commands, and to induce them to use conceptual options that match the robotic system. Our earlier results already proved that speakers change their conceptualizations, and therefore their linguistic choices, based on the robot's utterances, to the degree that they can integrate this information suitably with their own conceptions [4]. Thus, there are limits to users' adaptations; robotic output must therefore be carefully controlled and tested.

3.2 Procedure

This time, the task was carried out within various conditions (mostly concerning the participants mother tongue in relation to the language used in the study) that will not be analysed in detail here. Participants were 17 German native speakers talking German, 11 English native speakers talking English, 14 German native speakers talking English, and 8 German native speakers talking English in a sub-condition in which the Rolland gave far less information concerning the desired strategy choice and level of granularity than in the other conditions. Here I consider only the native-speaker data (17 German and 11 English speakers).

3.3 First Results

Here is a typical example dialogue, showing the very beginning of a user's interaction with the robot:

(21) ROBOT: you can take us now to a place you want to name.
USER: to the sofa, turning right moving straight ahead
ROBOT: where are we going to?
USER: we're going to the sofa. We're passing the table on our right

Generally, as in this example, the goal-based strategy turned out to be much more dominant than in Study I. Speakers included information about movements, but only to supplement their goal-based utterances. Also, the level of granularity was much more suitable for the robot. Many details (functional or object-related) did not occur at all or were mentioned far less frequently.

As in Study I, speakers again frequently treated new entities linguistically as "Given". The robotic output was not specifically tailored to change this particular aspect of speakers' choices. Also, spatial descriptions are still vague and partly underspecified. In this respect, the robot's questions sometimes triggered spatial descriptions when none had been given earlier, or even a re-adjustment of the actual positioning with respect to the described object:

(22) ROBOT: am I in front of it?
USER: almost [4 seconds pause with movement] yes

In subsequent utterances, speakers in some cases re-used the kind of spatial description that had been asked for:

(23) ROBOT: am I in front of it?
USER: yes
ROBOT: oh okay, I see which object you mean.
USER: and now the couch, now I'm in front of it

In phase 4, the robot asked clarification questions if utterances were underspecified. As expected, this led to improved clarity, as in the following:

(24) USER: turn left, follow the hallway and then take another left
ROBOT: Is this the first possibility to turn?
USER: yes. Then continue left continue down that hallway
ROBOT: Up to which point do I go straight?
USER: you go straight until the last room on the left side

Especially in phase 1, the speech act variability is greatly reduced. This seems to indicate that speakers are no longer uncertain to the same degree as in Study I in addressing the robot, since they get feedback. This time, the syntax of users' utterances is often reduced to sparse constructions:

(25) ROBOT: where are we going to?
USER: table

Further analyses of this study, specifically with respect to the dialogue flow in relation to the robot utterances, are published in [15]. Also, this study is still under analysis with respect to a range of details that will be published elsewhere. However, these first results already indicate that the robot's utterances have a great impact on the users' linguistic choices, concerning the more limited range of variability as well as the decisively increased proportion of utterances that match suitably with the robot's assumed knowledge.

4 Conclusion

I have presented a qualitative linguistic analysis of one experimental study in monologic HRI together with first results of a follow-up study involving dialogue. Results show that there is a broad variability of possible choices and strategies available to speakers, which can be reduced decisively by suitable robotic output. Another interesting result is a systematic difference between the German and English data with respect to the information structure in spatial descriptions (in Study 1): German speakers tend to begin with known objects, while English speakers start with the newly introduced entity.

The development of a dialogue system that incorporates our results is underway [10]. Also within our project group, empirical HRI investigations with a real system rather than Wizard-of-Oz are carried out (e.g., [7]). These incorporate detailed knowledge about spatial language usage and resolution of underspecified spatial reference. Technologically, the crucial point is to enable the robot to map linguistic and perceptual information with its internal knowledge. The contribution of linguistic analysis to this endeavour is based on the fact that intelligent HRI dialogue can solve many upcoming problems. Suitable clarification questions and triggers of the desired kind of language systematically help to meet the robot's requirements, if sufficient knowledge about speakers' spontaneous choices and typical reactions can be built on.

References

[1] H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

[2] K. Fischer. Linguistic methods for investigating concepts in use. In T. Stolz and K. Kolbe, editors, *Methodologie in der Linguistik*, pages 39–62. Frankfurt a.M.: Lang, 2003.

[3] K. Fischer and R. Moratz. From communicative strategies to cognitive modelling. In *Workshop Epigenetic Robotics*, Lund, 2001.

[4] K. Fischer and R. Wilde. Methoden zur Analyse interaktiver Bedeutungskonstitution. In C. Solte-Gresser, K. Struwe, and N. Ueckmann, editors, *Forschungsmethoden und Empiriebegriffe in den neueren Philologien, Forum Literaturen Europas*. LIT-Verlag, Hamburg, 2005.

[5] T. Herrmann and J. Grabowski. *Sprechen: Psychologie der Sprachproduktion*. Spektrum Verlag, Heidelberg, 1994.

[6] C. Mandel, U. Frese, and T. Röfer. Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, 2006.

[7] R. Moratz and T. Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, 6(1):63–106, 2006.

[8] M. J. Pickering and S. Garrod. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27(2):169–190, 2004.

[9] T. Röfer and A. Lankenau. Architecture and Applications of the Bremen Autonomous Wheelchair. In P. P. Wang, editor, *Proc. of the 4th Joint Conference on Information Systems*, volume 1, pages 365–368, 1998.

[10] R. Ross, J. Bateman, and H. Shi. Using generalized dialogue models to constrain information state based dialogue systems. In *Proc. of the Symposium on Dialogue Modelling and Generation*, 2005.

[11] H. Shi and T. Tenbrink. Telling Rolland where to go: HRI dialogues on route navigation. In *Proc. WoSLaD Workshop on Spatial Language and Dialogue, October 23-25, 2005*, 2005.

[12] T. Tenbrink. Identifying objects on the basis of spatial contrast: an empirical study. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel,

and T. Barkowsky, editors, *Spatial Cognition IV: Reasoning, Action, Interaction. International Conference Spatial Cognition 2004, Proceedings*, pages 124–146, Berlin, Heidelberg, 2005. Springer.

[13] T. Tenbrink. *Localising objects and events: Discoursal applicability conditions for spatiotemporal expressions in English and German. Dissertation.* University of Bremen, FB10 Linguistics and Literature, Bremen, 2005.

[14] T. Tenbrink. Semantics and application of spatial dimensional terms in English and German. Technical Report, SFB/TR8 Spatial Cognition 004-03/2005, University of Bremen, 2005.

[15] T. Tenbrink, H. Shi, and K. Fischer. Route instruction dialogues with a robotic wheelchair. In *Proc. brandial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue. University of Potsdam, Germany; September 11th-13th 2006*, 2006.

How To Talk to Robots: Evidence from User Studies on Human-Robot Communication

Petra Gieselmann
Interactive Systems Lab
University of Karlsruhe
Germany
petra@ira.uka.de

Prisca Stenneken
Physiological and Clinical Psycholgoy
Cath. University Eichstätt-Ingolstadt
Germany
stenneken.ku-eichstaett@gmx.de

Abstract

Talking to robots is an upcoming research field where one of the biggest challenges are misunderstandings and problematic situations: Dialogues are error-prone and errors and misunderstandings often result in error spirals from which the user can hardly escape. Therefore, mechanisms for error avoidance and error recovery are essential. By means of a data-driven analysis, we evaluated the reasons for errors within different testing conditions in human-robot communication and classified all the errors according to their causes. For the main types of errors, we implemented mechanisms to avoid them. In addition, we developed an error correction detection module which helps the user to correct problems. Therefore, we are developing a new generation strategy which includes detecting problematic situations, helping the user and avoiding giving the same information to the user several times. Furthermore, we evaluate the influence of the user strategy on the communicative success and on the occurrence of errors within human-robot communication. In this way, we can increase user satisfaction and have more successful dialogues within human-robot communication.

1 Introduction

We developed a household robot which helps users in the kitchen [9]. It can get something from somewhere, set the table, switch on or off lamps or air conditioners, put something somewhere, tell the user what is in the fridge, tell some recipes, etc. The user can interact with the robot in natural language and tell it what to do. A first semantico-syntactic grammar has

been developed and we now enhance this dialogue grammar by means of user tests and data collections.

Since the real robot consists of many different components, such as the speech recognizer, the gesture recognizer, the dialogue manager, the motion component, etc., we decided to restrict the user tests for the beginning to the dialogue management component. This means that we do not use a real robot to accomplish the tasks, but only a text-based interface where the dialogue manager informs the user what the robot is doing. In this way, we can skip problems resulting from other components and can focus on understanding and dialogue problems. We are aware of the fact that the findings cannot be directly applied to spoken communication with the real robot. However, this text-based paradigm was used for a first systematic investigation and is transferred to spoken robot communication in future studies.

In this paper, we discuss two methods how to improve human-robot communication: By analysing human-robot dialogues and avoiding the most important problems and on the other hand by changing the communicative strategy of the user. The second section deals with related work. Section three explains our household robot, the dialogue system and its particular characteristics. The fourth section is about user tests within different testing conditions which results in an error classification. Section five addresses the question whether communicative strategies affect the human-robot communication both in the subjective evaluation by the users and in the objectively measurable task success. Section six gives a conclusion and an outlook on future work.

2 Related Work

2.1 Errors in Man-Machine Dialogues

Most of the research about errors within man-machine dialogues deal with speech recognition errors: Some researchers evaluate methods for dialogue state adaptation to the language model to improve speech recognition [21, 11]. Work on hyperarticulation concludes that speakers change the way they are speaking when facing errors in principle so that the language model has to be adapted [19, 12]. Also Choularton et al. and also Stifelman are looking for general strategies on error recognition and repair to prepare the speech recognizer for the special needs of error communication [4, 19].

Furthermore, Schegloff et al. came up with a model which describes the mechanisms the dialogue partners use to handle errors in human-human

dialogue [17]. Also, within conversation analysis dialogues are evaluated concerning the rules and procedures how an interaction takes place [16]. These insights from human-human communication are essential for a natural human-robot communication.

However, the present study concentrates on semantic errors and classify them according to their reasons. For every error class, we develop methods to avoid it. Furthermore, we examine repair dialogues and their similarity to human-human repair dialogues in order to be able to perform efficient error handling strategies so that it will be easier for the user to correct errors which could not be avoided.

2.2 Effects of the User Strategy on Dialogue Success

In the field of humanoid robots and human-robot interaction the researchers concentrate on questions such as how to design the robot as similar as possible to a human regarding its outer appearance as well as its communicative behaviour [2, 1, 5]. In contrast, the present study concentrates on the human user and his communication strategies. This in turn would shape the expectations on how the dialogue should work and how errors could be avoided by another user strategy.

Furthermore, different evaluation methodologies of dialogue systems exist, starting from methodologies using the notion of a reference answer [13] to the most prominent approach for dialogue system evaluation which is Paradise [20] which uses a general performance function covering different measures such as user performance, number of turns, task success, repair ratio, etc. In the present study, objective measures were calculated from the participants' responses and success measures were assessed after each block in form of a questionnaire in order to get a deeper insight in the relationship between subjective and objective measures of success.

3 Our Household Robot

3.1 The Dialogue Manager

For dialogue management we use the TAPAS dialogue tools collection [14] which is based on the approaches of the language and domain independent dialogue manager ARIADNE [6]. This dialogue manager is specifically tailored for rapid prototyping. Possibilities to evaluate the dialogue state and general input and output mechanisms are already implemented which are applied in our application. We developed the domain and language depen-

Figure 1: Our Household Robot

dent components, such as an ontology, a specification of the dialogue goals, a data base, a context-free grammar and generation templates.

The dialogue manager uses typed feature structures [3] to represent semantic input and discourse information. At first, 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 our scenario, this ontology consists of all the objects available in the kitchen and their properties and all the actions the robot can do. The parse tree is then converted into a semantic representation and added to the current discourse. If all the necessary information to accomplish a goal is available in discourse, the dialogue system calls the corresponding service. But if some information is still missing, the dialogue manager generates clarification questions to the user. This is realized by means of generation templates which are responsible for generating spoken output.

3.2 Rapid prototype

We developed a rapid prototype system. This system includes about 32 tasks the robot can accomplish and more than 100 ontology concepts. Ontology concepts can be objects, actions or properties of these objects or actions. By means of this prototype we started user tests and continue to develop new versions of the grammar and domain model. The rapid prototype of our dialogue component is integrated in the robot (cf. figure 1) and also accessible via the internet for the web-based tests (cf. figure 2).

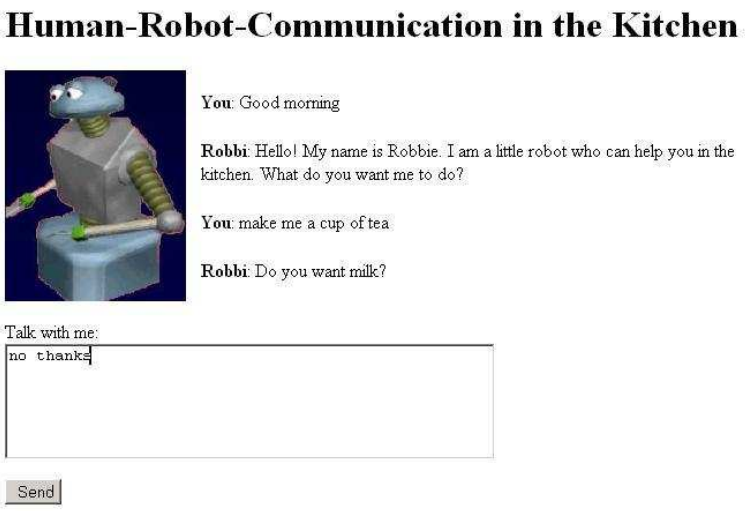

Figure 2: The web-based Interface of our Humanoid Robot

4 Analysis of Human-Robot Dialogues

4.1 Different Testing Conditions

As mentioned by Dybkjaer and Bernsen [7], predefined tasks covered in a user test will not necessarily be representative of the tasks real users would expect a system to cover. In addition, scenarios in user tests should not prime users on how to interact with the system which can only be avoided in a user test without predefined tasks or in a general user questionnaire. On the other hand, such a free exploration is much more complicated for the user and can be very frustrating, if the system does not understand the user intention. Therefore, we rely on two different testing conditions:

User tests with predefined tasks: Every user got five predefined tasks to accomplish by means of the robot. Since the tasks are given, it is easier for the user, but we do not get any information on the tasks a user really needs a robot for.

User tests without predefined tasks: The users were just told that they bought a new household robot which can support them in the household. They can freely explore and interact with the robot. This situation is much more realistic, but at the some time much harder for the user because he does not know what the robot can do in detail.

	Robot	Web-based
With Tasks	22.57%	49.94%
Without Tasks	57.03%	50.93%

Table 1: Turn Error Rates Within Different Testing Conditions.

In addition, we had two different testing conditions: Web-based user tests (see Figure 2) which have the advantage that lots of users all over the world can participate whenever they like to [18, 15] and also multimodal user tests with the robot (see Figure 1) to see how the user can get along with the real robot. The tests with the web-interface are of course different from the ones with the real robot, but within the web tests we can also use more dialogue capabilities concerning tasks the robot cannot accomplish until now.

4.2 Experimental Details and Results

We defined all the user turns which could not be transformed to the correct semantics by the dialogue system as *errors* so that the turn error rate gives the rate of error turns on the whole number of user turns. As expected, the turn error rate for tests with tasks is lower than without tasks (cf. Table 1) given the fact that the user has less clues what to say. Especially the tests with predefined tasks with the robot results in much less errors which might be due to the fact that these tasks were easier than in the web-based test and that the users could watch the robot interacting.

Nevertheless, within all the testing conditions, we can find the same error classes according to the following reasons for failure:

- **New Syntactic and Semantic Concepts**: New Formulations, New Objects, New Goals, Metacommunication
- **Ellipsis & Anaphora**: Elliptical Utterances, Anaphora, Missing Context
- **Concatenated Utterances**
- **Input Problems**: Punctuation & Digits, Background Noise, Grammatically Wrong Utterances

In addition, the rates for the error classes are very similar so that most of the errors can be found in the area of new syntactic and semantic concepts,

secondmost errors are input errors, thirdmost ellipsis and the fewest errors belong to the class of concatenated utterances.

Since the manual integration of new concepts is very time and cost-intensive, we developed a mechanism for dynamic vocabulary extension with data from the internet [10]. In addition, we implemented mechanisms to deal with ellipsis and anaphora [8] and handle complex user utterances. To resolve metacommunication, we grouped all the user utterances dealing with metacommunication according to the user intention:

- **Clarification Questions** from the user: The user wants to know, whether the robot understood him, what the robot is doing, etc.
- **Repair** of a user utterance: The user corrects the preceding utterance of the robot explicitly or implicitly.
- **Test** of the Robot: The user tests the abilities of the robot by giving instructions for tasks the robot can probably not accomplish; also insults are in this category.

Clarification questions from the user and tests of the robot indicate that the user does not know what the robot can do, has no idea on how to go on and what to say. Therefore, we implemented communication strategies so that the robot explains its capabilities to the users and help them in the case of problems. Different factors can indicate communication problems, such as that the user utterance is inconsistent with the current discourse, it cannot be completely parsed, it does not meet the system expectations, the user says the same utterance several times. These factors leads to an increase in error correction necessity and let the robot finally initiate a clarification dialog to help the user.

5 Influence of the User Strategy on the Communicative Success

5.1 Experimental Details

To evaluate the influence of the user strategy on the communicative success and the occurrence of errors, we conducted a web-based experiment with two different instructions for each participant:

- "Child instruction": The users were asked to talk to the robot in the same way as they would do to a little child.

- "Non-child instruction": The users got no detailed instruction on how to talk to the robot.

Each participant got predefined tasks. During the user interaction with the system, we measured the objective success per user by means of the turn error rate, the number of successfully accomplished tasks and the number of user turns necessary to accomplish resp. abort a task. After the participants had finished the task set under each instruction, they filled in a short user questionnaire about their general impression of the system and their experience during the experiment.

5.2 Results and Discussion

The effects of the instruction child vs. non-child are reflected in both qualitative and quantitative measures. Within quantitative measures, the instruction affected above all the mean utterance length, ie. number of words per user utterance. Participants had a numerically lower mean utterance length with instruction child (mean = 5.02) as compared to the non-child instruction (mean = 5.64). Interestingly, the effect of smaller mean utterance lengths in the child instruction occurs predominantly when the child instruction is given in the second block (the modulatory effect of the order of the instruction was marginally significant, $p = .053$). This might be due to the fact that participants who got the child instruction in the first block continued with this strategy also in the second block, irrespective of the instruction. This fact is also reported by some participants in the post-test questionnaires. Also within qualitative measures, about half of the participants reported to use short, simple sentences within the child instruction.

Pairwise comparisons were performed for possible effects of the instruction on subjective or objective measures of communicative success. For all variables, the effects of the instruction were non-significant, although we found a tendency towards more user satisfaction in the child instruction. This might be due to the fact that the present instructions were given rather implicitly and left some space for individual interpretations.

As expected, when comparing subjective and objective measures, a significant correlation was observed for the subjective measure "willingness to use the system again" and the objective measure "overall number of accomplished tasks" (p-value smaller than .05). Even though all other correlations did not reach significance, the numerical tendencies imply that the more tasks are accomplished, the higher the ratings are for subjective variables.

Findings from analyses of the user answers in free text also suggest that we have a rather strong influence of the participants' general attitude to-

wards robots which has a more dominant effect on the task success than the instruction. Since the conversation style of the user seems to be affected to a larger extent by the general attitude, future studies might address the question, how a dialogue system has to be designed to find out different user attitudes, support them and their different characteristics to improve the communication and avoid errors.

6 Conclusion and Outlook

We used a date-driven method to evaluate the reasons for errors in human-robot communication and implemented the following strategies to avoid resp. deal with them:

- dynamic extension of linguistic resources
- anaphora resolution
- handling complex as well as elliptical utterances
- meta communication

We evaluated the influence of the user strategy on the communicative success and found out that even though the user strategy had qualitative and quantitative effects on the communicative behavior, it was not systematically related to the communicative success in objective and subjective measures. However, the general attitude of the user towards robots has a more dominant effect on the task success than the instructed user strategy.

Future studies could further address the question, whether these findings are also true for extended grammars and tests with the real robot instead of the web interface.

References

[1] A. Billard and M. J. Mataric. A biologically inspired robotic model for learning by imitation. *Proceedings of the 4th conference on Autonomous Agents*, 2000.

[2] C. Breazeal. Robot in society: Friend or appliance? *Proceedings of the Agents99 workshop on emotion-based agent architectures*, 1999.

[3] B. Carpenter. *The Logic of Typed Feature Structures.* Cambridge University Press, 1992.

[4] S. Choularton and R. Dale. User responses to speech recognition errors: Consistency of behaviour across domains. *Proceedings of the Tenth Australian International Conference on Speech Science and Technology*, 2004.

[5] K. Dautenhahn and A. Billard. Bringing up robots or the psychology of socially intelligent robots: from theory to implementation. *Proceedings of the 3rd conference on Autonomous Agents*, 1999.

[6] M. Denecke. Rapid prototyping for spoken dialogue systems. *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.

[7] L. Dybkjr and N. Bernsen. Usability issues in spoken language dialogue systems. *Kuppevelt, J. v., Heid, U. and Kamp, H. (Eds.): Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, Natural Language Engineering*, 6:243–272, 2000.

[8] P. Gieselmann. Reference resolution mechanisms in dialogue management. *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue (CATALOG)*, 2005.

[9] P. Gieselmann, C. Fügen, H. Holzapfel, T. Schaaf, and A. Waibel. Towards multimodal communication with a household robot. *Proceedings of the Third IEEE International Conference on Humanoid Robots (Humanoids)*, 2003.

[10] P. Gieselmann and A. Waibel. Dynamic extension of a grammar-based dialogue system: Constructing an all-recipes knowing robot. *To Appear in: Proceedings of the International Conference on Spoken Language Processing (ICSLP 06)*, 2006.

[11] G. Gorrell. Recognition error handling in spoken dialogue systems. *Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia*, 2003.

[12] J. Hirschberg, D. Litman, and M. Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43, 2004.

[13] L. Hirschmann, D. A.Dahl, D. P. McKay, L. M. Norton, and M. C. Linebarger. Beyond class a: A proposal for automatic evaluation of discourse. *Proceedings of the Speech and Natural Language Workshop*, pages 109–113, 1990.

[14] H. Holzapfel. Towards development of multilingual spoken dialogue systems. *Proceedings of the 2nd Language and Technology Conference*, 2005.

[15] U.-D. Reips. Standards for internet-based experimenting. *Experimental Psychology*, 49(4), 2002.

[16] H. Sacks, E. Schegloff, and G. Jefferson. A simple system for the organization of turn-taking im conversation. *Language*, 50(4):696–735, 1974.

[17] E. Schegloff, G. Jefferson, and H. Sacks. The preference for self-correction in the organization of repair in conversation. *Language 53*, 1977.

[18] W. C. Schmidt. World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments & Computers*, 29(2), 1997.

[19] L. J. Stifelman. User repairs of speech recognition errors: An intonational analysis. *Technical Report, Speech Research Group, MIT Media Lab*, 1993.

[20] M. A. Walker, D. Litman, C. A. Kamm, and A. Abella. Paradise: A framework for evaluating spoken dialogue agents. *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, 1997.

[21] W. Xu and A. Rudnicky. Language modeling for dialog system. *Proceedings of the International Conference of Speech and Signal Processing (ICSLP'00)*, 2000.

To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk.

Anton Batliner, Christian Hacker, and Elmar Nöth
Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg, Germany
batliner,hacker,noeth@informatik.uni-erlangen.de

Abstract

If no specific precautions are taken, people talking to a computer can – the same way as while talking to another human – speak aside, either to themselves or to another person. On the one hand, the computer should notice and process such utterances in a special way; on the other hand, such utterances provide us with unique data to contrast these two registers: talking vs. **not** talking to a computer. By that, we can get more insight into the register 'Computer-Talk'. In this paper, we present two different databases, SmartKom and SmartWeb, and classify and analyse On-Talk (addressing the computer) vs. Off-Talk (addressing someone else) found in these two databases.

Enter Guildenstern and Rosencrantz. [...]
Guildenstern My honoured lord!
Rosencrantz My most dear lord! *[...]*
Hamlet [...] You were sent for *[...]*
Rosencrantz To what end, my lord?
Hamlet That you must teach me *[...]*
Rosencrantz [Aside to Guildenstern] What say you?
Hamlet [Aside] Nay then, I have an eye of you! *[Aloud.]* If you love me, hold not off.
Guildenstern My lord, we were sent for.

1 Introduction

As often, Shakespeare provides good examples to quote: in the passage from *Hamlet* above, we find two **'Asides'**, one for speaking aside to a third person and by that, not addressing the dialogue partners; the other one for speaking

to oneself. Implicitly we learn that such asides are produced with a lower voice because when Hamlet addresses Guildenstern and Rosencrantz again, the stage direction reads *Aloud*.

Nowadays, the dialogue partner does not need to be a human being but can be an automatic dialogue system as well. The more elaborate such a system is, the less restricted is the behaviour of the users. In the early days, the users were confined to a very restricted vocabulary (prompted numbers etc.). In conversations with more elaborated automatic dialogue systems, users behave more natural; thus, phenomena such as speaking aside can be observed and have to be coped with that could not be observed in communications with very simple dialogue systems. In most cases, the system should not react to these utterances, or it should process them in a special way, for instance, on a meta level, as remarks about the (mal–) functioning of the system, and not on an object level, as communication with the system.

In this paper, we deal with this phenomenon **Speaking Aside** which we want to call **'Off-Talk'** following [15]. There Off-Talk is defined as comprising 'every utterance that is not directed to the system as a question, a feedback utterance or as an instruction'. This comprises reading aloud from the display, speaking to oneself ('thinking aloud'), speaking aside to other people which are present, etc.; another term used in the literature is 'Private Speech' [14]. The default register for interaction with computers is, in analogy, called **'On-Talk'**. On-Talk is practically the same as Computer Talk [9]. However, whereas in the case of other (speech) registers such as 'baby-talk' the focus of interest is on the way **how** it is produced, i.e. its phonetics, in the case of Computer Talk, the focus of interest so far has rather been on **what** has been produced, i.e. its linguistics (syntax, semantics, pragmatics).

Off-Talk as a special dialogue act has not yet been the object of much investigation [1, 8] most likely because it could not be observed in human–human communication. (In a normal human–human dialogue setting, Off-Talk might really be rather self–contradictory, because of the 'Impossibility of Not Communicating' [21]. We can, however, easily imagine the use of Off-Talk if someone is speaking in a low voice not *to* but *about* a third person present who is very hard of hearing.)

For automatic dialogue systems, a good classification performance is most important; the way how to achieve this could be treated as a black-box. In the present paper, however, we report classification results as well but want to focus on the prosody of On- vs. Off-Talk. To learn more about the phonetics of Computer-Talk, On-Talks vs. Off-Talk is a unique constellation

because all other things are kept equal: the scenario, the speaker, the system, the microphone, etc. Thus we can be sure that any difference we find can be traced back to this very difference in speech registers – to talk or not to talk with a computer – and not to some other intervening factor.

In section 2 we present the two systems SmartKom and SmartWeb and the resp. databases where Off-Talk could be observed and/or has been provoked. Section 3 describes the prosodic and part-of-speech features that we extracted and used for classification and interpretation. In section 4, classification results and an interpretation of a principal component analysis are presented, followed by section 5 which discusses classification results, and by section 6 which discusses impact of single features for all databases.

2 Systems

2.1 The SmartKom System

SmartKom is a multi–modal dialogue system which combines speech with gesture and facial expression. The speech data investigated in this paper are obtained in large–scaled Wizard-of-Oz-experiments [10] within the SmartKom 'public' scenario: in a multi–modal communication telephone booth, the users can get information on specific points of interest, as, e.g., hotels, restaurants, cinemas. The user delegates a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. This agent is called 'Smartakus' or 'Aladdin'. The user gets the necessary information via synthesized speech produced by the agent, and on the graphical display, via presentations of lists of hotels, restaurants, cinemas, etc., and maps of the inner city, etc. The dialogue between the system and the user is recorded with several microphones and digital cameras. Subsequently, annotations are carried out. The recorded speech represents thus a special variety of non–prompted, spontaneous speech typical for human–machine–communication in general and for such a multi–modal setting in particular. More details on the system can be found in [20], more details on the recordings and annotations in [15, 17].

In the annotation, two different types of Off-Talk are labelled: *read Off-Talk* **ROT** and *other Off-Talk* **OOT**; every other word is via default annotated with the label **NOT** as *no Off-Talk*, i.e. On-Talk. If the user reads aloud words presented on the display, this is labelled as ROT; it was decided not to tell apart all other types of Off-Talk, e.g., speaking aside to another person or speaking to oneself, because these decisions are often not easy to make. Moreover, such a detailed labelling system would end up with

a small number of tokens for each sub-type which anyhow had to be mapped onto higher classes in order to process them automatically.

In the following example, a user wants to reserve two tickets for the cinema; first, the German original is presented, then an English translation. In the German example, <P> denotes pauses, +/ and /+, respectively, beginning and end of a repair, and <hm> a filled pause. Off-Talk is labelled adjacent to the pertinent word, without blanks. Most of the annotated phenomena are, however, filtered out of the German example; the English translation contains only the wording; here, ROT is given in capitals, and OOT boldface and recte.

system: *in diesem Kino können Sie nur telefonisch reservieren*
user: *kannst du mich dahin verbinden <P> Aladdin*
system: *diese Funktion ist momentan leider nicht verfügbar <P> kann ich Ihnen auch anders helfen*
user: *<hm> <P> das<OOT> is'<OOT> doof<OOT> <hm> <P> <A> geh doch noch mal zurück <P> Aladdin <P> zurückgehen*
system: *einen Augenblick bitte*
user: *ja <P> +/wie sieht d=/+ wie sieht das aus mit dem Kino Schloss <P> dreiundzwanzig fünfundvierzig <P> zwei Karten Aladdin <P> hoffentlich<OOT> nich' <OOT> wieder<OOT> telefonisch<OOT> doch<OOT>*
system: *in diesem Kino können Sie nur telefonisch reservieren*
user: *<hm> das<OOT> is'<OOT> natürlich<OOT> auch<OOT> doof<OOT> <P> gibt es noch eine andere Möglichkeit <P> sicher<OOT> nicht<OOT> <P> zurück*
system: *Reservierungen können nur im Kino StudioEuropa durchgeführt werden*
user: *aha na dann <P> gib mir doch mal das <P> Kino Europa <P> Kino Europa ah das sieht doch schon gut aus gut dann nehmen wir mal <P> <hm> <P> AmericanHistoryX zwoundzwanzig fünfundvierzig <P> aha<OOT> Reservierung<ROT> bestätigen <ROT> da<OOT> kommen<OOT> wir<OOT> doch<OOT>*
schon<OOT> der<OOT> Sache<OOT> näher<OOT>

system: *For this cinema, only reservation by phone is possible.*
user: *Can you connect me, Aladdin?*
system: *I'm afraid, at the moment, this is not possible.*
user: **That's bad** – *well – please go back, Aladdin – go back!*
system: *One moment please!*
user: *Well, what about, what about the Castle Cinema, 10.45 pm, Armageddon – two tickets, Aladdin –* **I hope it's not again reservation by phone.**
system: *For this cinema, only reservation by phone is available.*
user: **Again, that's bad.** *Is there another possibility?* **I guess not!** *Go back!*

Table 1: *Cross-tabulation of On-/Off-Talk vs. On-/Off-View*

	On-View	Off-View
NOT (On-Talk)	On-Focus, Interaction with the system	*(unusual)*
ROT	Reading from the display	—
POT	*(unusual)*	Reporting results from SmartWeb
SOT	Responding to an interruption	Responding to an interruption

system: *Reservations are only possible for the Studio Europe.*
user: *Well, okay, Studio Europe, Studio Europe, that's fine, well, then let's take – uh – AmericanHistory, 10.45 pm,* **okay, CONFIRM RESERVATION, now we are coming to the point.**

At least in this specific scenario, ROT is fairly easy to annotate: the labeller knows what is given on the display, and knows the dialogue history. OOT, however, as a sort of wast-paper-basket category for all other types of Off-Talk, is more problematic; for a discussion we want to refer to [17]. Note, however, that the labellers listened to the dialogues while annotating; thus, they could use acoustic information, e.g., whether some words are spoken in a very low voice or not. This is of course not possible if only the transliteration is available.

2.2 The SmartWeb System

In the SmartWeb-Project [19] – the follow-on project of SmartKom – a mobile and multimodal user interface to the Semantic Web is being developed. The user can ask open-domain questions to the system, no matter where he is: carrying a smartphone, he addresses the system via UMTS or WLAN using speech [16]. The idea is, as in the case of SmartKom, to classify automatically whether speech is addressed to the system or e.g. to a human dialogue partner or to the user himself. Thus, the system can do without any push-to-talk button and, nevertheless, the dialogue manager will not get confused. To classify the user's focus of attention, we take advantage of two modalities: speech-input from a close-talk microphone and the

video stream from the front camera of the mobile phone are analyzed on the server. In the video stream we classify **On-View** when the user looks into the camera. This is reasonable, since the user will look onto the display of the smartphone while interacting with the system, because he receives visual feedback, like the n-best results, maps and pictures, or even web-cam streams showing the object of interest. **Off-View** means, that the user does not look at the display at all[1]. In this paper, we concentrate on On-Talk vs. Off-Talk; preliminary results for On-View vs. Off-View can be found in [11].

For the SmartWeb-Project two databases containing questions in the context of a visit to a Football World Cup stadium in 2006 have been recorded. Different categories of Off-Talk were evoked (in the SW_{spont} database[2]) or acted (in our SW_{acted} recordings[3]). Besides *Read Off-Talk* (**ROT**), where the subjects read some system response from the display, the following categories of Off-Talk are discriminated: *Paraphrasing Off-Talk* ((**POT**) means, that the subjects report to someone else what they have found out from their request to the system, and *Spontaneous Off-Talk* ((**SOT**) can occur, when they are interrupted by someone else. We expect ROT to occur simultaneously with On-View and POT with Off-View. Table 1 displays a cross-tabulation of possible combinations of On-/Off-Talk with On-/Off-View.

In the following example, only the user turns are given. The user first asks for the next play of the Argentinian team; then she paraphrases the wrong answer to her partner (POT) and tells him that this is not her fault (SOT). The next system answer is correct and she reads it aloud from the screen (ROT). In the German example, Off-Talk is again labelled adjacent to the pertinent word, without blanks. The English translation contains only the wording; here, POT is given boldface and in italic, ROT in capitals, and SOT boldface and recte.

user: *wann ist das nächste Spiel der argentinischen Mannschaft*
user: *nein <ähm> die<POT> haben<POT> mich<POT> jetzt<POT> nur<POT>*

[1] In [12] On-Talk and On-View are analyzed for a Human-Human-Robot scenario. Here, face detection is based on the analysis of the skin-color; to classify the speech signal, different linguistic features are investigated. The assumption is that commands directed to a robot are shorter, contain more often imperatives or the word "robot", have a lower perplexity and are easy to parse with a simple grammar. However, the discrimination of On-/Off-Talk becomes more difficult in an automatic dialogue system, since speech recognition is not solely based on commands.

[2] designed and recorded at the Institute of Phonetics and Speech Communication, Ludwig-Maximilians-University, Munich

[3] designed and recorded at our Institute

Table 2: *Three databases, words per category in %: On-Talk (NOT), read (ROT), paraphrasing (POT), spontaneous (SOT) and other Off-Talk (OOT)*

	# Speakers	NOT	ROT	POT	SOT	OOT	[%]
SW_{spont}	28	48.8	13.1	21.0	17.1	-	
SW_{acted}	17	33.3	23.7	-	-	43.0	
SK_{spont}	92	93.9	1.8	-	-	4.3	

darüber<POT> informiert<POT> wo<POT> der<POT> nächste<POT> Taxistand<POT> ist<POT> und<OOT> nicht<POT> ja<SOT> ja<SOT> ich<SOT> kann<SOT> auch<SOT> nichts<SOT> dafür<SOT>
user: *bis wann fahren denn nachts die öffentlichen Verkehrsmittel*
user: *die<ROT> regulären<ROT> Linien<ROT> fahren<ROT> bis<ROT> zwei<ROT> und<ROT> danach<ROT> verkehren<ROT> Nachtlinien<ROT>*

user: *When is the next play of the Argentinian team?*
user: *no uhm* ***they only told me where the next taxi stand is*** **and not – well ok – it's not my fault**
user: *Until which time is the public transport running?*
user: THE REGULAR LINES ARE RUNNING UNTIL 2 AM AND THEN, NIGHT LINES ARE RUNNING.

2.3 Databases

All SmartWeb data has been recorded with a close-talk microphone and 8 kHz sampling rate. Recordings of the $\mathbf{SW_{spont}}$ data took place in situations that were as realistic as possible. No instruction regarding Off-Talk were given. The user was carrying a mobile phone and was interrupted by a second person. This way, a large amount of Off-Talk could be evoked. Simultaneously, video has been recorded with the front camera of the mobile phone. Up to now, data of 28 from 100 speakers (0.8 hrs. of speech) has been annotated with NOT (default), ROT, POT, SOT and OOT. OOT has been mapped onto SOT later on. This data consists of 2541 words; the distribution of On-/Off-Talk is given in Table 2. The vocabulary of this part of the database contains 750 different words.

We additionally recorded acted data ($\mathbf{SW_{acted}}$, 1.7 hrs.) to investigate which classification rates can be achieved and to show the differences to realistic data. Here, the classes POT and SOT are not discriminated and combined in *Other Off-Talk* (OOT, cf. SK_{spont}). First, we investigated the

SmartKom data, that have been recorded with a directional microphone: Off-Talk was uttered with lower voice and durations were longer for read speech. We further expect that in SmartWeb nobody using a head-set to address the automatic dialogue would intentionally confuse the system with loud Off-Talk. These considerations result in the following setup: The 17 speakers sat in front of a computer. All Off-Talk had to be articulated with lower voice and, additionally, ROT had to be read more slowly. Furthermore, each sentence could be read in advance so that some kind of "spontaneous" articulation was possible, whereas the ROT sentences were indeed read utterances. The vocabulary contains 361 different types. 2321 words are On-Talk, 1651 ROT, 2994 OOT (Table 2).

In the SmartKom ($\mathbf{SK_{spont}}$) database[4], 4 hrs. of speech (19416 words) have been collected from 92 speakers. Since the subjects were alone, no POT occurred: OOT is basically "talking to oneself" [7]. The proportion of Off-Talk is small (Table 2). The 16kHz data from a directional microphone was downsampled to 8kHz for the experiments in section 5.

3 Features used

The most plausible domain for **On-Talk** vs. **Off-Talk** is a unit between the word and the utterance level, such as clauses or phrases. In the present paper, we confine our analysis to the word level to be able to map words onto the most appropriate semantic units later on. However, we do not use any deep syntactic and semantic procedures, but only prosodic information and a rather shallow analysis with (sequences of) word classes, i.e. part-of-speech information.

The spoken word sequence which is obtained from the speech recognizer is only required for the time alignment and for a normalization of energy and duration based on the underlying phonemes. In this paper, we use the transcription of the data assuming a recognizer with 100 % accuracy.

It is still an open question which prosodic features are relevant for different classification problems, and how the different features are interrelated. We try therefore to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the statistical classifier to find out the relevant features and the optimal weighting of them. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favor of the end of a word because the word is a well–defined unit in word

[4]designed and recorded at the Institute of Phonetics and Speech Communication, Ludwig-Maximilians-University, Munich

Table 3: *100 prosodic and 30 POS features and their context*

	context size				
	-2	-1	0	1	2
95 prosodic features:					
DurTauLoc; EnTauLoc; F0MeanGlob			•		
Dur: Norm,Abs,AbsSyl		•	•	•	
En: RegCoeff,MseReg,Norm,Abs,Mean,Max,MaxPos		•	•	•	
F0: RegCoeff,MseReg,Mean,Max,MaxPos,Min,MinPos		•	•	•	
Pause-before, PauseFill-before; F0: Off,Offpos		•	•		
Pause-after, PauseFill-after; F0: On,Onpos			•	•	
Dur: Norm,Abs,AbsSyl	•			•	
En: RegCoeff,MseReg,Norm,Abs,Mean	•			•	
F0: RegCoeff,MseReg	•			•	
F0: RegCoeff,MseReg; En: RegCoeff,MseReg; Dur: Norm		•			
5 more in the set with 100 features:					
Jitter: Mean, Sigma; Shimmer: Mean, Sigma;			•		
RateOfSpeech			•		
30 POS-features:					
API,APN,AUX,NOUN,PAJ,VERB	•	•	•	•	•

recognition, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. Many relevant prosodic features are extracted from different context windows with the size of two words before, that is, contexts -2 and -1, and two words after, i.e. contexts 1 and 2 in Table 3, around the current word, namely context 0 in Table 3; by that, we use so to speak a 'prosodic 5-gram'. A full account of the strategy for the feature selection is beyond the scope of this paper; details and further references are given in [2]. Table 3 shows the 95 prosodic features used in section 4 and their context; in the experiments described in section 5, we used five additional features: global mean and sigma for jitter and shimmer (JitterMean, JitterSigma, ShimmerMean, ShimmerSigma), and another global tempo feature (RateOfSpeech). The six POS features with their context sum up to 30. The mean values DurTauLoc, EnTauLoc, and F0MeanGlob are computed for a window of 15 words (or less, if the

utterance is shorter); thus they are identical for each word in the context of five words, and only context 0 is necessary. Note that these features do not necessarily represent *the* optimal feature set; this could only be obtained by reducing a much larger set to those features which prove to be relevant for the actual task, but in our experience, the effort needed to find the optimal set normally does not pay off in terms of classification performance [3, 4]. A detailed overview of prosodic features is given in [5]. The abbreviations of the 95 features can be explained as follows:

duration features 'Dur': absolute (Abs) and normalized (Norm); the normalization is described in [2]; the global value DurTauLoc is used to scale the mean duration values, absolute duration divided by number of syllables AbsSyl represents another sort of normalization;

energy features 'En': regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalized (Norm) values; the normalization is described in [2]; the global value EnTauLoc is used to scale the mean energy values, absolute energy divided by number of syllables AbsSyl represents another sort of normalization;

F0 features 'F0': regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis; all F0 features are logarithmised and normalised as to the mean value F0MeanGlob;

length of pauses 'Pause': silent pause before (Pause-before) and after (Pause-after), and filled pause before (PauseFill-before) and after (PauseFill-after).

A Part of Speech (POS) flag is assigned to each word in the lexicon, cf. [6]. Six cover classes are used: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns). For the context of +/- two words, this sums up to 6x5, i.e., 30 POS features, cf. the last line in Table 3.

4 Preliminary Experiments with a Subset of the SmartKom Data

The material used for the classification task and the interpretation in this chapter is a subset of the whole SmartKom database; it consists of 81 dialogues, 1172 turns, 10775 words, and 132 minutes of speech. 2.6% of the words were labelled as ROT, and 4.9% as OOT.

We computed a Linear Discriminant (LDA) classification: a linear combination of the independent variables (the predictors) is formed; a case is classified, based on its discriminant score, in the group for which the posterior probability is largest [13]. We simply took an a priori probability of 0.5 for the two or three classes and did not try to optimize, for instance, performance for the marked classes. For classification, we used the leave-one–case-out (*loco*) method; note that this means that the speakers are seen, in contrast to the LDA used in section 5 where the leave-one-speaker-out method has been employed. Tables 4 and 5 show the recognition rates for the two–class problem Off-Talk vs. no–Off-Talk and for the three–class problem ROT, OOT, and NOT, resp. Besides recall for each class, the *CL*ass–wise computed mean classification rate (mean of all classes, unweighted average recall) CL and the overall classification (*R*ecognition) *R*ate RR, i.e., all correctly classified cases (weighted average recall), are given in percent. We display results for the 95 prosodic features with and without the 30 POS features, and for the 30 POS features alone – as a sort of 5-gram modelling a context of 2 words to the left and two words to the right, together with the pertaining word 0. Then, the same combinations are given for a sort of uni-gram modelling only the pertaining word 0. For the last two lines in Tables 4 and 5, we first computed a principal component analysis for the 5-gram- and for the uni-gram constellation, and used the resulting principal components PC with an eigenvalue > 1.0 as predictors in a subsequent classification.

Best classification results could be obtained by using both all 95 prosodic features and all 30 POS features together, both for the two–class problem (CL: 73.7%, RR: 78.8%) and for the three–class problem (CL: 70.5%, RR: 72.6%). These results are emphasized in Tables 4 and 5. Most information is of course encoded in the features of the pertinent word 0; thus, classifications which use only these 28 prosodic and 6 POS features are of course worse, but not to a large extent: for the two–class problem, CL is 71.6%, RR 74.0%; for the three–class problem, CL is 65.9%, RR 62.0%. If we use PCs as predictors, again, classification performance goes down, but not

Table 4: *Recognition rates in percent for different constellations; subset of SmartKom, leave–one–case–out, Off-Talk vs. no-Off-Talk; best results are emphasized*

constellation	predictors	Off-Talk	no-Off-Talk	CL	RR
	# of tokens	806	9969	10775	
5-gram	95 pros.	67.6	77.8	72.7	77.1
raw feat. values	95 pros./30 POS	67.7	79.7	**73.7**	**78.8**
5-gram, only POS	30 POS	50.6	72.4	61.5	70.8
uni-gram	28 pros. 0	68.4	73.4	70.9	73.0
raw feat. values	28 pros. 0/6 POS 0	68.6	74.5	71.6	74.0
uni-gram, only POS	6 POS	40.9	71.4	56.2	69.1
5-gram, PCs	24 pros. PC	69.2	75.2	72.2	74.8
uni-gram, PCs	9 pros. PC 0	66.0	71.4	68.7	71.0

drastically. This corroborates our results obtained for the classification of boundaries and accents, that more predictors – ceteris paribus – yield better classification rates, cf. [3, 4].

Now, we want to have a closer look at the nine PCs that model a sort of uni-gram and can be interpreted easier than 28 or 95 raw feature values. If we look at the functions at group centroid, and at the standardized canonical discriminant function coefficients, we can get an impression, which feature values are typical for ROT, OOT, and NOT. Most important is energy, which is lower for ROT and OOT than for NOT, and higher for ROT than for OOT. (Especially absolute) duration is longer for ROT than for OOT – we'll come back to this result in section 6. Energy regression is higher for ROT than for OOT, and F0 is lower for ROT and OOT than for NOT, and lower for ROT than for OOT. This result mirrors, of course, the strategies of the labellers and the characteristics of the phenomenon 'Off-Talk': if people speak aside or to themselves, they do this normally in lower voice and pitch.

5 Results

In the following all databases are evaluated with an LDA-classifier and leave-one-speaker-out (*loso*) validation. All results are measured with the class-

Table 5: *Recognition rates in percent for different constellations; subset of SmartKom, leave–one–case–out, ROT vs. OOT vs. NOT; best results are emphasized*

constellation	predictors	ROT	OOT	NOT	CL	RR
	# of tokens	277	529	9969	10775	
5-gram	95 pros.	54.9	65.2	71.5	63.9	70.8
raw feat. values	95 pros./30 POS	71.5	67.1	73.0	**70.5**	**72.6**
5-gram, only POS	30 POS	73.3	52.9	54.7	60.3	55.1
uni-gram	28 pros. 0	53.1	67.7	64.0	61.6	63.9
raw feat. values	28 pros. 0/6 POS 0	69.0	67.1	61.5	65.9	62.0
uni-gram, only POS	6 POS	80.1	64.7	18.2	54.3	22.1
5-gram, PCs	24 pros. PC	49.5	67.7	65.3	60.8	65.0
uni-gram, PCs	9 pros. PC 0	45.8	62.6	60.0	56.1	59.8

wise averaged recognition rate CL-N (N = 2, 3, 4) to guarantee robust recognition of all N classes (unweighted average recall). In the 2-class task we classify On-Talk (NOT) vs. rest; for $N = 3$ classes we discriminate NOT, ROT and OOT (= SOT $\cup$ POT); the $N = 4$ classes NOT, ROT, SOT, POT are only available in $\mathrm{SW}_{\mathrm{spont}}$.

In Table 6 results on the different databases are compared. Classification is performed with different feature sets: 100 prosodic features, 30 POS features, or all 130 features. For $\mathrm{SW}_{\mathrm{acted}}$ POS-features are not evaluated, since all sentences that had to be uttered were given in advance; for such a non-spontaneous database POS evaluation would only measure the design of the database rather than the correlation of different Off-Talk classes with the "real" frequency of POS categories. For the prosodic features, results are additionally given after speaker normalization (zero-mean and variance 1 for all feature components). Here, we assume that mean and variance (independent whether On-Talk or not) of all the speaker's prosodic feature vectors are known in advance. This is an upper bound for the results that can be reached with adaptation.

As could be expected, best results on prosodic features are obtained for the acted data: 80.8 % CL-2 and even higher recognition rates for three classes, whereas chance would be only 33.3 % CL-3. Rates are higher for $\mathrm{SK}_{\mathrm{spont}}$ than for $\mathrm{SW}_{\mathrm{spont}}$ (72.7 % vs. 65.3 % CL-2, 60.0 % vs. 55.2 % CL-

Table 6: *Results with prosodic features and POS features; leave-one-speaker-out, class-wise averaged recognition rate for On-Talk vs. Off-Talk (CL-2), NOT, ROT, OOT (CL-3) and NOT, ROT, POT, SOT (CL-4)*

	features	CL-2	CL-3	CL-4
SK_{spont}	100 pros.	72.7	60.0	-
SK_{spont}	100 pros. speaker norm.	74.2	61.5	-
SK_{spont}	30 POS	58.9	60.1	-
SK_{spont}	100 pros. + 30 POS	74.1	66.0	-
SW_{spont}	100 pros.	65.3	55.2	48.6
SW_{spont}	100 pros. speaker norm	66.8	56.4	49.8
SW_{spont}	30 POS	61.6	51.6	46.9
SW_{spont}	100 pros. + 30 POS	68.1	60.0	53.0
SW_{acted}	100 pros.	80.8	83.9	-
SW_{acted}	100 pros. speaker norm	92.6	92.9	-

3).[5] For all databases results could be improved when the 100-dimensional feature vectors are normalized per speaker. The results for SW_{acted} rise drastically to 92.6 % CL-3; for the other corpora a smaller increase can be observed. The evaluation of 30 POS features shows about 60 % CL-2 for both spontaneous databases; for three classes lower rates are achieved for SW_{spont}. Here, in particular the recall of ROT is significantly higher for SK_{spont} (78 % vs. 57 %). In all cases a significant increase of recognition rates is obtained when linguistic and prosodic information is combined, e.g. on SW_{spont} three classes are classified with 60.0 % CL-3, whereas with only prosodic or only POS features 55.2 % resp. 51.6 % CL-3 are reached. For SW_{spont} 4 classes could be discriminated with up to 53.0 % CL-4. Here, POT is the problematic category that is very close to all other classes (39 % recall only).

Fig. 1 shows the ROC-evaluation for all databases with prosodic features. In a real application it might be more "expensive" to drop a request that is addressed to the system than to answer a question that is not addressed to the system. If we thus set the recall for On-Talk to 90 %, every third Off-Talk word is detected in SW_{spont} and every second in SK_{spont}. For the SW_{acted} data, the Off-Talk recall is nearly 70 %; after speaker normalization

[5]The reason for this is most likely that in SmartKom, the users were alone with the system; thus Off-Talk was always talking to one-self – no need to be understood by a third partner. In SmartWeb, however, a third partner was present, and moreover, the signal-to-noise ratio was less favorable than in the case of SmartKom.

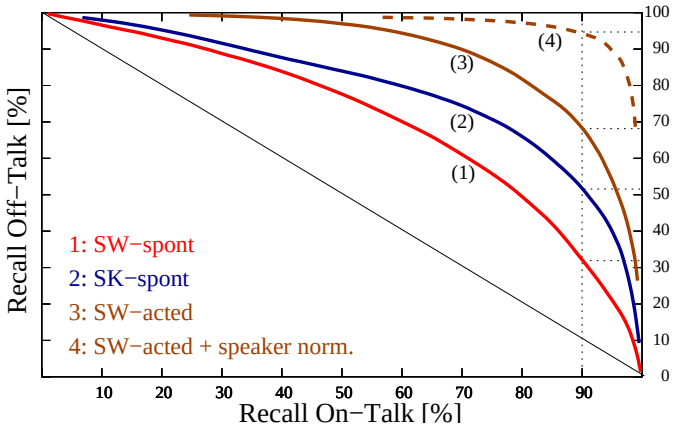

Figure 1: *ROC-Evaluation On-Talk vs. Off-Talk for the different databases*

Table 7: *Cross validation of the three corpora with speaker-normalized prosodic features. Diagonal elements are results for Train=Test (leave-one-speaker-out in brackets). All classification rates in % CL-2*

		Test		
		SW_{acted}	SW_{spont}	SK_{spont}
Training	SW_{acted}	93.4 (92.6)	63.4	61.9
	SW_{spont}	85.2	69.3 (66.8)	67.8
	SK_{spont}	74.0	61.1	76.9 (74.2)

it rises to 95 %.

To compare the different prosodic information used in the different corpora and the differences in acted and spontaneous speech, we use cross validation as shown in Table 7. The diagonal elements show the *Train=Test* case, and in brackets the *loso* result from Table 6 (speaker norm.). The maximum we can reach on $\mathrm{SW_{spont}}$ is 69.3 %, whereas with *loso*-evaluation 66.8 % are achieved; if we train with acted data and evaluate with $\mathrm{SW_{spont}}$, the drop is surprisingly small: we still reach 63.4 % CL-2. The other way round 85.2 % on $\mathrm{SW_{acted}}$ are obtained, if we train with $\mathrm{SW_{spont}}$. This shows, that both SmartWeb corpora are in some way similar; the database most related to $\mathrm{SK_{spont}}$ is $\mathrm{SW_{spont}}$.

Table 8: SW_{spont}: *Best single features for NOT vs. OOT (left) and NOT vs. ROT (right). Classification rate is given in* CL-2 *in %. The dominant feature group is emphasized. " •" denotes that the resp. values are greater for this type given in this column*

SW_{spont}	NOT	OOT	CL-2	SW_{spont}	NOT	ROT	CL-2
EnMax	•		61	*EnTauLoc*	•		60
EnTauLoc	•		60	***DurAbs***		•	58
EnMean	•		60	*F0Max**Pos***		•	58
PauseFill-before		•	54	*F0On**Pos***	•		57
JitterSigma	•		54	***DurTauLoc***	•		57
EnAbs	•		54	*EnMax**Pos***		•	56
F0Max	•		53	*EnMean*	•		56
ShimmerSigma	•		53	*EnAbs*		•	56
JitterMean	•		5	*F0Off **Pos***	•		55
Pause-before		•	53	*F0Min**Pos***		•	53

Table 9: SW_{acted}: *Best single features for NOT vs. OOT (left) and NOT vs. ROT (right)*

SW_{acted}	NOT	OOT	CL-2	SW_{acted}	NOT	ROT	CL-2
EnTauLoc	•		68	***DurTauLoc***		•	86
EnMax	•		68	*EnMax**Pos***		•	73
RateOfSpeech	•		65	***DurAbs***		•	71
F0MeanGlob	•		65	*EnMean*	•		71
EnMean	•		63	*F0Max**Pos***		•	69
ShimmerSigma	•		63	*EnMax*	•		69
F0Max	•		61	***DurAbsSyl***		•	68
EnAbs	•		61	*F0On**Pos***	•		68
F0Min		•	60	*F0Min**Pos***		•	65
ShimmerMean	•		60	*RateOfSpeech*	•		62

6 Discussion

As expected, results for spontaneous data were worse than for acted data (section 5). However, if we train with SW_{acted} and test with SW_{spont} and vice versa, the drop is just small. There is hope, that real applications can be enhanced with acted Off-Talk data. Next, we want to reveal similarities in the different databases and analyze single prosodic features. To discriminate On-Talk and OOT, all ROT words were deleted; for On-Talk vs. ROT, OOT is deleted. The top-ten best features are ranked in Table 8 for SW_{spont}, Table 9 for SW_{acted}, and Table 10 for SK_{spont}. For the case NOT vs. OOT the column CL-2 shows high rates for SW_{acted} and SK_{spont} with energy features; best results for NOT vs. ROT are achieved with duration features on SW_{acted}.

Most relevant features to discriminate On-Talk (**NOT**) vs. **OOT** (left column in Table 8, 9, 10) are the higher energy values for On-Talk, as well for the SW_{acted} data as for both spontaneous corpora. Highest results are achieved for SK_{spont}, since the user was alone and OOT is basically talking to oneself and consequently with extremely low voice. Also jitter and shimmer are important, in particular for SK_{spont}. The range of F0 is larger for On-Talk which might be caused by an exaggerated intonation when talking to computers. For SW_{acted} global features are more relevant (acted speech is more consistent), in particular the rate-of-speech that is lower for Off-Talk. Further global features are *EnTauLoc* and *F0MeanGlob*. Instead, for the more spontaneous SW_{spont} data pauses are more significant (longer pauses for OOT). In SK_{spont} global features are not relevant, because in many cases only one word per turn is Off-Talk (swearwords).

To discriminate **On-Talk** vs. **ROT** (right columns in Tables 8, 9, 10) duration features are highly important: the duration of read words is longer (cf. F0Max, F0Min). In addition, the duration is modeled with *Pos*-features: maxima are reached later for On-Talk.[6] Again, energy is very significant (higher for On-Talk). Most features show for all databases the same behavior but unfortunately there are some exceptions, probably caused by the instructions for the acted ROT: the global feature *DurTauLoc* is in SW_{acted} smaller for On-Talk, and in SW_{spont} and SK_{spont} smaller for ROT. Again, jitter is important in SK_{spont}.

To distinguish **ROT** vs. **OOT**, the higher duration of ROT is significant

[6]Note that these *Pos*-features are prosodic features that model the position of prominent pitch events on the time axis; if F0MaxPos is greater this normally simply means that the words are longer. These features should not be confused with POS, i.e. part-of-speech features which are discussed below in more detail.

as well as the wider F0-range. ROT shows higher energy values in SW_{spont} but only higher absolute energy in SW_{acted} which always rises for words with longer duration.[7] All results of the analysis of single features confirm our results from the principal component analysis in section 4.

For all classification experiments we would expect a small decrease of classification rates in a real application, since we assume a speech recognizer with 100 % recognition rate in this paper. However, when using a real speech recognizer, the drop is only little for On-Talk/Off-Talk classification: in preliminary experiments we used a very poor word recognizer with only 40 % word accuracy on SK_{spont}. The decrease of CL-2 was 3.2 % relative. Using a ROC evaluation, we can set the recall for On-Talk to 90 % as above by higher weighting of this class. Then, the recall for Off-Talk goes down from $\sim$ 50 % to $\sim$ 40 % for the evaluation based on the word recognizer.

Using all 100 features, best results are achieved with SW_{acted}. The classification rates for the SK_{spont} WoZ data are worse, but better than for the SW_{spont} data since there was no Off-Talk to another Person (POT). Therefore, we are going to analyze the different SW_{spont} speakers. Some of them yield very poor classification rates. It will be investigated, if it is possible for humans to annotate these speakers, without any linguistic information. We expect further that classification rates will rise if the analysis is performed turn-based. Last but not least, the combination with On-View/Off-View will increase the recognition rates, since especially POT, where the user does not look onto the display, is hard to classify from the audio signal. For the SW_{spont} video-data, the two classes On-View/Off-View are classified with 80 % CL-2 (frame-based) with the Viola-Jones face detection algorithm [18]. The multimodal classification of the focus of attention will result in *On-Focus*, the fusion of On-Talk and On-View.

The most important difference between ROT and OOT is not a prosodic, but a lexical one. This can be illustrated nicely by Tables 11 and 12 where percent occurrences of POS is given for the three classes NOT, ROT, and OOT (SK_{spont}) and for the four classes NOT, ROT, POT, and SOT (SW_{spont}). Especially for SK_{spont} there are more content words in ROT than in OOT and NOT, especially NOUNs: 54.9% compared to 7.2% in OOT and 18.9% in NOT. It is the other way round, if we look at the function words, especially at PAJ (particles, articles, and interjections): very few for ROT (15.2%), and most for OOT (64.7%). The explanation is straightforward: the user only reads words that are presented on the screen, and

[7] In this paper, we concentrate on Computer-Talk = On-Talk vs. Off-Talk; thus we do not display detailed tables for this distinction **ROT** vs. **OOT**.

these are mostly content words – names of restaurants, cinemas, etc., which of course are longer than other word classes. For SW_{spont}, there is the same tendency but less pronounced.

7 Concluding Remarks

Off-Talk is certainly a phenomenon the successful treatment of which is getting more and more important, if the performance of automatic dialogue systems allows unrestricted speech, and if the tasks performed by such systems approximate those tasks that are performed within these Wizard-of-Oz experiments. We have seen that a prosodic classification, based on a large feature vector yields good but not excellent classification rates. With additional lexical information entailed in the POS features, classification rates went up.

Classification performance as well as the unique phonetic traits discussed in this paper will very much depend on the types of Off-Talk that can be found in specific scenarios; for instance, in a noisy environment, talking aside to someone else might display the same amount of energy as addressing the system, simply because of an unfavourable signal-to-noise ratio.

We have seen that on the one hand, Computer-Talk (i.e. On-Talk) in fact is similar to talking to someone who is hard of hearing: its phonetics is more pronounced, energy is higher, etc. However we have to keep in mind that this register will most likely depend to some – even high – degree on other factors such as overall system performance: the better the system performance turns out to be, the more 'natural' the Computer-Talk of users will be, and this means in turn that the differences between On-Talk and Off-Talk will possibly be less pronounced.

Acknowledgments: This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the SmartKom project under Grant 01 IL 905 K7 and in the framework of the SmartWeb project under Grant 01 IMD 01 F. The responsibility for the contents of this study lies with the authors.

References

[1] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. Dialogue Acts in

VERBMOBIL-2 – Second Edition. Verbmobil Report 226, Juli 1998.

[2] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, pages 106–121. Springer, Berlin, 2000.

[3] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. ICPHS99*, pages 2315–2318, San Francisco, 1999.

[4] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In *Proc. of Eurospeech01*, pages 2781–2784, Aalborg, 2001.

[5] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication* , 40:117–143, 2003.

[6] A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proc. of Eurospeech99*, pages 519–522, Budapest, 1999.

[7] A. Batliner, V. Zeissler, E. Nöth, and H. Niemann. Prosodic Classification of Offtalk: First Experiments. In *Proc. of the Fifth International Conference on Text, Speech, Dialogue*, pages 357–364, Berlin, 2002. Springer.

[8] J. Carletta, N. Dahlbäck, N. Reithinger, and M. Walker. Standards for Dialogue Coding in Natural Language Processing. Dagstuhl-Seminar-Report 167, 1997.

[9] K. Fischer. *What Computer Talk Is and Is not: Human-Computer Conversation as Intercultural Communication*, volume 17 of *Linguistics - Computational Linguistics*. AQ, Saarbrücken, 2006.

[10] N. Fraser and G. Gilbert. Simulating Speech Systems. *CSL*, 5(1):81–99, 1991.

[11] C. Hacker, A. Batliner, and E. Nöth. Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In *Proc. of the Ninth International Conference on Text, Speech, Dialogue*, page to appear, Berlin, 2006. Springer.

[12] M. Katzenmaier, R. Stiefelhagen, and T. Schultz. Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech. In *Proc. of the Sixth International Conference on Multimodal Interfaces (ICMI 2004)* , pages 144–151, 2004.

[13] W. Klecka. *Discriminant Analysis*. SAGE PUBLICATIONS Inc., Beverly Hills, 9 edition, 1988.

[14] R. Lunsford. Private Speech during Multimodal Human-Computer Interaction. In *Proc. of the Sixth International Conference on Multimodal Interfaces (ICMI 2004)*, page 346, Pennsylvania, 2004. (abstract).

[15] D. Oppermann, F. Schiel, S. Steininger, and N. Beringer. Off-Talk – a Problem for Human-Machine-Interaction. In *Proc. Eurospeech01*, pages 2197–2200, Aalborg, 2001.

[16] N. Reithinger, S. Bergweiler, R. Engel, G. Herzog, N. Pfleger, M. Romanelli, and D. Sonntag. A Look Under the Hood - Design and Development of the First SmartWeb System Demonstrator. In *Proc. of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, Trento, Italy, 2005.

[17] R. Siepmann, A. Batliner, and D. Oppermann. Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, pages 147–150, Red Bank, N.J., 2001.

[18] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[19] W. Wahlster. Smartweb: Mobile Application of the Semantic Web. *GI Jahrestagung 2004*, pages 26–27, 2004.

[20] W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In *Proc. Eurospeech01*, pages 1547–1550, Aalborg, 2001.

[21] P. Watzlawick, J. Beavin, and D. D. Jackson. *Pragmatics of Human Communications*. W.W. Norton & Company, New York, 1967.

Table 10: SK_{spont}: *Best single features for NOT vs. OOT (left) and NOT vs. ROT (right)*

SK_{spont}	NOT	OOT	CL-2
EnMax	•		72
EnMean	•		69
JitterMean	•		69
JitterSigma	•		69
F0Max	•		69
ShimmerSigma	•		68
ShimmerMean	•		68
F0OnPos		•	67
EnAbs	•		66
EnNorm	•		61

SK_{spont}	NOT	ROT	CL-2
JitterMean	•		62
DurAbs		•	61
DurTauLoc	•		61
*F0Max**Pos***		•	61
EnTauLoc	•		69
*F0Min**Pos***		•	59
JitterSigma	•		59
EnMean	•		59
EnMax	•		58
F0Max	•		58

Table 11: SK_{spont}: *POS classes, percent occurrences for NOT, ROT, and OOT*

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
NOT	19415	18.1	2.2	6.6	9.6	8.4	55.1
ROT	365	56.2	7.1	18.1	2.2	2.2	14.2
OOT	889	7.2	2.6	10.7	8.9	6.7	63.9
total	20669	18.3	2.3	7.0	9.4	8.2	54.7

Table 12: SW_{spont}: *POS classes, percent occurrences for NOT, ROT, POT, and SOT*

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
NOT	2541	23.2	5.1	3.8	6.9	8.5	52.5
ROT	684	27.2	5.7	18.6	7.4	7.6	33.5
POT	1093	26.3	5.1	10.3	5.4	9.5	43.3
SOT	893	8.1	1.5	5.7	11.5	10.3	62.9
total	5211	21.8	4.6	7.4	7.5	8.9	49.8

How People Talk to a Virtual Human - Conversations from a Real-World Application

Stefan Kopp
A.I. Group, University of Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

Abstract

This paper describes a study on the kinds of dialogue human users are willing to have with the virtual human Max in a real-world scenario. Max is employed as guide in a public computer museum, where he engages with visitors in embodied face-to-face communication and provides them with information about the museum or the exhibition. Visitors can input natural language input using a keyboard. Logfiles from interactions between Max and museum visitors were analyzed. Results show that Max engages people in interactions where they are likely to use a variety of normal human communication strategies and the language this entails, also indicating attribution of sociality to the agent.

1 Introduction

During the last 15 years or so natural language interaction with computer systems has been increasingly augmented with ways of using non-verbal modalities along with speech. Embodied conversational agents (ECA, in short) can be seen as the most ambitious form of such interfaces, namely, virtual humans that are to be capable of understanding and generating all of the communicative behaviors that humans show in natural face-to-face dialog. When we ask how people talk to computers, it makes thus sense to further ask how they would interact with such virtual humans, and how the embodied appearance and multimodal behavior of a virtual interlocutor affects user behavior. Unfortunately, current ECAs have very rarely made the step out of the laboratories into real-world settings so that we have only little data on how people would interact with these agents in real-world applications. In this paper we present results on how human users interact

with the virtual human Max, under development at the A.I. group at Bielefeld University [7]. The interactions that have been analyzed took place not under controlled laboratory condtions, but in public place and without being monitored by experimenters: Max is applied as an information kiosk in the Heinz-Nixdorf-MuseumsForum (HNF; see Fig. 1), a public computer museum in Paderborn (Germany), where he engage visitors in face-to-face smalltalk conversations and provides them with information about the museum, the exhibition, and other topics daily since January 2004. Visitors can give natural language input to the system using a keyboard, whereas Max is to respond with a synthetic German voice and appropriate nonverbal behaviors like manual gestures, facial expressions, gaze, or locomotion. Using log files from more than 3.500 conversations we have studied the communications that take place between Max and the visitors. In particular, we were interested in the kind of dialogs that the museum visitors – unbiased people with various backgrounds, normally not used to interact with an ECA – are willing to have with Max and whether these bear some resemblance with human-human dialogues.

Figure 1: Max in the Heinz-Nixdorf-MuseumsForum.

1.1 How people talk to computers

Several studies have shown social effects of embodied agents, i.e. emotional, cognitive, or behavioral reactions similar to those reactions shown during interactions with human beings. In general, humans tend to apply their strategies of perceiving and understanding other people also when interacting with computers. For example, just like other humans, agents are evaluated to be more intelligent when they criticise others, or to be more likeable when giving positive feedback [8] (Nass, Steuer & Tauber, 1994). Trust and credibility of a computer system is increased when an anthropomorphic interface is used [10, 12, 9]. Also, effects of impression management and self-presentation were shown to be present in interactions with computers. That is, people tend to present themselves in a more favourable way [10, 5], when being observed by an artificial character. Likewise, they try harder and perform better when a computer has human-like features [12], but can also by more anxious and tend to make more mistakes when feeling monitored by an agent [9].

As for how people communicate with computer systems, it has been noted that humans are willing to apply human-like communication strategies in such interactions. This occurs even when talking to *disembodied* chatterbots [2, 6], although such dialogues vary in length, topic, and style, and people tend to use a simpler language. Nevertheless, one finds greetings, thanks, direct and indirect expressions of courtesy, attribution of moods, feelings and intentions to the system. Further, people ask intimate questions, assuming that the system has inner states to reveal (self-disclosure). These effects are even increased when embodied agents with a human-like appearance are encountered as interlocutors. It has been shown that such agents prompt communication per se and trigger the use of natural language interaction, as opposed to other direct forms of operating the system [4]. That is, embodied agents lead to higher expectations as to what interacive capabilities the system may have, as evident, e.g., in reciprocal communication attempts such as correcting comments or resignation utterances. When the agents make good use of nonverbal behavior, a facilitative effect on the communication has been reported. For example, the face of an agent is being attended to and interpreted for communicative feedback [11]. Remarkably, when the agent gives turn-taking feedback, displays attentional cues, and marks utterances with beat gestures, human users give higher subjective ratings of the system's language capability and communicate more smoothly, i.e. with fewer repetitions and hesitations [1]. These findings clearly show the benefits that embodied characters could potentially have

for spoken language man-machine interaction when they show consistent and pertinent nonverbal behaviors.

2 The virtual human Max

This section briefly explains the model of interactive behavior that underlies Max's behavior in the multimodal dialogues he has with visitors (see [3, 7] for more details). Max is construed as a general cognitive agent, based on an architecture that allows perception, action, and deliberative reasoning to run in parallel. Perception and action are directly connected through a reactive component, affording reflexes and immediate responses to situation events or input by a dialogue partner. Reactive processing is realized by a behavior generation component that is in charge of realizing all behaviors requested by other components. This includes feedback-driven reactive behaviors like gaze tracking the current interlocutor, or secondary behaviors like eye blink and breathing. Moreover, to realizes multimodal utterances by combining the synthesis of prosodic speech and animation of emotional facial expressions, lip-sync speech, and coverbal gestures, with the scheduling and synchronous execution of all verbal and nonverbal behaviors.

Deliberative processing of all events takes place in a central component. It determines when and how the agent acts, either driven by internal goals and intentions or in response to incoming events which, in turn, may originate either externally (user input, persons that have newly entered or left the agent's visual field) or internally (changing emotions, assertion of a new goal etc.). These deliberative processes are carried out by a BDI interpreter, which continuously pursues multiple, possibly nested plans (*intentions*) to achieve goals (*desires*) in the context of up-to-date knowledge about the world (*beliefs*). It draws on long-term knowledge about former dialogue episodes with visitors as well as a dynamic knowledge base that includes a discourse model, a user model, as well as a self model that comprises the agent's world knowledge as well as current goals and intentions.

All capabilities of dialogue management, language interpretation and behavior generation are represented as plans of two kinds. *Skeleton plans* realize the agent's general, domain-independent dialogue skills like negotiating initiative or structuring a presentation. These plans are adjoined by a larger number of smaller plans implementing *condition-action rules* that define both, the broad conversation knowledge (e.g., dialogue goals that can be pursued, possible interpretations of input, small talk answers) as well as the deeper knowledge about possible presentation contents. Condition-

action rules test either user input or the dynamic memories; their actions can alter dynamic knowledge structures, raise internal goals and thus invoke corresponding plans, or trigger the generation of an utterance by stating the words, semantic-pragmatic aspects, and a markup of the focus part. Using these rules, the deliberative component interprets an incoming event, decides how to react depending on current context, and produces an appropriate response. It is thereby able to conduct longer, coherent dialogues and to act proactively, e.g. to take over the initiative, instead of being purely reactive as classical chatterbots are. In its current state, Max is equipped with roughly 900 skeleton plans and 1.200 rule plans of conversational and presentational knowledge.

Max is further equipped with an emotion system that continuously runs a dynamic simulation to model the agent's emotional state. The emotional state is available anytime and modulates subtle aspects of the agent's behaviors, namely, the pitch, speech rate, and band width of his voice and the rates of breathing and eye blink. The weighted emotion category is mapped to Max's facial expression and is sent to the agent's deliberative processes, thus making him cognitively aware of his own emotional state and subjecting it to his further deliberations. The emotion system, in turn, receives input from both the perception (e.g., seeing a person triggers a positive stimulus) and the deliberative component. For example, obscene or politically incorrect wordings in the user input lead to negative impulses on Max's emotional system.

3 How Humans Talk To Max

In the HNF scenario, we were able to unobtrusively gather a tremendous amount of data on the interactions between Max and the visitors to the museum. This data comprise transcripts of what Max and the human user said, as well as information about which nonverbal actions Max performed and when he did so. We analyzed these data to see (1) if Max's conversational capabilities suffice to fluent interactions with the visitors to the museum, and (2) whether the dialogs bear some resemblance with human-human dialogs, i.e. if Max is perceived and treated as human-like communication partner.

3.1 Study 1

A first screening was done after the first seven weeks of Max's employment in the Nixdorf Museum (15 January through 6 April, 2004). Statistics is based

on digital logfiles, which were recorded from dialogues between Max and visitors to the museum. During this period, Max on average had 47 conversations daily, where "conversation" was defined to be the discourse between an individual visitor saying hello and good bye to Max. Altogether there were 3351 conversations, i.e. logfiles screened. About two-thirds of these were conversations with male visitors and about one-third were conversations with female visitors, as identified by given names and Max's names dictionary. On the avarage, there were 15.33 visitor inputs recorded per logfile, totaling to 51,373 inputs recorded in the observation period.

Data were evaluated with respect to the successful recognition of communicative functions by Max, that is, whether Max associated a visitor's want (not necessarily correctly) with an input. We found that, Max was able to recognize a communicative function in 32,332 (i.e. 63%) cases. This finding suggests that in roughly two-thirds of all cases, Max conducted sensible dialogue with visitors, reverting to smalltalk behavior in the remaining cases where no communicative function could be recognized. Among those cases where a communicative function was recognized, with overlap possible, a total of 993 (1.9%) inputs were classified as polite ("please", "thanks"), 806 (1.6%) inputs as insulting, and 711 (1.4%) inputs as obscene or politically incorrect, with 1430 (2.8%) no-words altogether. In 181 instances (about 3 times a day), accumulated negative emotions resulted in Max leaving the scene "very annoyed".

A qualitative conclusion from the findings of this first screening is that Max apparently "ties in" visitors of the museum with diverse kinds of social interaction. Thus we conducted a second study with the particular aim to investigate in what ways and to what an extent Max is able to engage visitors in social interaction.

3.2 Study 2

We conducted a detailed content analysis of the users' statements during their dialogue with Max. Specifically, we wanted to know whether people would use human-like communication strategies (greetings, farewells, commonplace phrases), and whether they would use utterances or pose questions that indicate the attribution of sociality to the agent, e.g., by asking anthropomorphized questions that only make sense when directed to a human being. We analysed logfiles of one week in March 2005 (15th through 22nd) that contained all utterances of the agent as well as of the user. The data comprised 205 dialogs. The numbers of utterances, words, words per utterance, and specific words such as I/me or you were counted and compared

for agent and user. Additionally, the content of the users' utterances was coded according to psychological content analysis (Mayring, 2000). Using one third of the log file protocols, a category scheme was developed (e.g., questions, feedback to agent, requests to do something, etc., including corresponding values; see table 1). Subsequently, the complete material was coded by two coders and the frequency of each value was counted. Multiple selections were possible, e.g., one utterance may be coded as proactive as well as anthropomorphic question.

Quantitative analyses showed that the agent is more active than the user is. While the user makes 3665 utterances during the 205 dialogues (on average 17.88 utterances per conversation), the agent has 5195 turns (25.22 utterances per conversation). This is reflected in the words used. Not only does the agent use more words in total (42802 in all dialogues vs. 9775 of the user; 207.78 in average per conversation vs. 47.68 for the user), but he also uses more words per utterance (7.84 vs. 2.52 of the user). Thus, the agent in average seemed to produce more elaborate sentences than the user does, which may be a consequence of the use of a keyboard as input device. Against this background, it is also plausible that the users utters less pronouns such as I/me (user: 0.15 per utterance; agent: 0.43 per utterance) and you (user: 0.26 per utterance; agent: 0.56 per utterance). These results might be due to the particular dialogue structure that is, for some part, designed to be determined by the agent's questions and proposals (e.g., it includes an animal guessing game that leaves the user stating yes or no). On the other hand, the content analyses reveal that 1316 (35.9 %) of the user utterances are proactive (see table 1).

In order to analyse user reactions it is important to look at the content of user utterances. Table 1 shows the frequencies of different categories and the corresponding values. Concerning human-like strategies of beginning and ending conversations, it becomes apparent that especially greeting is also popular when confronted with an agent (used in 57.6% of dialogues). Greetings, which may be directly triggered by the greeting of the agent, are uttered more often than farewells. But, given that the user can end the conversation by simply stepping away from the system, it is remarkable that 29.8% of the people said goodbye to Max. This tendency to use human-like communicative structures is also supported by the fact that commonplace phrases,common small talk questions like 'How are you?' are still uttered 154 times (4.2% of utterances). As with all publicly available agents or chatterbots, we observed flaming (406 utterances; 11.1%) and implicit testing of intelligence and interactivity (303; 8.3%). The latter happens via questions (146; 4%), obviously wrong answers (61; 1.7%), answers in foreign languages

(30; 0.82%), or utterances to test the system (66; 1.8%). However, direct user feedback to the agent is more frequently positive (51) than negative (32). Most elucidating with regard to whether interacting with Max has social aspects are the questions addressed to him: There were mere comprehension questions (139; 18.6% of questions), questions to test the system (146; 19.6%), questions about the system (109; 14.6%), the museum (17; 2.3%), or something else (49; 6.6%). The vast amount of questions are social, either since they are borrowed from human small talk habits (commonplace phrases; 154; 20.6%) or because they directly concern social or human-like concepts (132; 17.7%). Thus, more than one-third of the questions presuppose that treating Max like a human is appropriateor try to test this very assumption. Likewise, the answers of the visitors (30% of all utterances) show that people seem to be willing to get involved in dialogue with the agent: 75.8% of them were expedient and inconspicuous, whereas only a small number gave obviously false information or aimed at testing the system. Thus, users seem to engage in interacting with Max and try to be cooperative in answering his questions.

4 Conclusion

Current embodied conversational agents have for the most part stayed within their lab environments and there is little data on how people interact with such conversational characters in real-world applications. One could expect the often described, general disposition of humans to approach an artifact like a social being, even more so when the artifact is an agent with human-like appearance and animated behaviour. Our study seems to support this. However, the behaviour of the users also shows that they are not at all sure in how far this expectation can be met by the system. It is an open question as to what degree the language employed by users accommodates these beliefs, and how it changes over discourse with growing evidence on Max's capabilities and limitations. A study is underway to take a more detailed look at the linguistic aspects of the user language. Nevertheless, we found evidence that the visitors to the HNF tend to apply a variety of human-like communication strategies when conversing with Max (greeting, farewell, smalltalk elements, insults), and they do so using short, yet close to everyday natural language utterances. This becomes apparent in particular when people try to fasten down the degree of Max's human-likeness employing normal language. It seems that they do not wonder about the language capability of the system as much as they wonder about its world

Catgeory and values	Examples	*N*
Proactive utterance		1316 (36%)
Reactive utterance		1259 (34%)
Greeting and farewell		
Informal greeting	Hi, hello	114
Formal greeting	Good morning!	4
No greeting		87
Informal farewell	Bye	56
Formal farewell	Farewell	5
No farewell		144
Flaming		406 (11%)
Abuse, name-calling	Son of a bitch	198
Pornographic utterances	Do you like to ****?	19
Random keystrokes		114
Senseless utterances	http.http, dupa	75
Feedback to agent		83 (2%)
Positive feedback	I like you; You are cool	51
Negative feedback	I hate you; Your topics are boring	32
Questions		746 (20%)
Anthropomorphic questions	Can you dance? Are you in love?	132
Questions concerning the system	Who has built you?	109
Questions concerning the museum	Where are the restrooms?	17
Commonplace phrases	How are you?	154
Questions to test the system	How's the weather?	146
Checking comprehension	Pardon?	139
Other questions		49
Answers		1096 (30%)
Inconspicuous answer		831
Apparently wrong answers	[name] Michael Jackson, [age] 125	61
Refusal to answer	I do not talk about private matters	8
Proactive utterances about oneself	I have to go now	76
Answers in foreign language		30
Utterances to test the system	You are Michael Jackson	66
Laughter		24
Requests		108 (3%)
General request to say something	Talk to me!	10
Specific request to say something	Tell me about the museum!	13
Request to stop talking	Shut up!	24
Request for action	Go away! Come back!	61

Table 1: Results of the content analysis of user dialogues with Max in the HNF.

knowledge or general intelligence. In how far this impression is induced by Max's appearance or the way he acts and reacts remains to be investigated in more controlled studies.

References

[1] J. Cassell and K. R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(45):519–539, 1996.

[2] A. De Angeli, G. Johnson, and L. Coventry. The unfriendly user: exploring social reactions to chatterbots. In K. Helander and Tham, editors, *Proceedings of The International Conference on Affective Human Factors Design*, London, 2001. Asean Academic Press.

[3] S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth. A conversational agent as museum guide – design and evaluation of a real-world application. In *Intelligent Virtual Agents*, LNAI 3661, pages 329–343. Springer-Verlag, 2005.

[4] N. Krämer. Social communicative effects of a virtual program guide. In *Intelligent Virtual Agents*, pages 442–543, 2005.

[5] N. Krämer, G. Bente, and J. Piesk. The ghost in the machine. the influence of embodied conversational agents on user expectations and user behaviour in a tv/vcr application. In G. Bieber and T. Kirste, editors, *IMC Workshop 2003, Assistance, Mobility, Applications*, pages 121–128, 2003.

[6] M. Leaverton. Recruiting the chatterbots. Cnet Tech Trends, 10/2/00 (http://cnet.com/techtrends/0-1544320-8-2862007-1.html.), 2000.

[7] I. W. N. Lessmann, S. Kopp. Situated interaction with a virtual human - perception, action, and cognition. In G. Rickheit and I. Wachsmuth, editors, *Situated Communication*, pages 287–323. Mouton de Gruyter, 2006.

[8] C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In B. Adelson, S. Dumais, and J. Olson, editors, *Human Factors in Computing Systems: CHI-94 Conference Proceedings*, pages 72–78. ACM Press, 1994.

[9] R. Rickenberg and B. Reeves. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In *Letters of CHI 2000*, pages 49–56. 2000.

[10] L. Sproull, M. Subramani, S. Kiesler, J. H. Walker, and K. Waters. When the interface is a face. *Human Computer Interaction*, 11(2):97–124, 1996.

[11] A. Takeuchi and T. Naito. Situated facial displays: towards social interaction. In *Human factors in computing Systems: CHI 95 Conference Proceedings*, pages 450–455, 1995.

[12] J. H. Walker, L. Sproull, and R. Subramani. Using a human face in an interface. In B. Adelson, S. Dumais, and J. Olson, editors, *Human Factors in Computing Systems: CHI94 Conference Proceeding*, pages 85–91. ACM, 1994.

The Role of Users' Preconceptions in Talking to Computers and Robots

Kerstin Fischer
University of Bremen, Germany
kerstinf@uni-bremen.de

Abstract

Communication with artificial interaction partners differs in many ways from communication among humans, and often so in the very first utterance. That is, in human-computer and human-robot interaction users address their artificial communication partner on the basis of preconceptions. The current paper addresses the nature of speakers' preconceptions about robots and computers and the role these preconceptions play in human-computer and human-robot interactions. That is, I will show that a) two types of preconceptions as opposing poles of the same dimension of interpersonal relationship can be distinguished, b) these types can be readily identified on the basis of surface cues in the users' utterances, b) these preconceptions correlate with the users' linguistic choices on all linguistic levels, and d) these preconceptions also influence the speakers' interactional behaviour, in particular, with respect to which their linguistic behaviour can be influenced, that is, in how far speakers align with the computer's and robot's linguistic output.

1 Introduction

When we look at the literature available on how people talk to computers and robots, it soon becomes clear that people talk to artificial communication partners differently from how they talk to other humans. This has led to the proposal that speech directed at artificial communication partner constitutes a register, so-called *computer talk* [27, 14]. When we keep looking, however, it turns out that in fact we know very little both about the exact nature of users' preconceptions about artificial communication partners and the effect these preconceptions have on human-computer, or human-robot, interaction situations.

In this paper I will propose that there are two prototypes of users' preconceptions, which can be reliably identified on the basis of linguistic surface

cues and which have systematic effects on the linguistic properties of users' utterances. Thus, I show that the speakers' recipient design, i.e. their choosing of linguistic properties on the basis of their concept of their communication partner, is pervasive and plays a central role in the formulation both of every single utterance and with respect to all linguistic levels.

Previous research has shown that recipient design [22, 23] and audience design [1] play a major role in the communication among humans. Recently, there is an ongoing debate about how much knowledge about the communication partner exactly speakers take into account [9, 10] and under what circumstances; however, it is clear that speakers take their communication partners into account to some degree [24]. How such models are being built up, what exactly speakers take into account when building up such models, and how these models influence the speech produced for the respective partner is so far an unresolved issue (see also the contributions by Branigan and Pearson, this volume; Wrede et al., this volume; Andonova, this volume). Thus, particularly in human-computer and human-robot interaction, we yet don't know much about the preconceptions on the basis of which users tailor their speech for their artificial communication partners and in which ways.

Moreover, users in human-computer interaction are usually treated as a homogeneous group (see, for example, the studies in [14] or Gieselmann and Stenneken, this volume; Kopp, this volume; Batliner et al., this volume; Porzel, this volume). If at all, external sociolinguistic variables, such as age or gender, domain knowledge or familiarity with computers are being considered: "Explicit data capture involves the analysis of data input by the user, supplying data about their preferences by completing a user profile. Examples of explicit data captured are: age, sex, location, purchase history, content and layout preferences." [2], where implicit data elicitation is taken to involve the examination of server logs and the implementation of cookies for the identification of users' "different goals, interests, levels of expertise, abilities and preferences" [12]. User modeling should however not be restricted to factors related to the task or domain, since, as I am going to show, the users' preconceptions about such interfaces themselves cause considerable differences in users' linguistic behaviour.

Another open issue is the influence of the speakers' preconceptions on the interactional dynamics; the question is whether, besides influencing the users' linguistic choices, their recipient design also determines the discourse flow. I am going to demonstrate that such concepts have considerable influence on the users' alignment behaviour [20, 19], see also Branigan and Pearson, this volume.

2 Methods and Data

The procedure taken here is to analyse first speakers' preconceptions of their artificial communication partners as they become apparent in several corpora of human-computer and human-robot interaction. There are various possibilities to study speakers' concepts about their communication partner; one is to elicit speakers' ideas about their communication partner by means of questionnaires; this method is used, for instance, by Andonova, this volume, and by Wrede et al., this volume. In contrast, the methodology used here is essentially ethnomethodological; that is, I focus on speakers' common sense reasoning underlying their linguistic behaviour by orienting to their own displays of their understanding of the affordances of the situation. For instance, speakers will produce displays of their concepts about the communication partner in their clarification questions, but also in their reformulations. For example, the question directed at the experimenter *does it see anything?* shows that the user suspects the robot to be restricted in its perceptual capabilities and, moreover, that the speaker regards the robot as an *it*, a machine, rather than another social interactant. The reformulation in example (1) shows that the speaker suspects the robot to understand an extrinsic spatial description if it doesn't understand a projective term:

(1) S: go left
R: error
S: go East

From such displays, especially if they turn out to be systematic and recurrent both between speakers as well as within the same speaker's speech through time, we can infer what preconceptions the speakers hold about their artificial communication partner and what strengths and weaknesses they ascribe to it.

In addition, I use quantitative analyses to identify differences in distributions of particular linguistic properties as effects of the speakers' differing preconceptions about computers and robots.

The corpora I use were elicited in Wizard-of-Oz scenarios in order to ensure that all users are confronted with the same computer or robot behaviour. That is, the linguistic and other behaviour of the artificial system is produced by a human wizard but on the basis of a fixed schema of behaviours. In this way I can control for inter- and even intrapersonal variation [6]. Speakers (just) get the impression that the system is not functioning well. Besides comparability, another advantage is therefore that

the repeated use of system malfunction encourages the users to reformulate their utterances frequently and thus to reveal their hypotheses about their artificial communication partner.

Human-Computer Appointment Scheduling Corpus This corpus consists of 64 German and 8 English human-computer appointment scheduling dialogues (18-33 min each). The corpus was recorded in a Wizard-of-Oz scenario in the framework of the Verbmobil project [26]. Speakers are confronted with a fixed pattern of (simulated) system output which consists of sequences of acts, such as messages of failed understanding and rejections of proposals, which are repeated in a fixed order. The fixed schema of sequences of prefabricated system utterances allows us to identify how each speaker's reactions to particular types of system malfunctions change over time. It also allows the comparison of the speakers' use of language interpersonally. The impression the users have during the interaction is that of communicating with a malfunctioning automatic speech processing system, and the participants were indeed all convinced that they were talking to such a system. The data were transcribed and each turn was labelled with a turn ID that shows not only the speaker number, but also the respective position of the turn in the dialogue. Subsequently, the data were annotated for prosodic, lexical, and conversational properties. <P>, <B>, <L> stand for pause, breathing, and sylable lengthening respectively.

Human-Robot Distance Measurement Corpus The second corpus used here was elicited in a scenario in which the users' task was to instruct a robot to measure the distance between two objects out of a set of seven. These objects differed only in their spatial position. The users typed instructions into a notebook, the objects to be referred to and the robot being placed on the floor in front of them. The relevant objects were pointed at by the instructor of the experiments. There were 21 participants from all kinds of professions and with different experience with artificial systems. The robot's output was generated by a simple script that displayed answers in a fixed order after a particular 'processing' time. Thus, the dialogues are also comparable regarding the robot's linguistic material, and the users' instructions had no impact on the robot's linguistic behaviour. The robot, a Pioneer 2, could not move either, but the participants were told that they were connected to the robot's dialogue processing system by means of a wireless LAN connection. Participants did not doubt that they were talking to an automatic dialogue processing system, as is apparent from their answers

to the question: "If the robot didn't understand, what do you think could have been the cause?". The robot's output was either "error" (or a natural language variant of it) or a distance in centimeters. Since by reformulating their utterances the users display their hypotheses about the functioning of the system (see above), error messages were given frequently.

The user utterances are typed and thus transcription was not necessary; typos were not corrected. The turn IDs show the speaker number, for instance, usr-20, and the number of the turn in the dialogue.

Human-Robot Spatial Instruction Corpus This corpus was elicited with three different robots, Sony's Aibo, Pioneer, another commercially available robot, and Scorpion, built by colleagues at the University of Bremen [25]. Since we used a Wizard-of-Oz scenario, we were able to confront all users again with identical non-verbal robot behaviours, independent of the users' utterances. We elicited 30 English dialogues, using the same speakers, scheduling the recordings at least three months apart, and 66 German dialogues, in which we recruited naive users for each scenario. Here, we elicited 12 dialogues with Aibo, 33 with pioneer and 21 with scorpion.

The users' task was to instruct the respective robot to move to objects which were placed on the floor in front of them and which were pointed at by the experimenter. All robots moved between the objects in the same, predefined, way (there was no linguistic output).

The dialogues were transcribed and analysed with respect to their linguistic properties. Each turn ID shows whether the robot addressed was Aibo (A), Scorpion (S), or Pioneer (P). Transcription conventions are the following: (at=prominent) word (/a) means that the word is uttered in a prosodically prominent way, + indicates a word fragment, - means a short pause, – a longer pause, and (1) indicates a pause of one second; punctuation indicates the intonation contour with which the utterance was delivered.

Human-Aibo Interaction with and without Verbal Feedback For the comparison with the human-Aibo dialogues from the previous corpus, we elicited another corpus in the same scenario as before, just that Aibo also replied with verbal behaviours. The robot utterances were pre-synthesized and were played in a fixed order. The utterances were so designed as to give no clue as to what may have gone wrong in order to avoid prompting particular error resolution strategies from the users. However, in these utterances, three design features were used which previous studies [15, 3, 6] had revealed to be quite rare in human-robot interaction if the robot does not give feed-

back: First, we made the robot ask for and propose spatial references using object naming strategies. Second, we made the robot use an extrinsic reference system. Third, as an indicator of high linguistic capabilities, the robot made extensive use of relative clauses.

The robot's utterances are, for instance, the following: Ja, guten Tag, wie geht es Ihnen? (*yeah hello, how do you do?*) Soll ich das blaue Objekt ansteuern? (*do you want me to aim at the blue object?*) Soll ich mich zu dem Objekt begeben, das vorne liegt? (*do you want me to move to the object which lies in front?*) Meinen Sie das Objekt, das 30 Grad westlich der Dose liegt? (*do you mean the object that is 30 degrees west of the box?*) Ich habe Sie nicht verstanden. (*I did not understand.*) Entschuldigung, welches der Objekte wurde von Ihnen benannt? (*excuse me, which object was named by you?*) Ich kann nicht schneller. (*I can't go faster.*)

The corpus comprises 17 German human-Aibo dialogues recorded under circumstances exactly as in the corpus described above, just that the fixed schema of robot behaviours was paired with a fixed schema of robot utterances, both independent of what the speaker is saying.

3 Concepts about Computers and Robots

There are some beliefs about computers and robots that surface frequently and in all of the corpora under consideration. The first one is the concept of the computer or robot as linguistically restricted. This view of the artificial communication partner is in fact only encouraged in the human-computer interaction corpus when the system produces *I did not understand*. In the corpora in which the robot does not produce any speech, no such clues are given. Similarly, in the distance measurement corpus, only *error*-messages are produced, and thus the idea that the robot could be linguistically challenged is likely to stem from the speakers' preconceptions. Even more crucially, also in another corpus in which the linguistic capabilities of the robot were actually very good and in which communicative failure resulted from mismatches in instruction strategies [15], not in restricted linguistic capabilities, speakers overwhelmingly suspected the problem to have been that they weren't able to find those words that the robot would have been able to understand.

This preconception of artificial communication partners as linguistically restricted can turn out to be very problematic in the future; if our systems are getting better and the interfaces more natural, yet users continue to expect great linguistic problems, the interactions with such systems may

turn out very strange, as can be seen in the following example:

(2) R: yes, hello, how do you do?
A031: (4) oh okay. - um - um go forward, to, -

Here, the user does not react at all to the polite interaction proposed by the system. The rejection of such speech acts has to be attributed to the user's preconceptions, since at that point there is no evidence of miscommunication or communicative failure. This corresponds to findings by Krause [13] as well as to observation regarding politeness by [16, 21] and [11].

Another aspect is the suspected formality of artificial communication partners. In the following example, the speaker reformulates her utterance by using exact measurements:

(3) A003: nun zu den, zwei, Dosen, – links. (5) (now to the, two, boxes, – left)
R: Ich habe Sie nicht verstanden. (I did not understand.)
A003: (1) links zu den zwei Dosen circa 30 (at=lengthening) Grad(/a) Drehung (22) (left to the two boxes about 30 degrees turn)

In the appointment scheduling dialogues, often the year is added:

(4) e4012101: what about Monday, the fourth of January? <P> from eight <P> till fourteen-hundred.
s4012102: blurb appointment right blurb mist. [*nonsense*]
e4012102: okay. what about Tuesday, the fifth of January? <P> from<L> <P> eight to fourteen-hundred?
s4012103: please make a proposal.
e4012103: <Smack> <P> okay. <;low voiced> do you have time on Monday, the eleventh of January nineteen-ninety-nine?
s4012201: this date is already occupied.
e4012201: what about Tuesday, the twelfth of January nineteen-ninety-nine?

These preconceptions seem to be very common in HCI and HRI. In [6], I furthermore show that speakers generally believe that robots can be easily disturbed by orthographical matters, that they have problems with basic level and colloquial terminology and metaphorical concepts, and that they have to learn skills in the same order as humans do. Besides these generally shared ideas, users also seem to have very different concepts of their artificial communication partner and the situation, e.g. in the human-robot dialogues:

(5) P075: I was g+ I was wondering, whether it whether it understood English. - (laughter)

(6) S037: scorpion, - turn - ninety - left. (2) turn left (at=prominent)ninety(/a). - - now is that one command or two, - -

(7) A001: good (at=laughter)dog(/a), (1) now pee on 'em (laughter) – sit, (laughter) –

(8) A004: go on, - you are doing fine,

Such utterances indicate two fundamentally different attitudes towards robots, one in which the robot is treated as a mechanical device that needs commands and which is not expected to understand natural language, and the other in which the robot is expected to function like an animal or needs positive encouragement. Similar differences can be found in the distance-measurement corpus:

(9) usr1-2: wie weit entfernt ist die rechte Tasse? (*how far away is the right cup?*)
sys:ERROR
usr1-3: Tasse (*cup*)
sys:ERROR 652-a: input is invalid.
usr1-3: die rechte (*the right one*)

(10) usr3-3: wie heißt du eigentlich (*what's your name, by the way*)

(11) usr4-25: Bist du für eine weitere Aufgabe bereit? (*are you ready for another task?*)

Examples from the appointment scheduling corpus are the following:

(12) e0045206: können Sie denn Ihre Mittagspause auch erst um vierzehn Uhr machen? (*could you take your lunch break as late as 2pm?*)

(13) e0387103: Sprachsysteme sind dumm. (*language systems are stupid*)

An important observation is that these different attitudes towards the computer or robot correspond to different ways of opening the dialogue with the artificial communication partner. These different dialogue openings reveal different preconceptions about what the human-computer or human-robot situation consists in. For example, one such first move is to ignore the contact function of the system's first utterance completely and to start with the task-oriented dialogue immediately:

(14) S: ja, guten Tag, wie geht es Ihnen? (*yes, hello, how do you do?*)
e0440001: ich möchte gerne einen Termin einen Arzttermin mit Ihnen absprechen. (*I want to schedule an appointment a doctor's appointment with you.*)

This group of speakers only minimally reacts to the interpersonal information provided by the system or even refuse communication at that level. Instead they treat the computer as a tool, at best, in any case not as a social actor. I refer to this group as the *non-players*.

In contrast, the players will take up the system's cues and pretend to have a normal conversation. I call these speakers *players* because the delivery of the respective utterances show very well that the speakers find them unusual themselves, as in the following example where the user breathes and pauses before asking back:

(15) S: ja, guten Tag, wie geht es Ihnen? (*hello, how do you do?*)
e0110001: guten Tag. danke, gut. <B> <P> und wie geht's Ihnen? (*hello, thanks, fine. <B> <P> and how do you do?*)

Thus, it is not the case that these users would mindlessly [18, 17] transfer social behaviours to the human-computer situation. For them, it is a game, and eventually it is the game system designers are aiming at. Thus, these users talk to computers *as if* they were human beings.

Also in the human-robot dialogues with written input in which the user has the first turn, the same distinction can be found:

(16) usr17-1: hallo roboter (*hello robot*)
sys:ERROR
usr17-2: hallo roboter (*hello robot*)
sys:ERROR
usr17-3: Die Aufgabe ist, den Abstand zu zwei Tassen zu messen. (*The task is to measure the distance between two cups.*)

In this example, the speaker proposes a greeting himself and even repeats it. Then, he provides the system with an overview of the task. In contrast, user 19 in the following example first types in the help command, which is current practice with unix tools; when he does not get a response, he starts with a low-level, task-oriented utterance without further elaboration or relation-establishing efforts:

(17) usr19-1: hilfe (*help*)
sys:ERROR
usr19-2: messe abstand zwischen zweitem becher von links und zweitem becher von rechts (*measure distance between second mug from left and second mug from right*)

The same two prototypes can be found in our human-robot dialogues in which Aibo uses the same initial utterance as in the appointment scheduling corpus:

(18) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A011: (1) äh, geradeaus gehen. (breathing) – (*uh, going straight*)
R: Welches Objekt soll ich ansteuern? (*which object should I aim at?*)
A011: (1) links. (7) (*left*)

In this example, the speaker immediately produces a very basic spatial instruction. The next utterance is not syntactically or semantically aligned with the robot's question. In contrast, in the next example, the speaker asks the robot back politely. Her next utterance takes up both the term and the syntactic construction of the robot's utterance, and thus her utterance can be understood as the second part of an adjacency pair:

(19) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A014: Mir geht es sehr gut und selbst? (laughter) (1) (*I'm fine and how about you?*)
R: Welches Objekt soll ich ansteuern? (*which object should I aim at?*)
A014: (2) das Objekt äh hinten links. (6) (*the object uh at the back left.*)

Further examples of dialogue beginnings illustrate the spectrum of possible dialogue openings. Thus, the two behaviours identified, the task-oriented response (by the non-players) and the polite complementary question about the system's well-being (by the players) constitute prototypes, which are located at the opposite poles of the same dimension of social relationship:

(20) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A009: (laughter) - guten Tag, - ähm, vorwärts, (2) losgehen? (1) (*hello, um, straight, start?*)

(21) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A022: (1)(at=quiet)gut?(/a) (1) (laughter) (1) (*fine?*)
R: Welches Objekt soll ich ansteuern? (*which object should I aim at?*)
A022: (1) äh vorne links? (4) stopp, - links, (*uh front left? stop, left,*)
R: Soll ich mich zu dem Objekt begeben, das vorne liegt? (*do you want me to move to the object which is in front?*)
A022: (2) nein, - weiter links, (2) (*no, - further left,*)

(22) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A012: (1) gut, danke, (2) (*fine, thanks*)
R: Welches Objekt soll ich ansteuern? (*which object should I aim at?*)
A012: (1) die Schale, - ganz links. (6) (*the bowl, very far left.*)

(23) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A025: (at=prominent)ja,(/a) (hnoise) ganz gut. (at=quiet) und du? - äh(/a) - so, getz, (*yes, quite fine. and how about you? - uh - so, now,*)
R: Welches Objekt soll ich ansteuern? (*which object should I aim at?*)
A025: (1) ähm dieses Müslischälchen was da ganz links steht. - da sollst du hingehen. (*um this muslibowl which is very much to your left - there you have to go to.*)

In general, then, irrespective of particular communication situations between humans and artificial communication partners, we can distinguish two different prototypes of preconceptions: the computer as a tool versus the computer as a social actor. These prototypes are easily classifiable with automatic means since they correlate with a set up surface cues [8].

4 Effects of the Users' Preconceptions

Now that we have established the prototypical preconceptions in human-computer and human-robot interaction, the question is whether and how these preconceptions influence the way users talk to their artificial communication partners.

4.1 The Predictability of Linguistic Features from Preconceptions

For the appointment scheduling dialogues, it was found that the occurrence of conversational and prosodic peculiarities is significantly related to the users' preconceptions as evident from the different dialogue openings [6]. That is, there are significant correlations between dialogue beginning and the use of linguistic strategies on the conversational as well as the prosodic level. The conversational peculiarities comprise reformulations, meta-linguistic statements, new proposals without any relevant relationship to the previous utterances, thematic breaks, rejections, repetitions, and evaluations. In contrast to, for instance, sociolinguistic variables, such as gender, the distinction between players and non-players has a consistent effect on the use of the above conversational strategies. Similarly, the occurrence of phonetic and prosodic peculiarities, in particular, hyper-articulation, syllable lengthening (e.g. Mon<L>day), pauses (between words and syllables, e.g. on <P> Thurs <P>day), stress variation, variation of loudness, and the variation of intonation contours, can be predicted by the dialogue beginnings [6].

Also in the distance-measurement corpus, the dialogue openings can be used to predict the linguistic strategies used. In this case, we have found a systematic relationship with the occurrence of clarification questions [7]. That is, whether speakers began dialogues with a greeting or some other kind of contact-establishing move, as in the following example, or whether they started the task immediately could be used to predict the occurrence of clarification questions, in particular questions concerning the recipient design, such as the robot's perception, functionality and linguistic capabilities, for instance:

(24) usr11-1: hallo# (*hello#*)
sys:ERROR
usr11-2: siehst du was (*do you see anything*)
sys:ERROR
usr11-3: was siehst du (*what do you see*)

Also for the three German human-robot corpora with Aibo, Scorpion and Pioneer, results show a very significant effect between dialogue opening and emotional expression, sentence mood, structuring cues, and reference to the robot. Emotional expression was coded by distinguishing interjections, e.g. *oh, ah*, contact signals, e.g. *hello*, and displays of relationship,

e.g.*my friend*. Regarding structuring cues, we distinguish implicit, such as *now*, from explicit cues, e.g. *the first task*. For sentence mood, particularly relevant are imperative vs. infinitive vs. declarative mood vs. no verb at all. Finally, we coded whether speakers talked about the robot at all, and if so whether they referred to the robot as *he* or as *it*. For these linguistic features, significant differences could be found, depending on the first utterance in the dialogue [5].

To sum up, in the emotional HCI-corpus, the users' concept of the communication partner as tool or as a conversation partner was significantly related to the prosodic properties of their utterances and to the conversational strategies taken. In the distance-measurement corpus, the number of clarification questions asked can be significantly related to different dialogue openings as well [7]. In the three comparable HRI-dialogues, the effect of the conceptualisation of the robot as a tool or as a conversation partner (as apparent from the dialogue opening) was found to be the most important predicting factor for features as diverse as emotional expression, the sentence mood chosen, the kind of structuring cues used and the way the robot is referred to.

We can conclude that the preconception of the artificial communication partner as a tool versus as a social actor plays an important role in predicting the linguistic features empoyed by the users.

4.2 The Constraining Effect of Preconceptions for Alignment and Shaping

We may now want to ask how stable these preconceptions are and whether they may influence the course of the dialogue and particularly the users' alignment behaviour. This is not only theoretically interesting, but also of great practical interst since due to the restrictedness of current systems, it may be very useful to be able to subtly guide users into using those linguistic structures and strategies that the system can process best [28]. As predicted by [20], as well as Branigan and Pearson, this volume, speakers may align with the robot's output. In particular, we find:

Lexical Alignment

(25) R: Was kann ich für Sie tun? (*what can I do for you?*)
A004: (1) geh zur linken Tasse (...) links nach links (1) (*go to the left cup (...) left to the left*)
R: Soll ich mich zu dem Objekt begeben, das vorne liegt? (*do you want me to move to the object which is in front?*)

A004: – nein ähm zu dem Objekt was links liegt (*no um to the object which is to your left*)

In the example, user A004 initially employs a basic level term to describe the object, namely *cup*. After the robot uses the more abstract term *object*, the user aligns with that term. She furthermore aligns with the syntactic constrction the robot employs, namely the relative clause. In the following example, the speaker also aligns with the robot's construction by expanding it in the reply:

Constructional Alignment

(26) R: Welches Objekt soll ich ansteuern? (*which object should I aim at?*)
A003: (2) (at=breathing)hm, (/a) (3) (...) (at=quiet)ähm, (/a) – ja (2) das (3) zweite. – (*um, well the second one.*)

In the example below, the user employs the extrinsic reference system that the robot had introduced turns before:

Alignment of Reference System

(27) A003: (2) zu der Tasse, nord-östlich. (2) (*to the cup, north-east*)
R: Soll ich mich zu dem Objekt begeben, das vorne liegt? (*do you want me to go to the object that is in front?*)
A003: (4) nord-west. (laughter) mein Fehler. (laughter) (*north-west. my mistake.*)

Alignment of Instructional Strategies

(28) A058: gehe vorwärts. (*go straight.*)
Robot: Soll ich mich zu dem Glas begeben? (*do you want me to go to the jar?*)
A058: nein, geh zu dem Plastikbehälter in der Mitte vor Dir. (*no, go to the plastic container in the middle in front of you.*)

In the previous example, the speaker changes from his previous path-based instructional strategy to a goal-based strategy, aligning with the robot's orientation towards objects. In the follwong example, the user picks up the robot's formal form of address:

Alignment of Form of Address

(29) R: Welches Objekt meinten Sie? (*which object did you mean? (formal form of address)*)

A029: - bitte, fahren Sie erstmal geradeaus. (1) (*please, first drive straight (formal form of address)*)

Thus, speakers may take up the linguistic structures presented to them by the system. Moreover, a comparison between human-Aibo interaction with and without linguistic output shows that after the robot's initial utterance *hello, how do you do* many linguistic differences can be found, some of which can be attributed to alignment, some of which must however be due to changes in the conceptualization of the robot due to the fact that the robot produces verbal output. Thus, it seems that the concept of a language-generating robot is more sophisticated than that of an only language-understanding robot. In [4], I have shown the effect on spatial language; for example, the amount of consistent use of higher level spatial instructions increased from 15.4% to 41.2% in the comparison of the two conditions without and with verbal robot output. Similarly, the number of relative clauses rises significantly as well as the complexity of the syntactic structures used. Thus, the robot's linguistic behaviour contributes to speakers' conceptualization about it.

However, the speakers' preconcepts may also define the limits to this kind of adaptation, as can be seen in the following example:

(30) R: Ja guten Tag, wie geht es Ihnen? (*yes hello, how do you do?*)
A008: (2) geh vorwärts. – (*go straight*)
R: Was kann ich für Sie tun? (*what can I do for you?*)
A008: - gehe vorwärts. (7) (*go straight*)
R: Soll ich mich zu dem Objekt begeben, das vorne liegt? (*do you want me to go to the object that is in front?*)
A008: (1) nein. gehe vorwärts. (10) (*no. go straight.*)

The speaker does not adapt to the robot's utterances from the start. A bit later in the dialogue, the effect persists such that the speaker takes up neither the linguistic constructions nor the object-naming strategy presented by the robot. In the last utterance of the excerpt, he minimally aligns with the first part of the adjacency pair produced by the robot by providing the answer 'the box', but immediately after that he switches back to path-based instructions:

(31) R: Soll ich mich zum Glas begeben? (*do you want me to move to the jar?*)
A008: (3) gehe vorwärts. - (*go forward*)

R: Entschuldigung, welches der Objekte wurde von Ihnen benannt? (*excuse me, which of the objects did you name?*)
A008: (1) die Dose. (5) gehe links. (5) gehe links. (2) (*the box. go left. go left.*)

We can thus conclude that alignment, though a natural mechanism in HRI as much as in human-to-human communication, crucially depends on the users' concepts of their communication partner. That is, the less they regard the computer or robot as a social actor, the less they align. This is generally in line with the reasoning in Branigan and Pearson's article (this volume), who also argue that alignment is affected by speakers' prior beliefs. However, they hold users to align with computers only because they consider them to be lingistically limited in their linguistic capabilities, not because they would treat computers as social actors. In contrast, the findings presented here show that users do not constitute a homogeneous group, since speakers' beliefs about their artificial communication partners may vary considerably; those who regard computers as social actors will indeed align with them.

5 General Conclusions

To sum up, the users' concepts of their communication partner turned out to be a powerful factor in the explanation of inter- and intrapersonal variation with respect to linguistic features at all linguistic levels. In particular, two prototypical preconceptions could be identified, one of the artificial communication partner as a tool, one as another social actor. These prototypes can be reliably identified on the basis of the speakers' first utterances which display their orientation towards a social communication or a tool-using situation. These preconceptions have significant correlations with linguistic behaviour on all linguistic levels. Thus, speech directed to artificial communication partners is not constitute a homogeneous variety, and should thus not be referred to as a *register* [14], unless it is captured in terms of microregisters as suggested by Bateman (this volume). Moreover, depending on their attention to social aspects even in the human-computer or human-robot situation, speakers are inclined to align to their artificial communication partners' utterances. Thus, the users' preconceptions constrain the occurrence of, and define the limits for, alignment in human-computer and human-robot interaction.

References

[1] H. H. Clark. *Arenas of Language Use*. Chicago: University of Chicago Press, 1992.

[2] G. de la Flor. User modeling & adaptive user interfaces. Technical Report 1085, Institute for Learning and Research Technology, 2004.

[3] K. Fischer. Notes on analysing context. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Perspectives on Dialogue in the New Millennium*, number 114 in Pragmatics & Beyond New Series, pages 193–214. Amsterdam: John Benjamins, 2003.

[4] K. Fischer. Discourse conditions for spatial perspective taking. In *Proceedings of WoSLaD Workshop on Spatial Language and Dialogue, Delmenhorst, October 2005*, 2005.

[5] K. Fischer. The role of users' concepts of the robot in human-robot spatial instruction. In *Proceedings of 'Spatial Cognition '06'*, 2006.

[6] K. Fischer. *What Computer Talk Is and Isn't: Human-Computer Conversation as Intercultural Communication*. Saarbrücken: AQ, 2006.

[7] K. Fischer and J. A. Bateman. Keeping the initiative: An empirically motivated approach to predicting user-initiated dialogue contributions in HCI. In *Proceedings of the EACL'06, April 2006, Trento, Italy*, 2006.

[8] M. Glockemann. Methoden aus dem Bereich des Information Retrieval bei der Erkennung und Behandlung von Kommunikationsstörungen in der natürlichsprachlichen Mensch-Maschine-Interaktion. Master's thesis, University of Hamburg, 2003.

[9] B. Horton and B. Keysar. When do speakers take into account common ground? *Cognition*, 59:91–117, 1996.

[10] W. Horton and R. Gerrig. Conversational common ground and memory processes in language production. *Discourse Processes*, 40:1–35, 2005.

[11] A. Johnstone, U. Berry, T. Ngyuen, and A. Asper. There was a long pause: Influencing turn-taking behaviour in human-human and human-computer spoken dialogues. *International Journal of Human-Computer Studies*, 41:383–411, 1994.

[12] A. Kobsa. User modeling and user-adapted interaction. In *Proceedings of CHI'94*, 1994.

[13] J. Krause. Fazit und Ausblick: Registermodell versus metaphorischer Gebrauch von Sprache in der Mensch-Computer- Interaktion. In J. Krause and L. Hitzenberger, editors, *Computertalk*, number 12 in Sprache und Computer, pages 157–170. Hildesheim: Olms, 1992.

[14] J. Krause and L. Hitzenberger, editors. *Computer Talk*. Hildesheim: Olms Verlag, 1992.

[15] R. Moratz, K. Fischer, and T. Tenbrink. Cognitive modelling of spatial reference for human-robot interaction. *International Journal on Artificial Intelligence Tools*, 10(4):589–611, 2001.

[16] M.-A. Morel. Computer-human communication. In M. Taylor, F. Neel, and D. Bouhuis, editors, *The Structure of Multimodal Communication*, pages 323–330. Amsterdam: North-Holland Elsevier, 1989.

[17] C. Nass and S. Brave. *Wired for Speech. How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA., London: MIT Press, 2005.

[18] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

[19] J. Pearson, J. Hu, H. Branigan, M. Pickering, and C. Nass. Adaptive language behavior in hci: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, April 2006*, pages 1177–1180, 2006.

[20] M. J. Pickering and S. Garrod. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27:169–225, 2004.

[21] M. Richards and K. Underwood. Talking to machines: How are peoplpe naturally inclined to speak? In *Proceedings of the Ergonomics Society Annual Conference*, 1984.

[22] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

[23] E. A. Schegloff. Notes on a conversational practise: Formulating place. In D. Sudnow, editor, *Studies in Social Interaction*, pages 75–119. New York: Free Press, 1972.

[24] M. F. Schober and S. E. Brennan. Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, and S. R. Goldman, editors, *Handbook of Discourse Porcesses*, pages 123–164. Hillsdale: Lawrence Erlbaum, 2003.

[25] D. Spenneberg and F. Kirchner. Scorpion: A biomimetic walking robot. *Robotik*, 1679:677–682, 2002.

[26] W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin etc.: Springer, 2000.

[27] M. Zoeppritz. Computer talk? Technical Report TN 85.05, IBM Heidelberg Scientific Center, 1985.

[28] E. Zoltan-Ford. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34:527–647, 1991.

On Changing Mental Models of a Wheelchair Robot

Elena Andonova
SFB/TR8 Spatial Cognition, University of Bremen
andonova@uni-bremen.de

1 Introduction

Human-robot interaction has emerged as a field of investigation in its own right in which the more basic questions relating to how people converse with robots have been explored with a view to designing and improving specific applications. Given the combination of theoretical and applied concerns, it is not surprising that the field has been evolving rapidly in an attempt to go beyond the mere description of interactions and into investigation of how people can be influenced to conduct those in particular and predictable ways. One line of research that is currently pursued explores the phenomena of speaker adaptation, or influencing users into adapting to the robotic dialogue system, as well as vice versa. While a number of interactive phenomena are well-established by now, e.g., lexical overlap across speakers, referring expressions becoming shorter and more similar over time, the exact sources of these effects are still being debated. Thus, the key tenet of the theory of interactive alignment [4] is that alignment occurs primarily via an automatic psychological priming mechanism. On this view, mental models are not involved much in this process as they are costly to update and unnecessary in the default case. However, the assumption that mental models are strategically maintained and consciously accessed during interlocutors' interactions may be undermined by the lack of clear empirical evidence that mental models exact a cognitive cost during interaction, on the one hand, and by studies of speech accommodation as a form of adaptive behaviour. The jury is still out on the issue of the automatic vs. strategy-based character of accommodative verbal behaviour; for example, studies have suggested that the actual focus of accommodation may not be the addressees' communicative style in the specific interaction but a rather stereotypical model of the interlocutor which would be important in both convergent and divergent acts [2]. From this perspective, the degree to which interactive alignment

– as a form of adaptive behaviour – is mediated by speakers' mental models remains an open question and is part of a long-term research agenda in human-robot interaction where variability in users' speech patterns, including degrees and forms of alignment, can be examined with respect to their mental models of interlocutors.

While most research has focused on interactive alignment in human-to-human dialogue, recently the relationship between alignment and mental models was explored in the domain of human-computer interaction in an experimentally controlled setting where participants were shown to display greater alignment with a computer program when they were led to believe that their conversational partner was a computer rather than a human being, and further, when they thought that their computer interlocutor had rather basic capabilities instead of advanced ones [1]. Clearly, speakers' mental models provide at least a partial source of variability in aligning one's speech with a computer agent.

These considerations bring to the fore the need for systematic examination of speakers' mental models and their co-relationship with features of dialogic speech, an area that has remained under-researched. In Bremen, a long-term agenda on human-robot interaction has developed around a small set of highly specific spatially-embedded interactional scenarios such as route instructions or internal map augmentation. Within this programme, speakers' mental models have been inferred from the specific features, choices, and constraints attested during their interaction with robots (e.g., Aibo, the robotic wheelchair Rolland, the non-axial robotic Box, etc.). As part of this programme, the study described here aimed at examining mental models by means of explicit assessment of their features. Mental models refer to people' s conceptual frameworks which support their reasoning about the world, about other people, and in the case of human-robot interaction (HRI), about robots as well. Users' mental models can be manipulated by implicit means (variations in the appearance, voice, speech, other capacities, etc. of the robot), or more overtly, by explicit instructions preceding or accompanying the HRI situation, e.g., by providing a name for the robot (female, male), origin (Hong Kong vs. New York), definition of capacities, etc. Mental models of robots have recently been investigated more directly via users' behavioural responses to targeted assessment tools. A series of studies, Kiesler and Goetz [3] have contributed to developing a methodology of measurement and increasing our understanding of the involvement of mental models in human-robot interaction. [5] have also conducted an assessment of robotic mental models based on the Big Five personality model.

In this study, we focus on the relationship between mental models and

users' experience of an HRI situation. The specific situation involved interaction with the Bremen robotic wheelchair called Rolland which could, allegedly, understand and produce speech. The mental models' assessment took place twice – before and after the HRI task – so that both the initial pre-conceptions of robots and the impact of the interaction on the participants' perception of a specific robot could be examined. Our first research question concerned the contents of users' mental models of robots. We also aimed at establishing the relative stability or flexibility of participants' mental models as a function of the specific human-robot interaction that they were involved in – how do mental models of robots compare before and after the interaction with a talking wheelchair robot? Finally, the relationship between participants' general assessment of the HRI and the mental model features was examined.

2 Method

The study used a before-after questionnaire procedure where participants were asked to provide their judgments on how accurately each of a number of features describes what they think of robots in general before the human-robot interaction and of the specific robot after the HRI session on a five-point Likert scale where a score of 5 was associated with 'highly accurately' and a score 'highly inaccurately.'

The interaction was conducted in a Wizard-of-Oz setup in a sequence of spatially-embedded scenarios involving participants describing a room, a corridor environment, and offering Rolland route directions to locations in that same corridor area. This was done while each participant was seated in and navigated manually the wheelchair. Pre-designed and pre-synthesized male-voice robotic utterances were heard by participants as originating with Rolland.

The participants in the experiment were 11 English native speakers (7 women, 4 men, average age 36.5, age range 20-60), 11 German native speakers (8 women, 3 men, average age 23.4, age range 19-40), and 9 German-English bilinguals (8 women, 1 man, average age 21.7, age range 20-25). The bilinguals were asked to use their second language (English) in their communication with Rolland, and the others used their native language.

The mental model measures were partially based on the Big Five inventory (the most widely accepted taxonomy of personality traits since the late 1980s) with its five scales of extraversion, agreeableness, conscientiousness, emotional stability, and creativity/openness to new experiences, and

partially designed specifically for the scenarios of HRI, including scales of sociability, intelligence, partnership, mechanistic vs. anthropomorphic models. Participants also rated the robot' s accuracy and logic. There was also an additional section only following the HRI session which covered more general questions on mutual liking, degree of difficulty of the task and the interaction, degree of stress, enjoyment, satisfaction, interest involved and readiness to participate again.

3 Results

The analysis of participants' judgments in this study shows that they were perfectly able to distinguish among the five scales of personality in their mental representations of robots. In both before- and after-session questionnaires, participants gave high ratings for conscientiousness (4.03 and 4.23, respectively) and emotional stability (3.87/3.74) of robots in general and of Rolland in particular. At the same time, they had rather low expectations of robots on the openness/creativity (1.89/1.98) and agreeableness (1.97/2.50) scales. High estimates of robots' accuracy (3.47/3.90) and logic (4.42/4.06) were accompanied by low values on the anthropomorphism scale (1.58/1.94). Thus, the fact that participants differentiated among the five personality scales and did not provide similarly bland and non-committing estimates of robotic traits indicates that they entered and left the HRI situation with a mental model of robots and of Rolland. This also applies to their estimates of the additional measures on sociability, logic, accuracy, etc. Establishing participants' ability to differentiate among the five personality scales and the additional measures is in line with previous research (Goetz and Kiesler, 2002, Kiesler and Goetz, 2002). The before-after procedure, however, also allowed us to go one step further and assess the dynamics of these mental models and to what extent they were influenced by the specific HRI interaction that participants were involved in. Their initial expectations could thus be teased apart from the effect of the HRI experience. The analyses revealed that participants entered their interaction with Rolland with mental models of robots that already at that point indicated differentiation among the five personality scales, including high expectations of robots' conscientiousness (see above), emotional stability or lack of emotional instability (see above), and of their logic (M=4.42). On the contrary, robotic openness/creativity (see above), agreeableness (see above) and the additional measure of human-likeness or anthropomorphism (M=1.58) were estimated rather conservatively.

A comparison of the ratings given by participants before the HRI with the score following the interaction provides an insight into how stable or unstable such estimates and mental models of robots are. The analyses revealed a set of stable features which remain unchanged as a result of the HRI, namely, the highly positive estimates of emotional stability (or lack of instability) and conscientiousness, as well as the relatively positive values for accuracy and logic, on the one hand, and the negative perception of robots on the openness/creativity and extraversion scales, on the other hand. Stable values show consistency across time of assessment and the modest degree of impact produced by the specific HRI. Obviously, these are deeply entrenched beliefs about robots which are shared at least by the participants in the study. They may also be shared by the designers of robotic interactants, of their verbal output, embedded in the design of the HRI scenarios as such. However, these beliefs appear to be shared on an even wider scale by the society and culture at large. After all, cultural artefacts involving robots, past experience with robotic applications, etc., have taught us that expected and desirable robotic features include mostly accuracy, logic, conscientiousness, and not behaviour which is errorful, random, emotional or humorous, as in our own everyday system of beliefs on intelligent agents, a higher value is placed on utility. Naturally, on the basis of this study alone, we cannot say if people' s mental models vary across interactions with different robots. We expect, however, to find a subset of stable features in these models in addition to features that are more malleable by the particular circumstances of the HRI, the robotic appearance, etc.

In this study, the most fluid features of the mental models were those on the agreeableness scale and the measure of anthropomorphism (how machine-like vs. human-like the robotic wheelchair was perceived to be) on both of which more positive evaluations were received after the interaction than before. In fact, there was no re-arrangement at the bottom of the evaluation hierarchy – the measures with initial low estimates continued to occupy similar bottom ranks in the same hierarchical order as before. The 'movers and shakers' produced changes at the top ranks of the hierarchy: (a) conscientiousness instead of logic became the most positively perceived robotic feature; (b) emotional stability was rank-demoted at the expense of accuracy and partnership.

All in all, participants' perception of the robotic wheelchair was more favourable after participation in the HRI tasks with Rolland than their expectations of robots prior to the interaction. Out of all 11 measures, only three (extraversion, emotional stability, and logic) suffered a numerical drop in scores as a result of the HRI session (.18, .13, and .35 points, respectively),

all other scales showed an improved opinion of Rolland in comparison with general perceptions of robots. However, some changes were quite dramatic while others seemed somewhat superficial. This was confirmed by a statistical analysis of the significance of these changes performed by means of a series of paired t-tests (used to compare two population means where the observations in one of the two samples can be paired with observations in the other sample as in a before-after procedure), revealing that significant changes in participants' perception of robots before and after the HRI occurred on the measures of agreeableness (mean values of 1.97 and 2.50, respectively; paired t-test, $t = 3.02$, $p = .01$), anthropomorphism (mean values of 1.58 and 1.94, respectively; paired t-test, $t = 2.48$, $p = 0.02$), partnership (mean values of 3.37 and 3.84, respectively; paired t-test, $t = 3.41$, $p < 0.001$) and sociability (mean values of 2.89 and 3.21, respectively; paired t-test, $t = 2.06$, $p = 0.05$), all positive increases. As a whole, however, after the HRI, participants continued to maintain their negative stereotypical notions of robots while at the same time re-arranging the positive attributions in their evaluations.

Having established the contents and changes in participants' mental models of robots, we now turn to the general perception of Rolland, the task, and the overall experience of the HRI as assessed by the short end-of-session survey which included questions on mutual liking (How much did you like the robot? How much did the robot like you?), difficulty of the task and of working with Rolland, stress, enjoyment, interest, satisfaction, and willingness to participate in a similar experimental task later. The responses to these questions were moderately to highly correlated (coefficients ranging from .27 to .72). For example, a positive correlation was established between responses on the questions referring to mutual liking – the more the participants liked the robot, the more they thought the robot liked them, too ($r = .33$). Similarly, stress was associated with the difficulty of the task and how hard it was to work with the robot; enjoyment, satisfaction, interest, willingness to participate again were highly correlated, etc. However, is there a relationship between participants' general perception of the HRI task and of Rolland and their pre-conceived ideas of robotic personality and capabilities as established in the before-HRI assessment? Do their initial expectations of a cold and rational robotic assistant affect how they feel about their experience with human-robot interaction at the end of the experimental session? To answer this question, an analysis of correlations between responses on each of the general survey questions and each of the before-HRI measure ratings was conducted. The results of the analysis revealed that almost all of the survey general responses were moderately correlated with a measure

from the initial assessment (the one exception were responses to the question regarding how hard the task was). They were, however, significant correlations with only three of the measures used in the assessment before the HRI, i.e., the ratings of accuracy, anthropomorphism, and the openness/creativity scale. Note that it is only the latter that belongs to the Big Five personality inventory, that is, how extravert, agreeable, conscientious, and emotionally stable robots were in participants' mental models did not affect their general reactions to the HRI experience. Furthermore, it became evident that positive evaluations at the end of the experimental session were associated with lower ratings on the initial assessment of mental models. Thus, initial estimates of openness/creativity were negatively correlated with the degree to which participants liked our robot (r = -.37), or thought that the robot liked them (r = -.31), as well as the level of fun (r = -.35), interest (r = -.34), and willingness to repeat (r = -.34) that they had (the probability level was set to .05 for all correlations reported in the paper). On the other hand, initial low ratings of robots' accuracy were associated with higher levels of overall satisfaction (r = -.30) and participants liking the robot (r = -.39). Anthropomorphism or human – likeness estimates were negatively correlated with how hard it was to work with Rolland and the general level of stress they had during the HRI task. Perhaps somewhat paradoxically, the worse participants thought of robots' potential for creativity/openness to new experiences, accuracy and human-likeness, the more impressed and satisfied they were with the human-robot experience. Their initial schema of most robotic personality traits (robots seen as highly conscientious and unemotional, rather introverted and disagreeable) was not obviously involved in their general HRI assessment at the end. Whether this pattern can be generalized to account for interactions with further robotic partners and in other scenarios, is an open question that remains to be explored.

4 Conclusion

The conclusions that emerge from the analysis of the data on mental models from this study lead to our understanding of the existence of a stable set of features which remain unchanged as a result of the HRI, namely, highly positive estimates of the emotional stability and conscientiousness, accuracy and logic of robots, and at the same time, negative perceptions of robots' openness/creativity, extraversion, and agreeableness. Generally, robots are perceived as machine-like (not particularly anthropomorphic) both before and after interactions with Rolland in the scenarios used here. The general

image of robots is one of cold rationality, lacking in emotion and flexibility. To reiterate, such stable judgments may be representative of shared and deeply entrenched beliefs about robots not only by the participants here, but more widely, as part of the cultural expectations in our society at large.

In this study, the relative flexibility of some features of robotic mental models was established, namely, agreeableness, anthropomorphism, partnership (cooperation and reliability) and sociability. The significant changes observed were all in the positive direction; Rolland was not rated down after the HRI session in comparison with the initial conception of robots. With very few exceptions (openness/ creativity, accuracy), participants' initial mental models were hardly involved in their general perception of the human-robot interaction. However, the more machine-like they thought robots were to begin with, the higher their satisfaction level rose after the interaction. It remains to be seen if this is a 'novice user' effect with all the surprise and excitement which would wear off with repeated interactions by long-term users.

Finally, the next step on our research agenda would take us to the investigation of the relationship between mental models and dialogic feature patterns, including individual and group variability. This will bring us closer to an understanding whether interactive alignment can be enhanced by manipulating speakers' mental models of robots and whether increased levels and scope of alignment are beneficial for efficiency and success in human-robot interaction beyond the enhancement of dialogue.

References

[1] H. Branigan and J. Pearson. Alignment in human-computer interaction. In K. Fischer, editor, *Workshop on How People Talk to Computers, Robots, and other Artificial Communication Partners, Delmenhorst, April 21-23*, 2006.

[2] H. Giles and N. Coupland. *Language: Contexts and Consequences.* Keynes: Open University Press, 1991.

[3] S. Kiesler and J. Goetz. Mental models of robotic assistants. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2002), Minneapolis, Minnesota*, 2002.

[4] M. J. Pickering and S. Garrod. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27(2):169–190, 2004.

[5] B. Wrede, S. Buschkämper, C. Muhl, and K. Rohlfing. Analysing feedback in human-robot interaction. In K. Fischer, editor, *Workshop on How People Talk to Computers, Robots, and other Artificial Communication Partners, Delmenhorst, April 21-23*, 2006.

Alignment in Human-Computer Interaction

Holly Branigan and Jamie Pearson
University of Edinburgh
holly.branigan;jamie.pearson@ed.ac.uk

Abstract

There is strong evidence that speakers in Human-Human Interaction (HHI) are influenced by their interlocutors, both directly via the linguistic content of their interlocutors utterances (alignment), and indirectly via their beliefs about their interlocutors knowledge state, interests and so on (audience design). We discuss a series of experiments that investigated whether alignment effects also occur in Human-Computer Interaction (HCI). Our results suggest that not only does alignment occur in HCI, it is in many circumstances stronger than in HHI. Differences in alignment in HCI versus HHI appear to arise from differences in speakers' a priori beliefs about the capabilities of their interlocutor, suggesting a strategic component to alignment. Furthermore, speakers do not update their a priori beliefs about computer interlocutors on the basis of feedback, unlike in HHI, where feedback leads to the rapid updating of beliefs about (human) interlocutors.

1 Introduction

In order to understand how people behave in Human-Computer Interaction (HCI), it is often valuable to examine how they behave in Human-Human Interaction (HHI). Understanding HHI can help us to predict and simulate human behaviour in HCI. Perhaps more interestingly, it may be able to help us modify human (user) behaviour in HCI. In this paper we are concerned with how a computers linguistic behaviour, specifically its lexical and syntactic choices, may impact on the lexical and syntactic choices made by a human user who interacts with it. We will begin by considering how a human addressee can influence a speakers choices, and examine how this might map onto HCI, before discussing a number of experiments that directly investigated these issues by comparing human linguistic behaviour in the same task in HCI versus HHI.

2 Audience Design

There is overwhelming evidence that addressees influence speakers' linguistic behaviour both indirectly and directly in HHI. Indirectly, they affect speakers through Audience Design, the process by which speakers design their utterances with their addressee in mind (Bell, 1984). Thus speakers take into account their beliefs about the addressees current state of knowledge, beliefs, abilities etc when they formulate their utterances. For example, Fussell and Krauss (1992) demonstrated that speakers used their a priori assumptions about the social distribution of knowledge (e.g., that people are more likely to know movie stars than industrialists) to alter the way in which they referred to entities. In this experiment, speakers participating in a referential communication task that involved describing people from various domains (e.g., politicians, film stars, business people) produced descriptions that reflected their a priori beliefs about how likely the addressee was to be able to identify the referent, using proper names when they judged a referent to be easily identifiable by their addressee (e.g., Clint Eastwood), but more detailed descriptions when they judged a referent to be less easily identifiable by their addressee (e.g., Ted Turner). Such a priori beliefs can affect the form of speakers utterances, as well as their content. For example, beliefs about the linguistic competence of the addressee may cause the speaker to speak more slowly or use less complex syntax when addressing a young child than when addressing another adult, for example (Ferguson, 1975).

Speakers may also dynamically accommodate their addressees' changing state of knowledge. Haywood, Pickering, and Branigan (2005) reported a study in which pairs of participants took turns directing each other to move an object in an array, such as moving a toy penguin into a cup. They manipulated the array such that it contained potential ambiguities. For example, when two penguins were present, the utterance Put the penguin in the cup... was ambiguous. Speakers were more likely to produce *that's* more often (Put the penguin that's in the cup...), thus removing the ambiguity, when there were two penguins than where was only one penguin. Hence speakers chose syntactic structures that were most easily understood by addressees, by accommodating the addressees' current state of knowledge.

Audience Design can also be based on direct evidence from the addressee (i.e., feedback) about the addressees' state of knowledge. In a referential communication task that involved describing New York City landmarks, Isaacs and Clark (1987) showed that a speakers' a priori assumptions about an addressees' knowledge can be dynamically adjusted as their addressees

level of knowledge becomes apparent. Non-native New Yorkers were more likely to initially use a description based on visual cues, such as building with a tall pointy roof and a spike on top, but become more likely to use the name of the landmark over the course of the dialogue if the addressee gave evidence of being a native New Yorker. By contrast, native New Yorkers were more likely to initially use a name, such as Chrysler building, but become more likely to give extra identifying information over the course of the dialogue if the addressee gave evidence of being a non-native New Yorker.

3 Alignment in HHI

As well as addressees indirectly influencing a speakers linguistic behaviour through the speakers beliefs about an addressee, they may directly influence a speaker through their own linguistic behaviour. Evidence for this comes from demonstrations of alignment, the phenomenon whereby people tend to converge on the same linguistic features as a previous speaker. Alignment effects appear to be robust and highly pervasive in dialogue: Speakers have been found to align at many linguistic levels, including those as diverse as rhetorical structure, speech rate, pronunciation, word choice and syntactic structure (e.g., Giles, Coupland & Coupland, 1991; Schenkein, 1980), as well as at entirely non-linguistic levels, such as bodily movements, where it has been termed the chameleon effect (Chartrand & Bargh, 1999).

One important aspect of alignment is that it can be implicit. It almost always arises without explicit negotiation, and on those occasions where speakers do explicitly negotiate a term to use, they frequently end up aligning on a different expression (Garrod & Anderson, 1987). Furthermore, speakers are usually unaware of aligning with a conversational partner. Post-experimental debriefing has shown that speakers are very rarely aware of alignment of form; they sometimes though more frequently do not report awareness of alignment at levels related to meaning.

Alignment occurs at levels of structure concerned with meaning, such as choice of reference frame (Watson, Pickering, & Branigan, 2005) and situation models (Garrod and Anderson, 1987). Similarly, speakers align their lexical choices, using the same words in same ways (e.g., using *square* to refer to a single node or a configuration of nodes; Garrod and Anderson, 1987). In at least some circumstances, such alignment can occur even for lexical choices that are rare or unusual. Bortfield and Brennan (1997) showed that native speakers adjusted their preferred terminology to match non-native interlocutors' non-standard terminology (e.g., The chair that can go back

and forth to refer to a rocking chair) if the non-natives exhibited evidence of comprehension difficulties, although there was no difference in the degree of alignment to a non-native than a native partner.

Such alignment may be linked to differences in meaning. For example, aligning on a term such as rainbow trout versus coloured fish may reflect alignment of interlocutors perspectives, or ways of thinking about the world. But other alignment seems to be unrelated to convergence on types of meaning, such as alignment of speech rate, or alignment of syntax when both alternatives express the same meaning. Branigan, Pickering and Cleland (2000) showed that speakers align syntactic structure. A naive participant and a confederate (who followed a script) took turns to describe pictures to each other. Experimental pictures depicted ditransitive events and could be described using a Prepositional Object (PO) (e.g., The pirate handing the cake to the sailor), or a Double Object (DO) form (e.g., The pirate handing the sailor the cake). Nave participants tended to produce target descriptions that had the same syntactic structure as the confederates preceding prime description, even when the prime and target pictures involved unrelated events, though effects were larger when the same verb was repeated (77% aligned descriptions, versus 63% aligned descriptions when the verb was not repeated). Similar effects have been found for other structures (e.g., NP structure; Cleland & Pickering, 2003), in multi-party dialogues (Branigan, Pickering, McClean & Cleland, in press), and in special populations such as bilinguals, L2 learners, children etc. (Flett, Branigan & Pickering, submitted; Hartsuiker, Pickering & Veltkamp, 2004; Huttenlocher, Vasilyeva & Shimpi, 2004).

Alignment can co-occur alongside audience design. Haywood et al. (2005) found that not only did participants show audience design effects in their production of ambiguous versus disambiguated structures, they also showed alignment effects: participants were more likely to produce disambiguated instructions like Put the penguin that's in the cup after hearing the confederate produce an instruction like Put the sheep that's on the plate, independently of the content of the array.

Alignment effects have been explained in many ways. Some such effects may have a more or less consciously affective element; speakers who converge with respect to breadth of vocabulary are judged more favorably than those who do not, for example (Bradac, Mulac, & House, 1988). There is a substantial body of research that investigates alignment effects (termed accommodation effects) within such a social psychological framework (e.g., Giles, Coupland, & Coupland, 1991; Giles & Powesland, 1975; Giles & Smith, 1979). For example, reciprocity effects may explain why speakers align lin-

guistic form in the absence of differences in meaning (Gouldner, 1960). In such accounts, the perceived social identity of the addressee is critical. For example, alignment in order to display politeness towards an addressee is only relevant for addressees that are perceived as social agents.

Other research explains alignment as a manifestation of audience design. In such accounts, alignment is a strategic behaviour in which speakers choose to adopt the other persons perspective in order to enhance communication: by choosing the same description schema or referential expression as their conversational partner, the speaker maximises the chances of effective communication (e.g., Brennan & Clark, 1996). Such approaches provide a plausible explanation for alignment of aspects of language associated with differences in meaning (e.g., lexical choice), but do not adequately explain why alignment of linguistic form occurs (in the absence of meaning differences).

A third approach explains the effects primarily with reference to the cognitive processes that are involved in language processing. For example, Pickering and Garrod (2004) suggested that alignment is an automatic, default behaviour. In support of this proposal, they noted that children show a stronger tendency to align than adults; notably, they align linguistic form even when this leads to misunderstanding, such as using the same term with different reference (e.g., using *square* to mean different things; Garrod & Clark, 1994). Garrod and Clark therefore suggested that children align as their default behaviour, and that part of becoming a mature language user involves learning to suppress the tendency towards alignment when necessary. In keeping with this, Pickering and Garrod (2004) suggested that alignment is based on automatic priming mechanisms. That is, alignment reflects the facilitation of particular linguistic representations and processes following their prior use. For example, lexical alignment may reflect basic priming processes of the sorts that have long been identified in models of language processing. Similarly, syntactic alignment is hypothesised to occur because prior production or comprehension of a particular syntactic structure raises the activation of the relevant syntactic representations and/or processes, making them a better candidate for subsequent use (Branigan, Pickering & Cleland, 2000).

Pickering and Garrod argued that alignment is fundamental to efficient communication. In their account, efficient communication arises when interlocutors come to have the same understanding of relevant aspects of the world, through alignment of their situation models (e.g., Zwaan & Radvansky, 1998). Such alignment itself arises from alignment of other aspects of language (e.g., syntax, lexical choice): alignment is hypothesised to perco-

late upwards, such that alignment at one level promotes alignment at others. Hence lexical alignment promotes syntactic alignment, which in turn promotes semantic alignment.

Of course, these different types of explanation are not mutually exclusive. Rather, there is good reason to believe that multiple factors underline alignment. It seems most likely that there is at least some implicit element, given that participants generally report lack of awareness of alignment. But other factors may also contribute to the overall effect, such that the basic (automatic) alignment effect may be enhanced by other, social factors, such as the social status of an interlocutor. Such factors may influence alignment at some levels of structure more than at others. For example, in the same way that levels of structure associated with differences in meaning appear to be more amenable to audience design effect, such levels might also be more amenable to non-implicit or strategic alignment effects. Hence we suggest that observable alignment of linguistic behaviour, by which we mean convergence on common linguistic features, is most likely to contain both automatic and strategic components.

4 Possible Patterns of Alignment in HCI

All of the evidence reviewed above relates to alignment in HHI. But if alignment is a default linguistic behaviour whose occurrence may at least in part arise as a consequence of the architecture of human language processor, then it should occur in any communicative context. Hence we might expect to find alignment effects in HCI.

If alignment effects arise purely from automatic priming of linguistic representations, then alignment would occur whenever a linguistic structure is encountered, irrespective of context. However, there are reasons to expect that the pattern of any alignment in HCI might differ from that found in HHI. In particular, it seems likely that there may be a strategic component to alignment that would affect alignment differentially in HHI and HCI. Some element of this may relate to social factors such as community membership. In that case, speakers might be influenced by their a priori beliefs about the social identity of the computer. If systems are not treated as social agents just like humans, then alignment in HCI might differ from alignment in HHI; for example, in that case we might expect less alignment in HCI contexts if a substantial component of alignment relates to social factors such as reciprocity and politeness. Conversely, if systems are treated as social agents just like humans (Reeves & Nass, 1996), then alignment with

a computer could occur in the same way as it does with a human.

But as we have seen, speakers' linguistic choices in HHI, including their lexical and syntactic choices, are also influenced by both their a priori beliefs and the direct evidence that they encounter concerning, their addressees' knowledge, capability etc. Extrapolating from this, it seems plausible that peoples beliefs about the knowledge, capability etc. of a computer might influence the extent to which they align with it. For example, people might assume computers to be (generally and/or specifically linguistically) less capable than humans. This might increase their likelihood of aligning with computers for essentially strategic reasons (i.e., to increase the likelihood of successful communication), relative to their likelihood of aligning with another human, to the extent that people might overcome their default preferences to use particular terms or structures in order to align with a less preferred one that has just been used by a computer interlocutor. If there is such a strategic component to alignment, then we might find variations in magnitude of alignment associated with variations in the perceived capability of the computer, such that alignment is stronger with a computer that is perceived to be of lower capability than with one perceived to be of higher capability.

Research on HHI has shown that speakers can rapidly update their a priori beliefs on the basis of feedback from the addressee concerning communicative success (or lack thereof), so we might expect that a priori beliefs about the capability or otherwise of a computer might similarly be quickly overridden in the light of feedback. Hence we might expect an initial tendency towards stronger alignment in HCI to rapidly disappear if the computer gives evidence of successful comprehension.

In sum, then, alignment is potentially a highly important phenomenon in HCI but there are many factors that might affect patterns of behaviour. Specifically, there are many reasons why alignment in HCI might differ from alignment in HHI. One important issue that any study of such effects must address is the extent to which any differences between HCI and HHI are an artefact of the communicative situation, in other words, the involvement of a computer in the communication rather than arising from genuine differences between HCI and HHI.

5 Experimental Investigations of Alignment in HHI and HCI

Our research investigates lexical and syntactic alignment in HCI in a way that excludes such an explanation by using a modified version of the confederate scripting paradigm (Branigan, Pickering & Cleland, 2000), which allows investigation of alignment in dialogue under controlled conditions. Pairs of participants play a picture-matching and -describing game, alternately describing a picture to their interlocutor, and selecting a picture that matches their interlocutors description. In fact, only one participant is an experimental participant; unbeknownst to the naive participant, the other participant is a confederate of the experimenter who produces descriptions scripted by the experimenter. The form of the confederates' description is systematically manipulated and the form of the participants subsequent description is examined to see whether it has the same linguistic features (i.e., aligns) or not with the confederates immediately prior description. In experiments investigating syntactic alignment, we were concerned with whether the participant chose the same syntactic structure as the confederate had just used, when they had a choice of two denotationally identical alternatives (PO vs DO) to describe a ditransitive event; in experiments investigating lexical alignment, we were concerned with whether the participant chose the same word as the confederate had just used, when they had a choice of (at least) two quasi-synonymous words to describe a single object.

In our version of the confederate scripting paradigm, participants were led to believe that they were playing the picture-matching and -describing game with their interlocutor via a networked computer terminal, interacting with their unseen interlocutor by typing. We manipulated participants' beliefs about identity of their interlocutor: participants were led to believe that they were interacting with a computer interlocutor or with a human one. In fact, there was no interlocutor: participants always interacted with a computer program that produced pre-scripted utterances (Reverse Wizard-of-Oz). Using this methodology enables the experimenter to systematically control the interlocutors' utterances that participant encounters. In the studies we report here, the actual linguistic behaviour that they experienced from their interlocutor was always identical in all conditions. In other words, the human and computer interlocutors behave identically. Indeed, all aspects of the experiment were identical apart from the participants' beliefs about the interlocutor with which they were interacting. Clearly, then, any differences in participants' linguistic behaviour must be due to differences

in participants' beliefs about their interlocutor. In this way we can investigate how beliefs about the nature of ones interlocutor affect participants likelihood of aligning to their interlocutor.

In Branigan, Pickering, Pearson, McLean and Nass (2003), we investigated the role of a priori beliefs about an addressee on syntactic alignment. This study was similar to Branigan et al. (2000), but using typed communication. We manipulated the syntactic structure of the description that participants received, ostensibly from their interlocutor: experimental pictures depicting ditransitive events were described using two different syntactic forms, a PO or a DO form. We examined how this affected the syntactic structure that they produced for the immediately subsequent describing turn. We also manipulated whether these two descriptions involved the same verb or different verbs. In addition, we also manipulated participants beliefs about the nature of their interlocutor: Participants interacted with what they believed to be another person or a computer.

Given that Branigan et al. (2000) and other researchers have found a strong tendency in HHI for speakers to use the same structure as the utterance they had just heard, which increased when the verb was repeated between descriptions, what predictions might one make for syntactic alignment in HCI? Alignment at the level of syntactic form seems to occur without any awareness on the part of speakers (see Pickering & Branigan, 1999 for a review). Branigan et al. (2000) interpreted their results in terms of the activation of syntactic information: Comprehending a particular structure activates associated syntactic rules and thus raises the likelihood of their application in subsequent speech. If syntactic alignment is a largely automatic process, then we would expect it to be relatively impervious to beliefs about an interlocutor. That is, an utterance with particular syntactic characteristics should bring about the same effect on the addressee, regardless of the identity of the producer. For example, comprehending a PO sentence will automatically activate the syntactic rule(s) associated with the PO structure. However, we noted above that alignment in HCI might be subject to social factors (e.g., reciprocity, politeness) or to strategic effects related to differences in a priori beliefs about computers versus humans, either of which could give rise to different patterns of behaviour in HCI versus HHI.

In our study, a participant's description was coded as aligned if it had the same syntactic structure as the structure of their interlocutor's immediately preceding description (either PO or DO), or as misaligned if it had a different syntactic structure. We found that, as in earlier studies of HHI (Branigan et al., 2000), alignment occurred whether the verb in the interlocutors' descriptions and the verb in the participants' subsequent descrip-

tions were the same or different, but it was significantly stronger if the verb was repeated than if it was not. This suggests that alignment processes in typed dialogue involving no other visible interlocutor are broadly similar to alignment processes in dialogue between co-present interlocutors who use speech to communicate.

More interestingly, however, the results helped to distinguish between accounts of alignment in which a priori beliefs about the nature of one's interlocutor are not relevant, such that the magnitude of alignment is based solely on features of the utterances that have just been encountered; and accounts in which alignment is influenced by beliefs about the interlocutor, either because it is a strategy that people use because they believe it is beneficial in helping both interlocutors to reach mutual understanding or because it arises from social factors such as reciprocity and politeness, and which is, to at least some degree, under their control. In the study, participants encountered identical utterances in each condition (HHI vs HCI). When the interlocutor's description and the participant's description involved different verbs, alignment occurred to the same extent for human and computer interlocutors. Hence, participants aligned linguistically with what they believed to be a computer, and the strength of this alignment was broadly comparable with the alignment that occurred when participants believed themselves to be communicating with another person. By contrast, when the interlocutors description and the participants description involved the same verb, there was significantly greater alignment to a computer than to a human interlocutor.

The finding of comparable alignment to both computer and human interlocutors when the verb was not repeated is in line with accounts in which alignment has a non-strategic component, in keeping with accounts stressing that alignment is a basic organizing principle of dialogue (Pickering & Garrod, 2004). It is consistent with Reeves and Nass's (1996) claim that people respond mindlessly to social cues, irrespective of their origin. But the greater alignment to computer than human interlocutors when the verb was repeated provides evidence that when people may be more aware of the nature of their utterances, alignment can also involve strategic activation of a decision component. In this case, the lexical repetition, together with the use of typed responses in which their utterance was visible on-screen, may have made participants more aware of the differences between the PO and DO constructions, allowing for participants to chose to align or not. This suggests that beliefs about one's addressee can affect alignment when speakers are aware that a strategy of alignment is available.

Existing evidence suggests that speakers' lexical choices in HHI are af-

fected by beliefs about one's addressee (e.g., Fussell and Krauss, 1992). Our finding of greater alignment to computer than human addressees when the verb was repeated suggests that beliefs about an addressee affect syntactic alignment in HCI when speakers are aware that a strategy of producing aligned utterances is available. Thus, it seems likely that there may be a strategic component to the formulation of utterances that would affect lexical alignment in HCI. In Branigan, Pickering, Pearson, McLean, Nass and Hu (2004), we investigated lexical alignment using the typed version of the confederate scripting paradigm described above. In this study, participants saw two objects on-screen, and had to name one of them. Experimental objects were chosen to have one highly preferred name (e.g., bench) and one highly dispreferred but acceptable name (e.g., seat), on the basis of the pretest. We manipulated the lexical items that participants received, ostensibly from their interlocutor, so that they received with the highly preferred or the highly dispreferred but acceptable name. We examined the lexical form that participants produced when they subsequently named the same picture. As before, we also manipulated participants' beliefs about the nature of their interlocutor: participants interacted with what they believed to be another person or a computer.

Participants' responses were coded as aligned if they used the same word to name the picture as that just used by their interlocutor, or as misaligned if they used a different word. The results showed that speakers lexically aligned to both computer and human interlocutors. Hence, lexical alignment occurs in HCI just as in HHI. Moreover, participants aligned to a highly dispreferred term, overriding their own lexical preferences. However, there was significantly greater alignment to a computer than to a human interlocutor. This follows the pattern of results found in the repeated-verb condition of our previous study investigating syntactic alignment, and again implies that alignment is influenced by beliefs about one's addressee. It provides further evidence that alignment involves strategic activation of a decision component in contexts where speakers may be more aware of the linguistic characteristics of their utterances or the existence of alternative linguistic formulations for their intended message.

Why might speakers align more with computer interlocutors when they are aware that such a strategy is open to them? Clearly, social factors such as reciprocity and politeness are not a substantial component of alignment in such contexts. If computers are treated as social agents just like humans, then alignment based on reciprocity/politeness should occur in the same way with a computer as it does with a human. If computers are not treated as social agents just like humans, then alignment based on reciprocity/politeness

should occur to a much lesser extent with a computer than with a human. But we found neither such pattern; instead, we found more alignment with a computer than with a human, suggesting that even if such social factors do influence alignment, their influence is a relatively negligible determinant of alignment in these contexts.

We noted above that speakers' linguistic choices in HHI, including their lexical and syntactic choices, can be influenced by their a priori beliefs and the direct evidence that they encounter concerning their addressees knowledge, capability etc. (e.g., Bortfield and Brennan, 1997). Thus, a possible explanation for the greater alignment to computers than human addressees observed in our previous studies may be because people believe that computers are, in some respects, less capable (generally, or specifically linguistically) than people. This might increase their likelihood of aligning with computers for essentially strategic reasons (i.e., to increase the likelihood of successful communication). If there is such a strategic component to alignment, then we might find variations in magnitude of alignment associated with variations in the perceived capability of the computer, such that alignment is stronger with a computer that is perceived to be of lower capability than with one perceived to be of higher capability.

In a further study, we therefore manipulated participants beliefs about the capability of a computer interlocutor. In Pearson, Hu, Branigan, Pickering and Nass (2006), we used the same method as above to further investigate lexical alignment. Unlike in the previous studies, participants were always led to believe that they were interacting with a computer (i.e., there were no human interlocutor HHI conditions). We manipulated participants' beliefs about the capability of the computer. Because the manipulation through verbal instructions to induce different beliefs about an interlocutor generated strong effects, we employed a more subtle manipulation of the apparent sophistication of the computer by using a start-up screen that made the computer system appear old-fashioned and unsophisticated (basic computer condition) or up-to-date and sophisticated (advanced computer condition). The start-up screen for the basic condition displayed the term Basic version, bore a 1987-dated copyright, and displayed a fictional computer magazine review stressing its limited features but cheap price and value for money. In contrast, the start-up screen for the advanced condition displayed the term Advanced version: Professional edition, bore a current-year copyright, and displayed a fictional computer magazine review stressing its expense and its impressive range of features and sophisticated technology.

Participants' responses were coded as aligned if they used the same name to describe an object as their interlocutor had previously used to name the

object, or as misaligned if they used a different name. The results showed that participants lexically aligned to both basic and advanced computer interlocutors, producing the dispreferred name if their interlocutor had used it. However, there was significantly greater alignment when the interlocutor was a basic than advanced computer, even though the interlocutor produced identical behaviour in both conditions and even though the interlocutor gave evidence of understanding the participant's preferred name in both conditions. In other words, when participants were led to believe that a computer was of restricted capabilities, they aligned more than when they were led to believe that it was of greater capabilities, irrespective of the direct evidence they received about its capabilities. Hence participants made reference to their a priori beliefs about an interlocutor's capabilities when choosing how to name an object; they did not update these beliefs in the face of direct evidence that the interlocutor understood the alternative name. These results converge with our previous findings that beliefs about one's interlocutor affects alignment, and provide further evidence that alignment involves strategic activation of a decision component when speakers may be more aware of the existence of alternative ways of encoding the same meaning. This suggests that people believe that computers are, in some respects, less capable than people, and that people strategically align with computers to increase the likelihood of successful communication.

The previous study suggested that beliefs about a computer interlocutor's capability affect the magnitude of alignment in HCI. To examine whether the same is true with respect to beliefs about a human interlocutor's specifically linguistic capability in HHI. To investigate this, we conducted a further study that again manipulated participants beliefs about the capability of their interlocutor. In Pearson, Pickering, Branigan, Hu and Nass (2006), we investigated lexical alignment using a similar method as above, but this time participants always believed that they were interacting with another person. However, they were induced through verbal instructions to have different beliefs about the linguistic capability of their interlocutor. Specifically, participants believed that they were interacting either with a native English-speaking or with a non-native English-speaking interlocutor. (Note that unlike our previous studies, this study employed a within-participants design.)

Participants' responses were coded as aligned if they used the same name as that used prior by their interlocutor to name the picture, or as misaligned if they used a different name. The results showed that speakers lexically aligned to both native and non-native English-speaking interlocutors, and that there was no difference in alignment when the interlocutor was a na-

tive or non-native English-speaker. These results converge with previous findings (e.g., Isaacs and Clark, 1987) showing that a speaker's a priori assumptions about an addressees knowledge can be dynamically adjusted as their addressees level of knowledge becomes apparent: a priori beliefs that a non-native English-speaking interlocutor is linguistically less capable are rapidly updated on the basis of feedback from the interlocutor concerning communicative success. In this case, participants accommodated evidence that the interlocutor understood the preferred term (even if the interlocutor used the dispreferred term in their own descriptions) and continued to use that term in their utterances. This contrasts markedly with our previous finding in HCI that participants align more strongly with a computer that is perceived to be of lower capability than with one perceived to be of higher capability: a priori beliefs that a basic computer interlocutor is less capable were not updated on the basis of feedback from the interlocutor concerning communicative success.

6 Summary and Conclusions

To summarize our findings, we demonstrated alignment affects in HCI as well as HHI: there was a tendency for speakers to align both syntactically and lexically to both computer and human addressees. Hence in both HCI and HHI, the features of an utterance that the speaker has just encountered shape the utterances that the speaker subsequently produces. For example, after reading an utterance with a particular syntactic structure, participants tended to repeat that syntactic structure in a subsequent utterance involving a different verb. In such cases, alignment was the same whether the participants believed themselves to be interacting with a human or a computer. This suggests that in some respects alignment processes in typed dialogue involving no other visible interlocutor are broadly similar to alignment in dialogue between co-present interlocutors who use speech to communicate (e.g., Branigan et al., 2000).

However, and more importantly, we found that a speaker's linguistic behaviour, and specifically the extent to which it is affected by an addressee's linguistic behaviour, is influenced by beliefs about an addressee. In this respect, our results are important in demonstrating that alignment is not an entirely automatic behaviour, but rather a behaviour that may have a strong strategic component in addition to a basic automatic component. In contexts where they are aware of the availability of alternative linguistic realisations of a message, and hence of the availability of alignment as a

strategy, participants may choose to align in order to maximise the chances of successful communication when they believe that communication may otherwise fail. For example, participants aligned lexically and syntactically (for utterances containing the same verb) to a greater extent when they believed they were interacting with a computer than with a human. Our results suggest that computers are not treated as social agents just like humans; rather, people believe that computers are, in some respects, less capable than people. The finding of greater lexical alignment to basic than advanced computer addressees provides further support for this conclusion. Intriguingly, such a priori beliefs appear to be resistant to updating on the basis of behavioural evidence: whereas a priori beliefs about human addressees appear to be rapidly updated based on the addressees contributions throughout the dialogue, speakers do not appear willing to alter their beliefs about computers on the same evidence, suggesting that they may err on the side of caution with respect to designing utterances for computer interlocutors.

Overall, our results suggest that not only does alignment occur in HCI, it may be an even more important determinant of behaviour in HCI than in HHI, because it may involve a stronger strategic component that is designed to increase the likelihood of successful communication. It remains to be seen whether such alignment can be exploited to develop systems that are both robust and naturalistic.

7 References

Bell, A. (1984). Language style as audience design. Language in society, 13, 145-204.

Bortfeld, H. & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. Discourse Processes, 23, 119-147.

Bradac, J.J., Mulac, A., & House, A. (1988). Lexical diversity and magnitude of convergent versus divergent style shifting perceptual and evaluative consequences. Language & Communication, 8, 213-228.

Branigan, H.P., Pickering, M.J., & Cleland, A.A. (2000). Syntactic coordination in dialogue. Cognition, 75, B13-B25.

Branigan, H.P., Pickering, M.J., McLean, J.F., & Cleland, A.A. (in press). Syntactic alignment and participant role in dialogue. Cognition.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., & Nass, C.I. (2003). Syntactic. Alignment Between Computers and People: The Role of

Belief about Mental States. Presented at the 25th Annual Meeting of the Cognitive Science Society, Boston, MAS.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., Nass, C.I., & Hu, J. (2004). Beliefs about mental states in lexical and syntactic alignment: Evidence from Human-Computer dialogs. Poster presented at the 17th annual CUNY Human Sentence Processing Conference, Maryland, DC.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory and Cognition, 22, 1482-1493.

Chartrand, T.L., & Bargh, J.A. (1999). The chameleon effect: The perception-behaviour link and social interaction. Journal of Personality and Social Psychology, 76, 893-910.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun phrase structure. Journal of Memory and Language, 49, 214-230.

Ferguson, C. (1975). Toward a Characterization of English Foreigner Talk, Anthropological Linguistics, 17, 1–14.

Flett, S.J., Branigan, H.P., & Pickering, M.J. (submitted). Syntactic representation and processing in L2 speakers.

Fussell, S. E., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. Journal of Personality and Social Psychology, 62, 378-391.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. Cognition, 27, 181-218.

Garrod, S., & Clark, A. (1994). The development of dialogue co-ordination skills in schoolchildren. Language and Cognitive Processes, 8, 101-126.

Giles, H., Coupland, N. & Coupland, J. (1991) Accommodation theory: Communication, context, and consequence. In: Contexts of accommodation: Developments in applied sociolinguistics, ed. H. Giles, J. Coupland, & N. Coupland, pp. 168. Cambridge University Press.

Giles, H., & Powesland, P. (1975). Speech Style and Social Evaluation. San Diego: Academic Press.

Giles, H., & Smith, P.M. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. St. Clair. (Eds), Language and Social Psychology. Oxford: Blackwell.

Gouldner, A.W. (1960). The norm of reciprocity: A preliminary statement. American Sociological Review, 25, 161-178.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish/English bilinguals. Psychological Science, 15, 409-414.

Haywood, S.L., Pickering, M.J., & Branigan, H.P. (2005). Do speakers avoid ambiguities during dialogue? Psychological Science, 16, 362-366.

Huttenlocher, J., Vasilyeva, M., & Shimpi, P. (2004). Syntactic priming in young children. Journal of Memory and Language, 50, 182-195.

Isaacs, E. A., & Clark, H.H. (1987). References in conversations between experts and novices. Journal of Experimental Psychology: General, 116, 26-37.

Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J., & Nass, C.I. (2006). Adaptive Language Behavior in HCI: How Expectations and Beliefs about a System Affect Users Word Choice. Talk presented at the CHI 2006 conference, Montreal, Canada.

Pearson, J., Pickering, M.J., Branigan, H.P., Hu, J. & Nass, C.I., (2006). Influence of prior beliefs and (lack of) evidence of understanding on lexical alignment. Poster presented at the 12th annual Architectures and Mechanisms of Language Processing Conference, Nijmegen, Netherlands.

Pickering, M.J. and Branigan, H.P. (1999) Syntactic Priming in Language Production, Trends in Cognitive Sciences, 3, 136–141.

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences. 27, 169225.

Schenkein, J. (1980). A taxonomy for repeating action sequences in natural conversation. In B. Butterworth (ed.), Language production, Vol. 1, 21-47. London: Academic Press.

Watson, M. E., Pickering, M. J., & Branigan, H. P. (2006). An empirical investigation into spatial reference frame taxonomy using dialogue. Proceedings of the 26th Annual Conference of the Cognitive Science Society, Vancouver, Canada.

Zwaan, R.A. and Radvansky, G.A. (1998) Situation models in language comprehension and memory. Psychological Bulletin, 123, 162185.

A social-semiotic view of interactive alignment and its computational instantiation: a brief position statement and proposal

John A. Bateman
University of Bremen, Germany
bateman@uni-bremen.de

1 Introduction

Interactive alignment [25] is one of the currently most promising additions that have been made to our theoretical approaches to understanding dialogue. The empirical investigation of alignment in dialogue has made considerable progress in recent years and a broadening range of results is being gathered concerning both the nature of and conditions on alignment. Rather less attention has, however, been given to the possible implications that such results have for appropriate design decisions for dialogue systems capable of supporting alignment. Often alignment models that are proposed make little contact with large-scale computational language resources used for sophisticated dialogue systems such as lexicons, grammars, semantics and so on.

In this position paper, I sketch a proposal for an architecture for the computational modelling of alignment within dialogue systems that can be used as a repository for recording and evaluating empirical results/claims concerning alignment behaviour. The model requires that particular features of a linguistic system be made accessible to alignment mechanisms in order that alignment be enforceable. The precise nature of these features, as well as the determination of the scope of alignment over the course of a dialogue, must be established empirically. Explicitly capturing how speakers interact with artificial communication partners is then one crucial aspect of defining the space of possibilities within which alignment may operate. However, providing the level of detail required for driving such a model still presents significant challenges for empirical investigations. Just what collections of features are 'at risk' during alignment and which are not is

still largely unexplored. And yet, without answers to these questions, it will not be possible to construct naturally aligning dialogue agents. One focus will therefore be on the demands that computational modelling places on empirical investigation: what kind of empirical research is now necessary in order to support more sophisticated dialogue systems?

To start, I set out very briefly an alternative view of the nature of interactive alignment that draws on constructs from a socially-oriented view of language rather than a psychological one. The two approaches do not, in my view, necessarily conflict; the social processes also need to have a grounding in psychological processes and it is to be expected that there will be convergences in the functionalities achieved. The social orientation does, however, add a further set of considerations to the necessity and functionality of a phenomenon like alignment in discourse. In particular, we see from a sketch of how language is considered to function from the social semiotic perspective that it also predicts that alignment *must* take place to some respect—or, at least, that it would be extremely surprising if it did not occur. This follows from what is known about the relation of language use to situation in general and so if it were not also now available as a principle in psycholinguistics it would be necessary to invent it. Given this perspective, I also then sketch how this could find a computational instantiation in a natural language system drawing, again, on formalisable notions of how situation and language use can be related.

2 Language as social semiotic: Register

The position set out in [15] argues that language is essentially a social phenomenon. Language behaviour then unfolds in time and is simultaneously, in its unfolding, a structuring and restructing of the interpersonal situtation. Language is itself viewed as a stratified system (following [18]), with relations of 'meta-redundancy' holding between strata. The higher (more abstract) strata anchor directly into social context and situation; the lower (least abstract) strata are the traditional phonology, lexicogrammar, discourse semantics of linguistics. The model of language use relies crucially on a tight bidirectional relationship holding between contextual configurations and configurations in the semantics and lexicogrammar. That is: particular lexicogrammatical configurations are indicative of particular situational configurations.

This is already sufficent to see that something like alignment is strongly predicted. As shown in Figures 1 and 2, the situation for the individualistic

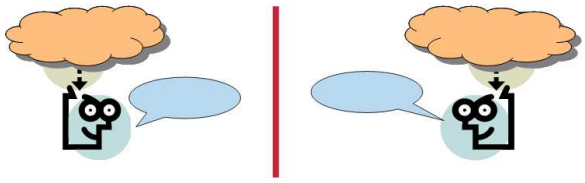

Figure 1: Individual view of linguistic interaction

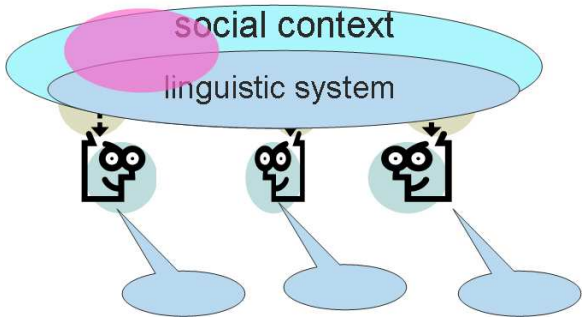

Figure 2: Social view of linguistic interaction

approach presents the mystery of how the two agents come to a common understanding; in contrast, in the social approach, language use necessarily enforces an overall common situatedness of the interlocutors. There can, of course, be variation and differences in the situation that each agent acts in, but this variation takes place against the backdrop of a general commonality rather than *vice versa*.

The linguistic accounts developed within this tradition, primarily but not only within systemic-functional linguistics, rely crucially on the notion of *register*. Register was suggested early on in studies of situated language [26, 27, 14] and has since become a major component of systemic theory [22, 20]. Register is typically divided into three areas of meaning: field, the social activities being played out; tenor, the interpersonal relationships and evaluations being enacted; and mode, the channel and rhetorical purposes of the interaction. Each of these areas is taken to be carried primarily by particular identifiable resources from the semantics and lexicogrammar. This is the explanatory mechanism suggested to explain why particular situational uses of language pattern together with particular selections of linguistic features.

In [2], drawing on data that also fed into forerunners of the interactive

Figure 3: Microregister and register

alignment perspective [12, 13], I extended the notion of register, holding for a situation as a whole, to a derivative notion of **microregister**. The essence of this idea is that there is nothing special about entire situation that differ from individual utterances in discourse. Each individual utterance is linked into a situational context in the traditional manner of register theory but, of necessity, can also change and modify that situational context. Thus, the trajectory of linguistic selections in a discourse is paralleled by trajectories of contextual development.

This draws strongly on Halliday's suggested meterological metaphor in which register corresponds to climate and microregister corresponds to weather. There is no difference in kind between these phenomena—simply one of time depth. The daily reoccurences that we experience as weather add up over time to be characterizeable as a climate. But the climate does not exist independently of the unfolding daily weather. Similarly, register is the contextual configuration holding for an entire 'text'; but this is nothing other than the result of the trajectory followed through and created by the individual contributions to that text. The proportionality at hand is depicted in Figure 3. This also makes the connection to alignment clear: alignment from the psychological perspective corresponds to microregister from the social perspective.

We already know a considerable amount about the general constraints that register, or contextual configurations, exert on language. Established studies of register, such as that of [8] have demonstrated the effectiveness and robustness of the constraint. If we model the language system as networks of possible choice, as is generally done within systemic-functional linguistics, then the consequences of register can be seen as the definition of *subgrammars* where certain choices are preferred and other dispreferred. This is suggested graphically in Figure 4.

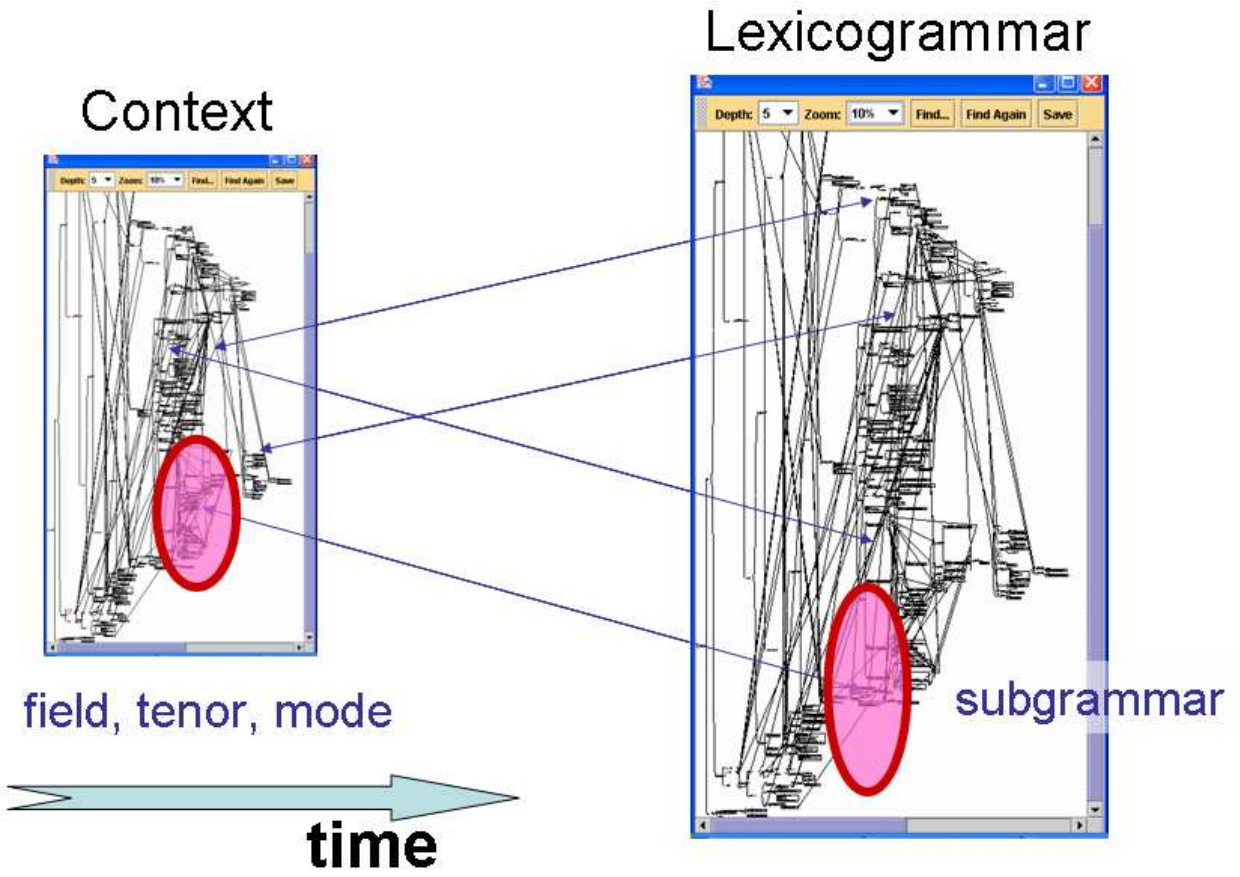

Figure 4: Register and subgrammars over time

3 Computational modelling

We have substantial compuational grammars available in the systemic-functional framework [21, 4]. These are expressed as networks of choices that capture functionally motivated distinctions. Formally, these networks correspond to large type lattices defined over feature structures [17]. Using register as a way of restricting the scale of these networks during actual use for generation or interpretation was first implemented computationally by [24]. This was initially achieved by defining networks of choices for a contextual description and relating features of the lexicogrammar directly to features of the context. A similar approach was also then carried out with the generation grammar of the Penman system by [10]. Although this kind of approach achieves a restriction of the language that occurs according to contexts, it also demands an extremely fine description of context: probably too fine for most purposes since all lexicogrammatical decisions were dependent on their being corresponding contextual decisions to drive them.

A further, more flexible account of the relation between register and semantics and lexicogrammatical expression was developed and implemented by [5, 6]. This approach combined the flexibility of full natural language generation according to semantic inputs and the restriction of register. In [1] we have developed this further and propose that we need 3 distinct mechanisms in a generation system in order to allow register to effectively control phrasing:

1. the selection of which 'size' (more technically, *rank*) of grammatical unit is to be used for given semantic classes;

2. the construction of a subgrammar, which controls the grammatical options available; and

3. a controlled mapping of instances in the world (i.e, concepts in a domain model) to a linguistic ontology which will guide the grammar during generation.

These mechanisms are quite general and are sufficient for providing a very rich and varied range of linguistic phrasing variation that nevertheless remains under functional control.

Systemic-functional grammars are very amenable to defining subgrammars by pruning the type lattice of unwanted or unused features. This is discussed from the perspective of pure engineering efficiency in [3]. Now we can consider using these techniques on a move-by-move basis in a dialogue. The

controlled mapping of instances in the world to linguistic ontology has also been explored on an experimental basis in previous generation systems [7]. In general, therefore, there are a number of techniques which can now be explored further for managing the move-by-move tracking of microregisters.

4 Relations to alignment

We can consider some established phenomena of alignment in terms of the mechanisms that are available for modelling microregisterial unfolding in texts and interaction. For example, we can adapt one of the examples given by [25]. If one speaker in a dialogue uses the phrase "the sheep that's red" rather than "the red sheep" to assign a colour to some sheep under discussion, then alignment predicts that, via priming, the other speaker will subsequently be more likely to use the first strategy rather than the second, too. Within the semantic formalism that we employ, the intended meaning for these alternatives has a common representation:[1]

```
(s / sheep
   :property-ascription (r / (color red)))
```

Then, within our linguistic model and the description of lexico-grammar employed (essentially systemic-functional grammar as set out in Halliday and Matthiessen [16] and described computationally for natural language generation in Matthiessen and Bateman [23]), we can characterise the production of an associated utterance as follows.

If we do not provide any further constraints, then both of the possible utterances above (and several others) can be generated with our English grammar. However a selection between these can be forced (in this case) by the choice between contrasting grammatical features: for example, somewhat simplified for the purposes of discussion, 'pre-modification' *vs.* 'post-modification'. By default the grammar tries to make a sensible choice between these on the basis of how much semantic material is to fit in the property ascription (e.g., 'the red sheep' *vs.* 'the sheep that used to be red every other day'), but we can also choose to *pre-select* the relevant feature in advance. Such pre-selection has precisely the effect of priming for one construction rather than another that Pickering and Garrod associate with alignment. This, then, is a minimal micro-register: we can state that the

[1] This semantic representation is based on the sentence planning language (SPL) originally defined by Kasper [19], and subsequently used in several natural language generation systems.

semantics	*lexicogrammar*

`(s / animal`
`  :property-ascription`
`  (r / colour ))`
	`post-modification`

Table 1: A simple microregisterial setting that pairs an underspecfied semantic expression with a grammatical constraint

production of this grammatical form primes for the actually selected lexico-grammatical features rather than those that would in principal be possible but which were not selected.

This can be made arbitrarily more complex. The actual example given by Pickering and Garrod draws on experimental results from Cleland and Pickering [9] which showed that the priming effect was much stronger when ascribing a colour to a *semantically similar entity*. That is, "the sheep that's red" was produced far more often after hearing "the goat that's red" than it was after hearing "the book that's red". This shows that the micro-register must consist not only of preselected lexico-grammatical features but, instead, of (at least) *pairs* of semantic:lexico-grammatical expressions that are contingently associated during an interaction. The micro-register established in the current case might then be summarised by the pair shown in Table 1.

The exact degree of specificity for the semantic types (i.e., any 'animal' or just 'mammals', any 'colour' or some particular range, etc.) must be ascertained empirically; the basic mechanism for the formation of such locally active 'routines' or micro-registers is, however, relatively clear. We therefore can import accounts of 'partially idiomatic' expressions and fixed phrases (all interpreted as more or less underspecified fragments of syntactic structure) and combine these with our notion of dynamically grown micro-register pairings for tracking spontaneously created routines during dialogue. This process is depicted graphically in Figure 5.

The description in terms of ontological partitions and lexicogrammatical features may well provide a convenient way of expressing ongoing alignment that is both very succinct and functionally relevant. This is, at present, a research hypothesis and will need to be explored further in concrete computational instantiations. Furthermore, although Pickering and Garrod argue that prioritising decontextualised sentences has made it more difficult for theoretical accounts to see the natural processes of alignment by which dialogue functions, since the functional view of register adopted here is drawn

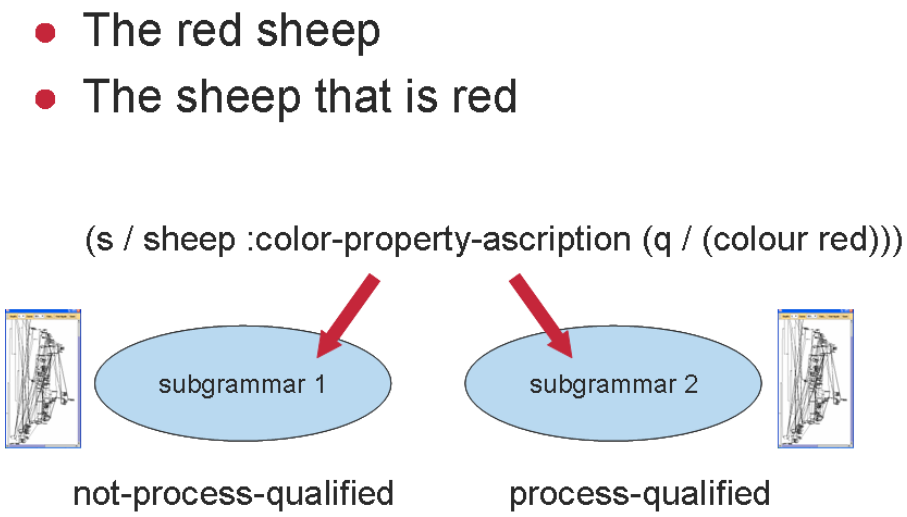

Figure 5: Microregisterial alignment

from a linguistic orientation which insists on the centrality of relating use of language to context, it becomes more natural to consider possible interconnections between its linguistic models and Pickering and Garrod's proposed architecture.

5 Open questions for computational alignment

I will end this brief position statement and research suggestion with an open question that arises very naturally in the context of concrete computational instantiation. Although alignment has been observed to hold in various circumstances, the kinds of linguistic descriptions that have been used in these studies are relatively unspecific compared to the more detailed descriptions necessary for computational use. Given the following dialogue extract, taken from our ongoing empirical work on HRI, we can suggest that alignment of some kind has taken place.

R043f[ol1] ROBOT Is this part of the kitchen?
R043f[ol1] USER This is part of the kitchen

Computationally the task looks a little different. In order to describe the first utterance, we need 59 features (for the clause rank alone) from our lexicogrammar. An extract of these features is shown in Figure 6.

The second utterance contains 62 features, many of which are identical to those of the first utterance. The question for our computational approach

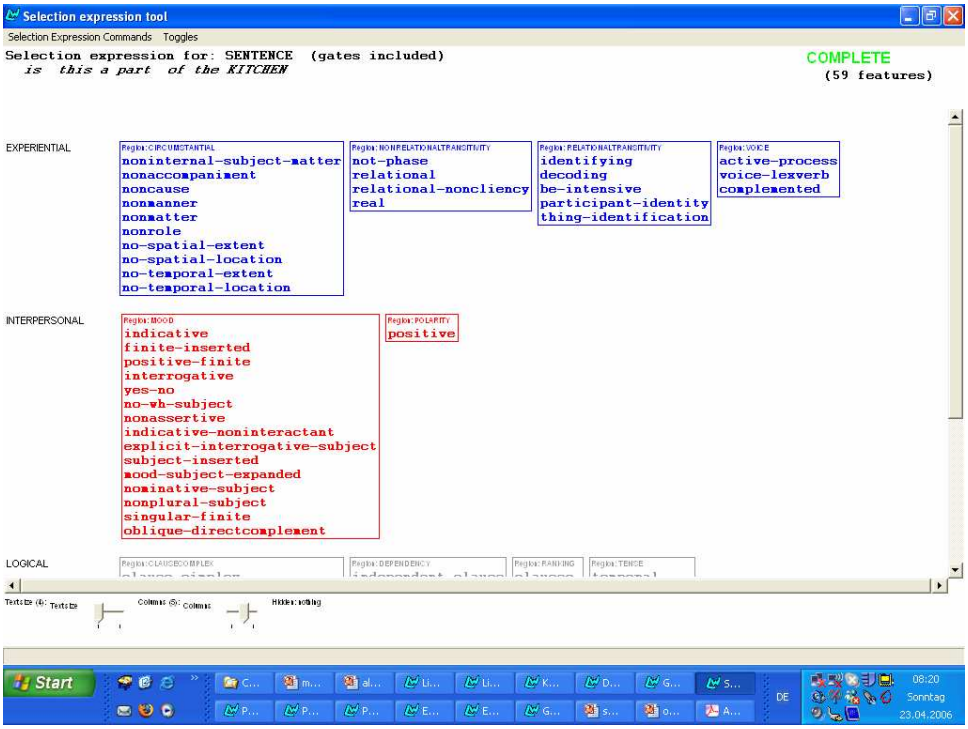

Figure 6: Clause-rank features for the clause “Is this part of the kitchen?”

to alignment is then: which of these features are indications of alignment and which not? We will need to explore in much more detail just how we can decide which features are being maintained across speakers because of alignment and which are reflecting other issues. We will also need to explore over what interactional extent groups of features might be held constant: this will most obviously be related to issues of discourse structure, topic maintenance and the like.

A further, similar example from our data is the following:

USER	and now to the table
USER	and straight
ROBOT	I'm sorry, which kind of table?
USER	a small table that has a cup
ROBOT	is this part of the living room area?
USER	yes
ROBOT	**am I in front of it?**
USER	yes
ROBOT	Oh okay, I see which object you mean.
USER	and now the couch
USER	**now I'm in front of it**

The utterances in bold might also be taken as indicating alignment. This makes it clear that we also need to consider the particular linguistic task that is being performed as the intervening utterance of the robot "I see which object you mean" is performing a completely different discourse task and so does not appear to interfere with the established alignment of the routine for expressing a spatial relationship to an object in the scene. Nevertheless, in the utterance "am I in front of it", we also have a significant number of linguistic features: 59 features in the clause, 17 features for the nominal phrase "I", 17 features for the prepositional phrase "in front of it", and a further 25 features for the nominal phrase "it". Just which of these features are negotiable? Under which circumstances? And for how long? We will also need to address issues of control: as is inherent in the systemic-functional view of language as choice, speakers make choices about what they say and how they say it. Often these choices are abstract and nondeliberative, but regardless of their status they necessarily bring about certain situational trajectories, or discursive positions, rather than others. Here the extent to which a speaker can 'choose' to align or not, or can 'choose' to cooperate in the situation that their interlocutor is pursuing to not, will need to be addressed. This will also no doubt vary according to a variety of situational conditions, some of which have already been revealed from empirical work [11]. This appears to be an issue for both the psychological

and socially oriented approaches as the 'mechanistic' nature of the original interactive alignment proposal is weakened. Was the speaker here choosing to cooperate with the robot or being subjected to alignment?

For a functioning dialogue system, for example, that exhibits alignment, these are all questions that we will need answers for.

One advantage of building such mechanisms into established natural language technology is then that we can explore in natural contexts the consequences of restricting the linguistic features that are available at a very fine level of detail. But, conversely, that very level of detail is itself a significant issue that we will need to learn how to deal with.

Acknowledgement

This research is partially funded by the German Deutsche Forschungsgemeinschaft within the scope of the SFB/TR8.

References

[1] J. Bateman and C. Paris. Adaptation to affective factors: Architectural impacts on natural language generation and dialogue. In *Proceedings of the Workshop on Adaptation to Affective Factors at the International User Modelling Conference (UM'05)*, Edinburgh, Scotland, 2005.

[2] J. A. Bateman. *Utterances in context: towards a systemic theory of the intersubjective achievement of discourse.* PhD thesis, University of Edinburgh, School of Epistemics, Edinburgh, Scotland, 1986. Available as Edinburgh University, Centre for Cognitive Science In-House Publication EUCCS/PhD-7.

[3] J. A. Bateman and R. Henschel. From full generation to 'near' templates without loosing generality. In S. Busemann and T. Becker, editors, *May I speak freely?: Proceedings of the KI'99 Workshop on Natural Language Generation*, pages 13–18, Bonn, Germany, 1999. Available as: DFKI document D-99-01; http://www.dfki.de/service/NLG/KI99.html.

[4] J. A. Bateman, I. Kruijff-Korbayová, and G.-J. Kruijff. Multilingual resource sharing across both related and unrelated languages: An implemented, open-source framework for practical natural language generation. *Research on Language and Computation*, 3(2):191–219, 2005.

[5] J. A. Bateman and C. L. Paris. Phrasing a text in terms the user can understand. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1511–1517, Detroit, Michigan, 1989. IJCAI'89.

[6] J. A. Bateman and C. L. Paris. Constraining the development of lexicogrammatical resources during text generation: towards a computational instantiation of register theory. In E. Ventola, editor, *Recent Systemic and Other Views on Language*, pages 81–106. Mouton, Amsterdam, 1991.

[7] J. A. Bateman and E. Teich. Selective information presentation in an integrated publication system: an application of genre-driven text generation. *Information Processing and Management: an international journal*, 31(5):753–767, Sept. 1995.

[8] D. Biber. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, 1988.

[9] A. Cleland and M. J. Pickering. The use of lexical and syntactic information in language production: evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49:214–230, 2003.

[10] M. Cross. *Choice in text: a systemic approach to computer modelling of variant text production*. PhD thesis, School of English and Linguistics, Macquarie University, Sydney, Australia, 1992.

[11] K. Fischer. *What Computer Talk Is and Is not: Human-Computer Conversation as Intercultural Communication*, volume 17 of *Linguistics – Computational Linguistics*. AQ-Verlag, Saarbrücken, 2006.

[12] S. C. Garrod and A. Anderson. Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, 27:181–218, 1987.

[13] S. C. Garrod and A. J. Sanford. Discourse models as interfaces between language and the spatial world. *Journal of Semantics*, 6:147–160, 1988.

[14] M. Gregory and S. Carrol. *Language and Situation: Language varieties and their social contexts*. Routledge and Kegan Paul, London, 1978.

[15] M. A. K. Halliday. *Language as social semiotic*. Edward Arnold, London, 1978.

[16] M. A. K. Halliday and C. M. Matthiessen. *An Introduction to Functional Grammar*. Edward Arnold, London, 3rd edition, 2004.

[17] R. Henschel. Compiling systemic grammar into feature logic systems. In S. Manandhar, W. Nutt, and G. P. Lopez, editors, *CLNLP/NLULP Proceedings*. 1997.

[18] L. Hjemslev. *Prolegomena to a theory of language*. University of Wisconsin Press, Madison, Wisconsin, 1961. Originally published 1943; translated by F.J.Whitfield.

[19] R. T. Kasper. A flexible interface for linking applications to PENMAN's sentence generator. In *Proceedings of the DARPA Workshop on Speech and Natural Language*, 1989.

[20] J. R. Martin. *English text: systems and structure*. Benjamins, Amsterdam, 1992.

[21] C. M. I. M. Matthiessen. The systemic framework in text generation: Nigel. In J. D. Benson and W. S. Greaves, editors, *Systemic Perspectives on Discourse, Volume 1*, pages 96–118. Ablex, Norwood, New Jersey, 1985.

[22] C. M. I. M. Matthiessen. Register in the round, or diversity in a unified theory of register. In M. Ghadessy, editor, *Register Analysis. Theory and Practice*, pages 221–292. Pinter, London, 1993.

[23] C. M. I. M. Matthiessen and J. A. Bateman. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press, London and New York, 1991.

[24] T. Patten. *Systemic Text Generation as Problem Solving*. Cambridge University Press, Cambridge, England, 1988.

[25] M. J. Pickering and S. Garrod. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27(2):169–190, 2004.

[26] J. Ure and J. Ellis. Language varieties – register. In *Encyclopedia of Linguistics (Information and Control)*, volume 12, pages 251–259. Pergammon Press, Oxford, 1969.

[27] J. N. Ure and J. Ellis. Register in descriptive linguistics and linguistic sociology. In O. Uribe-Villegas, editor, *Issues in Sociolinguistics*. Mouton, The Hague, 1977.

Reasoning on Action during Interaction

Structured Information State for Flexible Dialogue

Robert J. Ross
University of Bremen, Germany
robertr@tzi.de

Abstract

Dialogue based interaction with service robots of the near future will be based on one of two paradigms: the use of a tool, or interaction with a partner. In this talk I review some recent work which is based in a school of thought that believes that the former is simply a matter of engineering application, but that the latter is achievable, but only through continued research into dialogue systems that sufficiently leverages off linguistic knowledge. The work presented here attempts to overcome limitations of the Information State Update (ISU) approach to dialogue management through explicit dialogue modelling and rich multi-stratal representations of the information state which do not disregard detail for simplicity in canonical form. This work is being implemented in the context of Corella, an information-state based dialogue management toolkit, and has been used in the development of a spoken dialogue system for Rolland the autonomous wheelchair.

1 Introduction

The Information State Update (ISU) based approach to dialogue management [8, 17] advocates dialogue manager construction based around discourse objects (e.g., questions, beliefs) and rules which encode relationships between these objects. As such, ISU based systems may be viewed as practical instantiations of agent-based models, instantiations where the broad notions of beliefs, actions, and plans, are replaced with more precise semantic types and their inter-relationships. ISU modelling techniques provide an open palette of modelling choices and possibilities, which while being appealing in reducing constraints on system developers, also leave many questions left to be answered.

Following initial studies into the use of ISU based dialogue managers in producing dialogue systems for human-robot interaction [14], some deficien-

cies of ISU implementations and the dialogue models commonly developed upon them were identified:

- Opacity of Control – As with all declarative rule based systems, the use of a potentially large number of rules to define information state transitions can lead to systems that are difficult to design and debug, with unforeseen logic errors difficult to trace and leading to potentially serious side-effects.
- Over Simplicity of Modeling – Furthermore, many of the dialogue models applied to ISU systems take rather elementary views of either dialogue structure, language semantics, or the relationship between language and domain knowledge.
- Limited Tool Support – ISU based toolkits still provide a limited functionality, particularly with regard to rapid prototyping, code reuse, and debugging.

In the remainder of the talk I will describe ongoing work which attempts to overcome these issues by developing an Information State Update modelleting methodology which on one hand cleanly separates operational from dialogue structure, while on the other uses deep, fine-grained semantics to model linguistic and non-linguistic knowledge within the spoken dialogue engine. I will come to a close by describing Corella, a hybrid Information State based dialogue management library that has been built around these ideas, and which has been used in the development of a spoken dialogue system for Rolland the autonomous wheelchair.

2 Separating Control and Discourse Structure

The separation of control structure from dialogue structure has been a common theme in the evolution of dialogue system design [11]. Whereas finite state-based dialogue systems often encode both control structure and dialogue structure, this has been a tendency in frame-based and agent-based models to abstract control structure from the dialogue structure or models to be treated as resource.

However, rule based dialogue systems, including to some extent vanilla ISU models, have a tendency to represent all aspects of the dialogue modelling as the application of various 'update rules'. While these rules may often be classed into particular update sets which in tern can be sequenced

through a high-level control structure, the update rules themselves retain a mixture of control structure as well as purely dialogue structure. Thus, it is often difficult to separate out the dialogue structure or *resource* from system control or *process*. This in turn can lead to relative simple dialogue structures being employed in implementations simply to cut down on the complexity of the ISU model. An alternative approach pursued here is to explicitly extract the dialogue structure from ISU update rules, and guarantee that all dialogue structure may be modelled externally and implemented through dedicated domain specific plans which are in no way reliant on explicit rules. While this may seem a relatively trivial issue of design, we believe that this issue is symptomatic of a gulf between dialogue management and discourse modelling which is preventing dialogue system application from leveraging off empirical studies.

The mixed treatment of dialogue model and control model can even be seen where researchers have attempted to analyse the meaning of a dialogue model. In [18], Xu et al view dialogue models as being categorisable into two groups: pattern based models and plan based models. In pattern-based models, Xu includes recurrent interaction patterns or regularities in dialogue at the illocutionary force level of speech acts are identified [16]. While, in the second approach, i.e. plan-based models, dialogue is modelled in terms of speech acts and their relation to plans and mental states in the greater agent design [2]. Thus, in Xu's view, pattern based models describe what happens, but care little about why. Conversely, plan-based models contextualise speech acts within the greater agent plans and rationality, but are costly and care little about the actual patterns of dialogue identified in human-human or human-computer interaction. Instead, we view this distinction as one between Generalized Dialogue Models which describe the overall patterns of dialogue as a linguistic resource, and, from a computational perspective, dialogue plans, which inherently capture such generalized dialogue models within application.

To develop ISU based dialogue systems which separate out dialogue modelling from control and implementation issues, two questions must be addressed: (a) how do we capture dialogue models at an abstract level? and (b) how then may such models be related to traditional ISU based methodologies? The first question is an issue of modelling approach which has consequences both for formal linguistic analysis and to verification of the linguistic properties of a system. The second question is one of implementation methodology, and how the use of cleanly defined dialogue models can then be used to aid in the construction of flexible dialogue systems. I discuss the first of these issues below, while the second is addressed in Section 4.

2.1 Capturing ISU Based Dialogue Models

The structuring approaches used in Information State Update techniques do not in themselves lead easily to the capture of the multi-tired nature of dialogue, where clarification situations and multiple overlapping dialogue threads which may characterise a mixed initiative human-robot interaction [13]. We must first establish a distinction between the information state based dialogue management model or paradigm, and the dialogue models which can be implemented with such a paradigm. Broadly, we share the view that as a paradigm, the Information State based approach is extremely flexible and can support the implementation of a wide variety of dialogue models. Such implementations range from simple finite state models using registers and state transition rules, to what we refer to as *IS centric dialogue models* where the dialogue modelling approach is inextricably linked to the modelling of the dialogue's information state. Examples of such modes include those models behind GoDiS, and EDIS [17].

One alternative modelling approach which has been applied extensively for over three decades has been the use of recursive state transition networks. One well-known example of such a modelling is the '**Co**nversational **R**oles' COR Model of Sitter & Stein [16], which set out as a communicative-based approach to interaction in the relatively limited context of information-seeking dialogues. Individual dialogue moves at the interlocutionary force level may be achieved through either individual acts, or alternatively through a sub-traversal of the structure – corresponding to a sub-dialogue.

One set of dialogue models which arguably has the tightest computational link to the Information State paradigm are those underlying Larsson's IBiS systems [7]. These IBiS models, developed to explore the area of *Issue Based Dialogue Management*, place structural emphasis on conversation goals as issues and questions, using them as a basis of dialogue management. Such modelling, achieved through a rich structuring of dialogue in terms of information state and the range of *update* and *selection* rules, results in effective dialogue management for a wide range of discourse phenomena including grounding and accommodation – these phenomena not easily addressed by previous dialogue modelling approaches.

Despite the apparent complexity the IBiS system descriptions, the domain independence of IBiS1 through IBiS4 makes it possible to extract underlying dialogue structure. This can be done by examining selection and update rules with regard to the movement of information on and off the latest utterance record of the information state, i.e., `/SHARED/LU/MOVE` [7]. For example, Figure 1 depicts an abstraction of IBiS1's underlying dialogue

model. Once extracted, such a model can then be added to the information state, and used to supply context information where applicable.

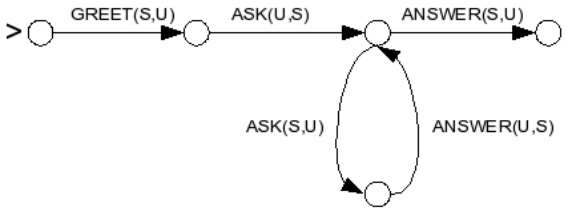

Figure 1: Abstraction of IBiS1 Dialogue Model

In recent works Hui has described the use of a recursive transition networks to capture the structure of interaction between users and robotic wheelchair in a shared control task [15]. While it would be possible to encode dialogue models through relatively arbitrary means, Hui has applied formal specification techniques based on Hoare's *Communicating Sequential Processes* (CSP) language [4, 12], to facilitate property analysis, model comparison and implementation verification. In Section 4 I describe how such a model can be used to improve ISU based human-robot interaction.

3 Fine Grained Information Structure

While linguistic and empirical studies of actual human-human or human-robot interaction attempt to capture the precise details of any given interaction in considerable detail, the same is rarely true of computational approaches to dialogue modelling and dialogue system construction. To the contrary, the use of *canonical form* is often seen as a key tool in producing practical dialogue systems [5].

Unfortunately however, such simplifications of the information state, if introduced at the wrong level of abstraction, can lead to considerable loss of reasoning and linguistic control. To illustrate, consider a simple example from the robotics domain where a user request that the robot *turn left* through one of the following three utterances:

(32) a. turn to the left
b. turn left
c. take the next turn left

All three utterances do of course seem to be equally applicable to achieving the goal of causing the system to turn to the left. Thus, a naive approach, but one ultimately assumed by some views of information state structuring would be to represent such commands within a dialogue system with a predicate such as `turn(left)`, and use keyword spotting of *turn* and *left* to extract such a command from a user's language. In practice such an assumption is predictably enough misguided since all three utterances will of course have very different meanings depending on their use in context: (1a) to the most part is used in static contexts to communicate a request for reorientation while planar location is effectively unaltered; conversely, (1c) may often be used in dynamic contexts to achieve a vector change thus resulting in a net planar motion; while, (1b) is slightly more ambiguous, taking on the meaning of (1c) in dynamic contexts, and sometimes taking on the meaning of (1a) in static contexts.

The fact that the three utterances above do not map directly to a single concept should not of course be surprising since language ultimately serves to facilitate some communicative goal, and the subtle differences that speakers make ultimately reflects the precise goal they wish to convey. This then is a strong argument for guaranteeing that we do not attempt to over simplify the ontological structuring within the dialogue systems which construct for HRI. Particularly when we strive toward interaction with un-trained users, the nuances in the language which is applied may be key to efficient communication and ultimately high user satisfaction.

4 Dialogue Management with Corella

To develop rich dialogue systems which possess a degree of flexibility which approach that which could provide *natural interaction*, we must be willing to put effort into the development of dialogue technologies that make use of and integrate available linguistic results on dialogue and knowledge structure, while remaining efficient practical implementations for engineers to apply to domain applications. To help address such requirements we have developed *Corella* as a hybrid dialogue management engine that extends the standard information state paradigm with greater emphasis on ontological and discourse structuring. Here, I give a brief overview of Corella and its use.

Corella came about through the need for a spoken dialogue system which could process detailed spatial language between naive users and an autonomous robotic wheelchair in shared-control tasks [6]. The wheelchair,

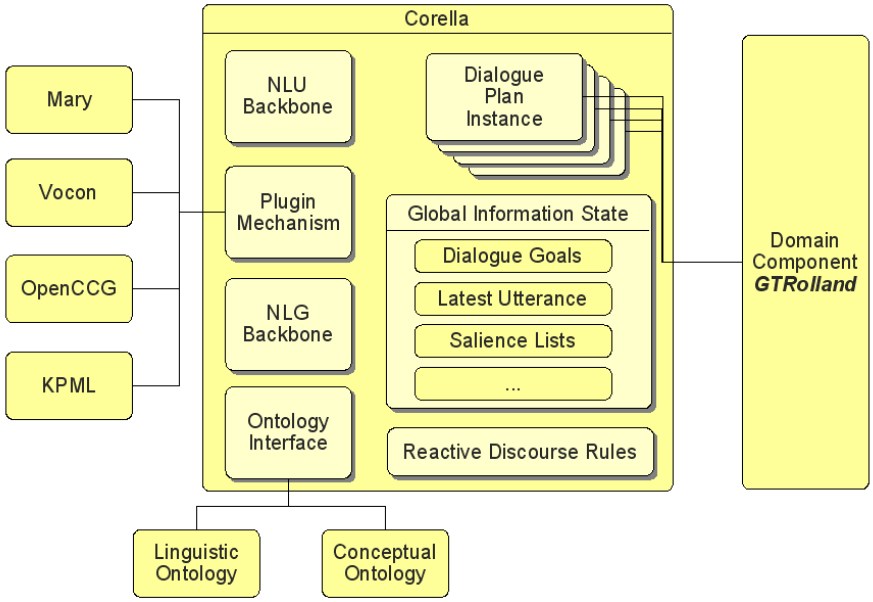

Figure 2: Spoken Dialogue System for Rolland III wheelchair.

Rolland III, is the latest in s series of intelligent wheelchairs at the University of Bremen [10], and should be capable of voice control for users who may suffer physical impairments which would limit either their visual sense or manual dexterity. Thus, the dialogue system must be: (a) capable of processing spatial expressions including spatial descriptions, basic and complex navigation instructions, and route descriptions; (b) must be adaptable to different user types depending on the particular abilities of individual users; and (c) should allow to the greatest degree possible to process *natural language* to maintain a low learning curve for users.

Figure 2 depicts Corella in the context of Rolland's spoken dialogue system at an architectural level. In comparison to some of our earlier work in dialogue system construction [6], a relatively tight coupling has been employed between the dialogue engine, the domain component, and external language technology components. Other notable features of the dialogue engine include the application of a *functional semantics* as the first level within a two-level semantics structuring of information state; the use of domain specific dialogue plans which may be verified against the abstracted dialogue models introduced earlier; and the management of multiple threads of interaction.

Corella's information state implements a two-level semantics model where the first level of semantics is a so-called linguistic semantics which acts as interface to language technology components, while the second semantics level is a conceptual semantics used for primary domain reasoning or interfacing with domain applications. Motivation for a two-level semantics comes from many different directions, and were reviewed extensively by Farrar & Bateman in [3]. Some motivations include the fact that users often make utterances that are not literally true with respect to an underlying model; that dialogue systems which mix linguistic and conceptual knowledge can become overly complex; and that adding an additional layer of representation allows us to cleanly provide a representation of surface form language which is sufficiently fine grained to facilitate flexible language structure. Two level semantics are often confused with issues of quasi-logical form (QLF) versus logical form (LF) issues as exemplified by the Core Language Engine [1]. We should make clear here that these are two operate issues, and that it is possible to have a single-level semantics system that employs both QLF and LF. Two-level semantics is best characterised by the use of two separate ontologies, one for the linguistic semantic categories, and another for the underlying conceptual and domain knowledge held by the agent. While the use of two-levels of representation within the ddialogue engine's information state can provide some clear advantages, it should of course be realized that these advantages do not come without their own costs.

Generalized dialogue plans are applied to encode particular dialogue phenonema at the implementation level and may be considered as a specialization of the generalized dialogue plans introduced earlier. We believe that the use of generalised dialogue models within the information state paradigm provides two advantages that would not be easily achieved otherwise. Firstly, a clear model of expected discourse moves can be extracted from the recursive transition network that encodes a generalised dialogue model. Thus, applying a similar approach to [9]'s use of allowed attachments, the search space for intention identification can be considerably reduced. Secondly, abstraction of the many declarative rules that constitute an information state implementation can make evaluation of the *quality* of the underlying dialogue model more straightforward. Furthermore, through simulation, the accuracy of rules in an information state based implementation can be judged against the sought after generalized dialogue model. Moreover, when considered in the light of the ever increasing application of SDS to *safety critical* applications such as service robotics and automotives, the need for analysis and verification of dialogue models underlying spoken dialogue systems becomes even more imperative.

In investigating the relationship between the IS paradigm and GDMs encoded as Recursive Transition Networks (RTNs), it is important to distinguish between the encoding of RTNs through information state, and the use of RTNs in information state. The former of these two approaches refers to the fact, as observed in [17], that recursive transition networks can be directly encoded through an information state based implementation through the use of a stack to record a history of nested state positions, and a collection of update rules to encode state transitions. The latter view, however, reflects the use of RTNs as part of the data types used to store information state; this being analogous to the use of queues, records or predicate sets. It is the latter view of using RTNs within the information state that can best leverage off existing generalised dialogue models.

While dialogue models such as COR are principally intended only to describe one *thread* of conversation, the nature of mixed-initiative systems, often viewed as favourable for human-robot interaction, places additional requirements of robustness in the event of parallel conversational threads, e.g., a robotic wheelchair might wish to inform a user of a system event in the middle of a route description task. By allowing multiple instantiations of GDMs within the information state, implementations can effectively track parallel conversational threads. Indeed, the application of models like this to multi-threaded dialogue systems might be considered essential to language understanding. A deeper investigation of the issues involved in the multi-threading is not investigated further here and is left for future work.

5 Summary & Future Work

Driven by the desire to achieve human-robot interaction based on natural discourse rather than the metaphoric use of a tool, we are looking at building dialogue systems which build upon rich resource models while yet guaranteeing that the system's operate in an effectively real-time manner. Specific factors motivating this approach have been the application of explicit dialogue structure within the control mechanisms of information state update dialogue systems, and the need for ontological sophistication in knowledge structuring to capture the true meaning of a user's utterance without over simplification. Such goals should not however remain lofty academic exercises. Thus, we developed Corella as a dialogue engine which makes use of rich ontological structuring and a modelling of dialogue plans which can be mapped to empirically derived generalized dialogue models.

Our application of these techniques to Rolland the autonomous wheelchair

continue. To this end, a more formal analysis of the resultant dialogue implemenation is underway.

References

[1] H. Alshawi, editor. *The Core Language Engine*. MIT Press, Cambridge, Massachusetts, 1992.

[2] P. R. Cohen and C. R. Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212, 1979.

[3] S. Farrar and J. Bateman. Linguistic ontology baseline. SFB/TR8 internal report I1-[OntoSpace]: D1, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany, 2006.

[4] C. A. R. Hoare. *Communicating Sequential Processes*. Prentice-Hall, 1985.

[5] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 2000.

[6] B. Krieg-Brückner, H. Shi, and R. Ross. A safe and robust approach to shared-control via dialogue. *Journal of Software*, 15(12):1764 –1775, 2004.

[7] S. Larsson. *Issue-Based Dialogue Management*. Ph.d. dissertation, Department of Linguistics, Göteborg University, Göteborg, 2002.

[8] S. Larsson and D. Traum. Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, 6(3-4):323–340, 2000. Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering.

[9] O. Lemon, A. Gruenstein, and P. Peters. Collaborative Activities and Multi-tasking in Dialogue Systems. *Traitement Automatique des Langues (TAL)*, 43(2):131–154, 2002. Special issue on dialogue.

[10] C. Mandel, U. Frese, and T. Röfer. Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, 2006.

[11] M. F. McTear. Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90 – 169, 2002.

[12] A. W. Roscoe. *The Theory and Practice of Concurrency*. Prentice-Hall, 1998.

[13] R. J. Ross, J. Bateman, and H. Shi. Using Generalised Dialogue Models to Constrain Information State Based Dialogue Systems. In *the Symposium on Dialogue Modelling and Generation, 2005*, Amsterdam, The Neterhands., 2005.

[14] R. J. Ross, H. Shi, T. Vierhuf, B. Krieg-Bruckner, and J. Bateman. Towards Dialogue Based Shared Control of Navigating Robots. In *Proceedings of Spatial Cognition 04*, Germany, 2004. Springer.

[15] H. Shi, R. J. Ross, and J. Bateman. Formalising control in robust spoken dialogue systems. In *Software Engineering & Formal Methods 2005*, Germany, Sept 2005.

[16] S. Sitter and A. Stein. Modeling information-seeking dialogues: The *Co*nversational *R*oles model. *Review of Information Science*, 1(1):n/a, 1996. (On-line journal; date of verification: 20.1.1998).

[17] D. Traum and S. Larsson. The information state approach to dialogue management. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer Academic Publishers, Dordrecht, 2003.

[18] W. Xu, B. Xu, T. Huang, and H. Xia. Bridging the gap between dialogue management and dialogue models. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 201–210, Philadelphia, USA, July 2002. Association for Computational Linguistics.