How Entrainment Increases Dialogical Effectiveness

Robert Porzel
European Media Laboratory, GmbH
Schloss-Wolfsbrunnenweg 33
D-69118 Heidelberg, Germany
porzel@eml.org

Annika Scheffler
European Media Laboratory, GmbH
Schloss-Wolfsbrunnenweg 33
D-69118 Heidelberg, Germany
scheffler@eml.org

Rainer Malaka
European Media Laboratory, GmbH
Schloss-Wolfsbrunnenweg 33
D-69118 Heidelberg, Germany
malaka@eml.org

ABSTRACT

Recent work on spoken and multimodal dialogue systems is aimed at more conversational and adaptive systems. We show that - in certain dialogical situations - it is important for such systems to adapt linguistically towards the users. We report ongoing work in addressing these tasks, describing the empirical experiments, investigating the feasibility of annotating the data reliably by human annotators as well as analyses of the collected data and results from a *Wizard of Oz* experiment. Additionally, entrainment, as such, is firstly examined in the domain of multimodal dialogical assistance systems.

Keywords

Entrainment, Dialog Systems, Adaptive Interfaces

1. INTRODUCTION

The complete understanding of specific characteristics of dialogical interaction is still an unresolved task for (computational) linguistics. Entrainment, i.e. linguistic adaptation, presents such a specific characteristic in dialogue. It has been explored by linguists [17] and recently came into focus of computational linguistics [8]. It is 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 users choice of terms rather than staying with their own fixed vocabulary.

Research on multimodal dialogue systems in the past has focused on engineering the various processing stages involved in multimodal human-computer interaction (HCI), e. g., robust automatic speech & gesture recognition, intention recognition, natural language generation or speech synthesis ([1], [9], [3]). Alongside these efforts the characteristics of computer-directed language have also been examined as a general phenomenon ([30], [29], [10]).

The adaptation to the user has been examined ranging from epistemic factors such as prior knowledge or cognitive competence [22] to multimodal preferences [11]. The notion of lexical entrainment was first established by Garrod and Anderson [17] and later explored by Brennan [4, 5].[1]

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.

The purpose of the work presented in this paper is to develop and test an annotation scheme for annotating entrained terms within a corpus of previously recorded and transcribed dialogues of a support hotline obtained in a study described in Sections 3 and 4. In order to find out if the implementation of entrainment is feasible in spoken dialogue systems, we needed to examine how reliably human annotators detect cases of entrainment within a human-human dialogue and to examine specific regularities and characteristics of entrainment behavior, e.g. when to lead and when to follow. The results of these findings - presented in Section 5 will then be used to perform a corresponding HCI experiment - described in Section 6, which, in turn, will be analyzed based on the findings shown and discussed below. Concluding remarks are given in Section 7.

2. ENTRAINMENT

The first studies and descriptions - as subsequent ones - of the particularities of dialogical human-computer interaction, then labeled as *computer talk* in analogy to *baby talk* by [30], focused on:

- proving that a regular register for humans conversing with dialogue system exists, e. g., those of [19] and [15],
- describing the regularities and characteristics of that register, as in [20] or [10].

The results of these studies clearly show that such a register exists and that its regularities can be replicated and observed again and

[1] 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

.

again. In general, this 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 explicitly asked or studied are:

- when and how entrainment occurs in the human interlocutor following the computer and how the analog human-human linguistic behavior functions in this respect
- what could be the effects of lexical adaptations in the form of entrainment on the computer's side in human-computer dialogues.

In this work we adopt the notion of lexical entrainment as defined by Garrod and Anderson [17] and Brennan [4, 5]. Both research teams found out 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. Furnas *et al.* [16] have labeled this phenomenon the *Vocabulary Problem*. Although there are no real synonyms, i.e. two words that in all contexts could be used interchangeably, people still have numerous options when referring to an object in a given context. For instance, in a user study conducted by Furnas *et al.* subjects used several different words for *to delete*: *change*, *remove*, *spell* or *make into*.

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* [6]. If people wish to align within a conversation and terms are adopted, Garrod and Anderson [17] denote the one who introduces a term as the *leader* and the one who adopts it as the *follower*.

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 [4]. After that they hedge, i.e. they mark the term as provisional, pending evidence of acceptance from the other [5]. Only then do they agree on the same choice of words. As a last step, entrained terms are no linger indefinite and can be shortened, e.g. via anaphora, one-pronominalization or gapping.

Core and Moore [8] point out that entrainment is of major importance in tutorial systems. They argue 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 [12] and searches for synonyms that match.

To get back to the notion of leader and follower, 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 the study introduced below, is whether in shorter exchanges, e.g. in an assistance, help-desk or hotline setting, we find specific cases of entrainment or not.

3. DATA COLLECTION

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 evaluation paradigm suggested by Paek [21]. 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 there are two basic levels of entrainment, phrasal entrainment and lexical entrainment. However, more specific kinds could be distinguished further which resulted in the entrainment model listed in Table 1. These kinds of entrainment are briefly explained in the following sections.

Phrasal Entrainment - Ad Hoc Entrainment - Later Phrasal Entrainment
Lexical Entrainment - Classical Entrainment - One Time Entrainment - Reviving Entrainment - Granularity Entrainment - External Entrainment - Cognitive Entrainment - Anaphoric Entrainment

Table 1: A fine-grained Entrainment Model

3.1 Phrasal Entrainment

This category deals with the adaptation of phrases. Additionally, two subcategories occurred, i.e. Ad Hoc Entrainment (see Examples (1) and (2)[2]) and Later Phrasal Entrainment. The first means that phrases are adopted within the same entrainment segment, while the latter denotes the adaptation of phrases that are adopted some time later during the same dialogue.

(1) *also* *dann* *tausch* *ich* *die* *beiden* *aus*
well then exchange I the two out
well then I exchange both

(2) *genau* *dann* *tauschen* *Sie* *die* *beiden* *aus*
exactly then exchange you the two out
exactly then you exchange both

[2]Underneath all German examples are gloss-by-gloss and semantic translations. Entrained items are underlined.

3.2 Lexical Entrainment

In contrast to phrasal entrainment, lexical entrainment deals with lexical terms that are adopted. Lexical entrainment occurs in various forms and uses different linguistic mechanisms, which are discussed below.

Classical Entrainment

This kind of entrainment comes closest to the phenomenon described in Section 2: a term is suggested by one of the interlocutors, whereupon one of them either acts as leader or follower. Both operator and subject stay with that term throughout the whole dialogue without suggesting a different one. In classical entrainment, the term can be adopted anywhere in the dialogue, regardless of the particular dialogue turn (or entrainment segment).

One Time Entrainment

One time entrainment means that a term is entrained once, but it is not kept throughout the dialogue. It is thus an instance of classical entrainment.

Clarifying Entrainment

This kind of entrainment is an adaptation that is linked to a clarifying question or statement. This, for example, occurs when one of the interlocutors employs a term in a question, which is then explicitly rejected by the other one.

Reviving Entrainment

If a term has so far been classically entrained, it happened that for indistinct reasons either the operator or the subject broke out of that pattern for one or two dialogue turns, but then reverted to the classically entrained term again.

Granularity Entrainment

This kind of entrainment only occurred in one of the ten dialogues that were recorded. However, it did occur which suggests that it might reoccur in further experiments. Especially in dialogues that deal with giving instructions there are different levels of granularity. For example you can imagine that the operator just explained how to access the Internet and referred to the mouse pointer as *cursor*. As a next step, she instructed the subject how to type in an Internet address in the address line. The tricky thing was that the cursor had to appear as a red arrow within the address line, otherwise typing was not possible. In this case the operator switched from the previously entrained *cursor* to *the red arrow*. This makes sense, because at this level, the *red arrow* is a much more salient hint than *the cursor*. When the operator came back to the previous level, she would revert to saying *the cursor*. Interestingly, the subject followed that pattern and adapted both terms, depending on the granularity level of the instruction process.

External Entrainment

The name of this kind of entrainment already suggests that in fact it is the kind that is farthest away from Brennan's findings. This is because in the strict sense adaptation in this case is not initiated by the dialogue partners themselves, but it is imposed on them from factors outside of the conversation. In the experimental setting external entrainment occurs if the subject reads a term that is given within the setting. This means that this word is not a term she chose herself, but rather a term that occurs in the TV vocabulary. For instance, the word *Channels* is part of the menu navigation. If the subject refers to that menu point and the operator uses the same word this is not because she adapts towards the subject, but because she knows that the menu says *Channels* and that any other word that might come to her mind would confuse the subject. In fact, both subject and operator adapt to the TV vocabulary rather than their own, which is entrainment, albeit motivated by external circumstances.

Cognitive Entrainment

Cognitive Entrainment presents a special case, since it is not the term itself that is entrained. Instead both interlocutors converge to the same cognitive level. There is an OK-key in the center of the remote control. In one of the dialogues the subject, who was a computer expert, referred to this key as the ENTER key simply because the functions of the OK-key are similar to the ones of the ENTER key of the computer. If the operator refers to the OK-key as the RETURN-key in future dialogue turns, this proves that an adaptation took place, albeit one that is related to cognitive patterns, rather than actual terms.

Anaphoric Entrainment

In Section 2 we noted that people tended to shorten terms after they had adopted them. The same phenomenon could be shown in the dialogue data of the MOS study. In Anaphoric Entrainment an entrained term is reduced - or rather shortened - to the definite article, which works at the same time as an anaphora. For instance, the entrained term is *die Taste* (the key), which is later referred to as *die*, omitting *Taste*. Strictly speaking the originally entrained term *Taste* has been shortened to *die* while at the same time *die* functions as an anaphora, co-referring to the same entity. The translation into English of *die* is *it*, which would as well be anaphoric, yet, it is not a direct shortening of the expression *the key* which means that in English it would not count as an entrained word.

4. ANNOTATION

The annotation was conducted in order to find out the following aspects:

- Can entrainment be detected?
- If yes, which kind of entrainment is it?
- Who was leader who was follower?

For that purpose a manual was created that contained instructions on how to mark the aspects mentioned above. The manual gave a definition of entrainment and also provided examples taken from the dialogues. It also gave information about the role of the interlocutors, i.e. that they could either act as leader or as follower. Additionally, it contained definitions of each kind of entrainment, e.g. what is classical entrainment, one time entrainment etc. and how to mark each kind of entrainment, e.g. CE, OE etc. Furthermore, the manual gave instructions on how to conduct the annotations, e.g. how to form entrainment segments and how to annotate them. If no entrainment occurred in a segment it was annotated as NE.

4.1 Annotation Scheme

For the annotation, any two consecutive dialogue utterances were coupled. The coupled dialogue utterance were grouped as one *entrainment segment*, encompassing an utterance i and its successor $i + 1$ (Examples 3 and 4). The next segment than repeats (uses) the successor $i + 1$ as i' with its successor $i' + 1$, e.g. Examples 4 and 5. Each entrainment segment was to be marked by the operator's role as follower or leader and which kind of entrainment could be detected. This is illustrated in examples (3 -6), in which

the capital letters correspond to the annotation tags. In this example the Operator is the follower (OP=F), *keyboard* is a classically entrained term, marked as CE. *Dann lieber das Keyboard* is a phrase that has been entrained ad hoc, marked as AE. Example (6) show an non-entrained pair marked as NE.

(3) *ja, gut, ok, dann lieber das Keyboard*
yes, good, o.k., then rather the keyboard

(4) *ja dann nehmen Sie mal lieber das Keyboard*
yes then take you once better the keyboard
... das Keyboard ist genauso wie
... the keyboard is exactly how
OP=F, CE=Keyboard, OP=F, AE=dann...Keyboard

(5) *ja dann nehmen Sie mal lieber das Keyboard*
yes then take you once better the keyboard
... das Keyboard ist genauso wie
... the keyboard is exactly how

(6) *richtig*
right
NE

5. RESULTS OF THE STUDY

5.1 Reliability of the Annotation

Annotations should always be performed by at least two human annotators. This is because for one thing an annotation, which is done by a human, is based on an individual opinion, which might differ considerably from the opinion of someone else. Another drawback is that humans make mistakes. For example, a word that has been entrained might be overlooked and thus not be marked as such. This is why the second annotator serves as a control medium, since it can be measured in how far the results of the annotations are reliable. This is done by means of the Kappa coefficient [7], which measures if the agreement of two annotations is reliable. The calculation of the Kappa coefficient is calculated by subtracting the expected agreement for an annotation task from the observed agreement and divide the result by one minus the expected agreement. Interpretations of Kappa vary, we follow the interpretation of Altman [2]: poor (< 0.2), fair ($0.2 - 0.4$), moderate ($0.4 - 0.6$), good ($0.6 - 0.8$) and very good agreement ($0.8 - 1$). When the agreement of annotations was evaluated, three different criteria were considered. The annotations were analyzed considering entrainment segments, phrases, terms and kinds of entrainment as markables.

Dialogue Turns

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. As example pair (3 and 4) showed, 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 2 shows the distribution of assigned values (N/NE) in percent. The measured agreement was K = 0.76, which showed a good reliability as for agreement between the annotators according to the interpretation by Altman [2].

	Annotator 1	Annotator 2
Segment with E	33%	28%
Segment with NE	67%	72%

Table 2: Annotated Segments

Phrases

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 3 and the agreement was K = 0.92, which shows an excellent reliability.

	Annotator 1	Annotator 2
Phrases with E	7%	6%
Phrases with NE	93%	94%

Table 3: Annotated Phrases

Terms

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 4. Again the reliability of agreement was excellent, since the Kappa result was K = 0.82.

	Annotator 1	Annotator 2
Terms with E	18%	15%
Terms with NE	82%	85%

Table 4: Annotated Terms

Kinds of Entrainment

What seemed to be very difficult was to assign the right kind of lexical entrainment. In fact the agreement was so poor that it resulted in a negative Kappa result. This shows that the annotation for the different kinds of entrainment is mainly a question of interpretation. Many times it was not obvious which kind of entrainment was suitable. This might be because the manual contained many extra conditions, which lead to difficulties regarding the distinction between different subcategories. For instance, an anaphora that referred to a classically entrained term, was counted as a classically entrained term rather than an anaphoric one, although, strictly speaking they should have been counted separately for both categories. Despite these overlaps the annotators did not find that the annotation procedure had been explained too vaguely. This shows that both annotators rather had different ideas about each kind of entrainment and the interpretation thereof. This may indicate that the entrainment model was designed too tightly and too complex in order to distinguish between different kinds. Two kinds that were often confused were classical entrainment and external entrainment. Obviously the dialogue context solely in form of textual transcripts did not reveal

enough information about the question whether a word had been adopted from the TV's vocabulary or if it had been a word adopted from one of the interlocutors. The same phenomenon can be applied to cognitive entrainment. On many occasions it was not at all apparent if both interlocutors had entrained on the cognitive level. This result confirms once more that human language is a complex issue and that an insufficient dialogue context does not always reveal which kind of entrainment was used.

5.2 Statistical Analysis

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.

Phrasal Entrainment

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

Lexical 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. 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 for the different kinds of entrainment, Figure 4 displays that classical entrainment and external entrainment are the most dominant kinds of entrainment. From the perspective of spoken dialogue systems we can safely conflate external and classical entrainment, as a system - capable of adopting lexically - can do so regardless of the fact whether external stimuli or the user introduces the term. All the other kinds do not even hit the 10% mark. One specific dialogue, which also showed the most cooperative behavior of the interlocutors, is the only dialogue that holds all kinds of entrainment.

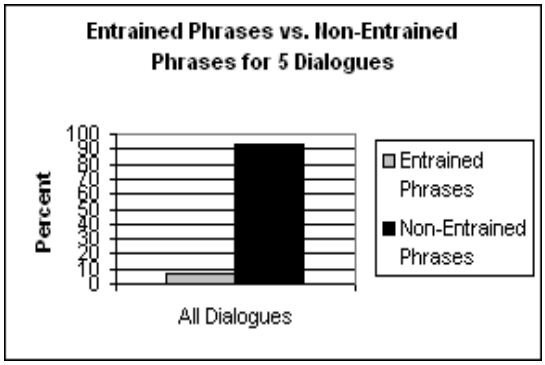

Figure 1: Entrained Phrases vs. Non-Entrained Phrases

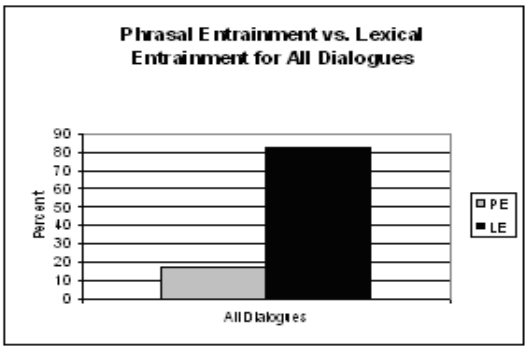

Figure 2: Phrasal Entrainment vs. Lexical Entrainment

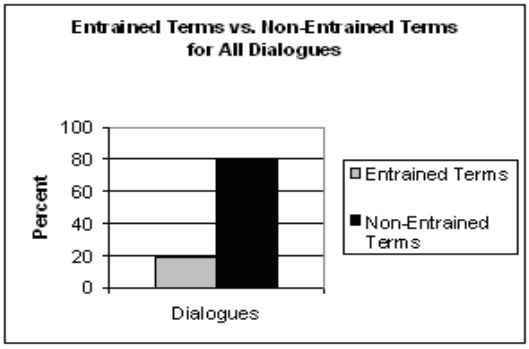

Figure 3: Entrained Terms vs. Non-Entrained Terms

As far as the interlocutors' roles as follower and leader are concerned, Figure 5 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, opera-

tors 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 [25].

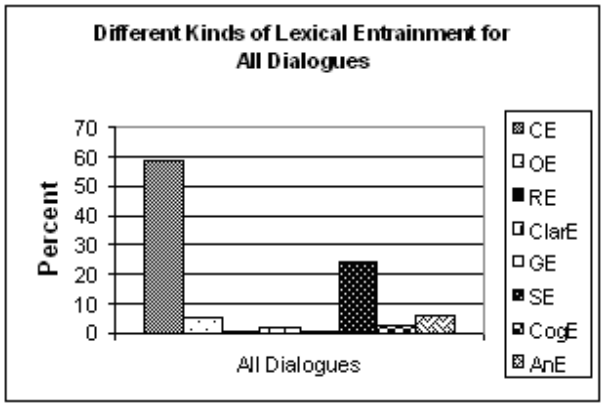

Figure 4: Different Kinds of Lexical Entrainment

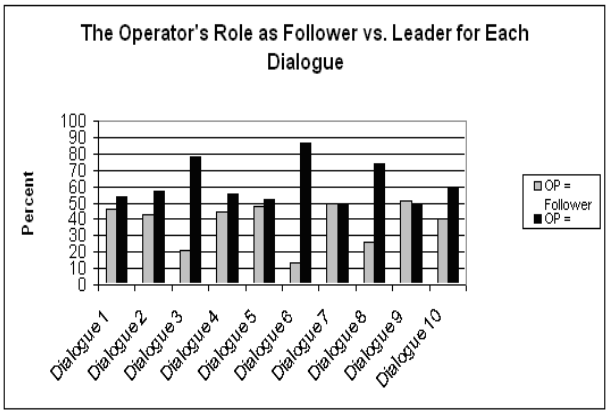

Figure 5: The Operator's Role as Follower vs. Leader

6. WIZARD OF OZ EXPERIMENT

Based on these and prior [24] 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.

6.1 Experimental Set-Up

In our experimental setup we created an entraining and a non-entraining *Wizard of Oz* system [14].

- 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 as in the MOS Study described in Section 3, i.e.:

Task 1: Assigning Channels to Stations

Task 2: Accessing the Internet

Task 3: Changing Mouse Speed

Task 4: Changing Font Size in Browser

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.

6.2 Experimental Results

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 6. 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 [28], the adaptive system - as one would expect - scored better in all respects.

Figure 7 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.

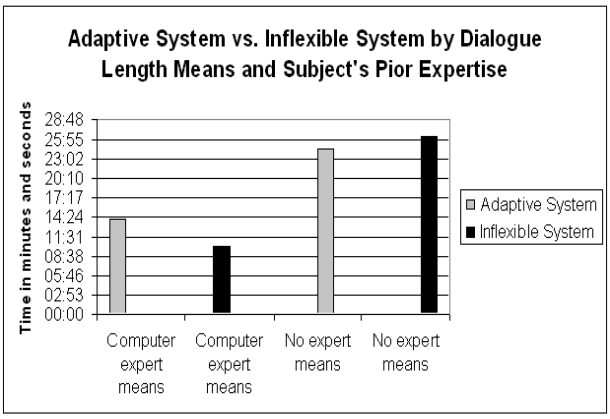

Figure 6: Task Completion Times

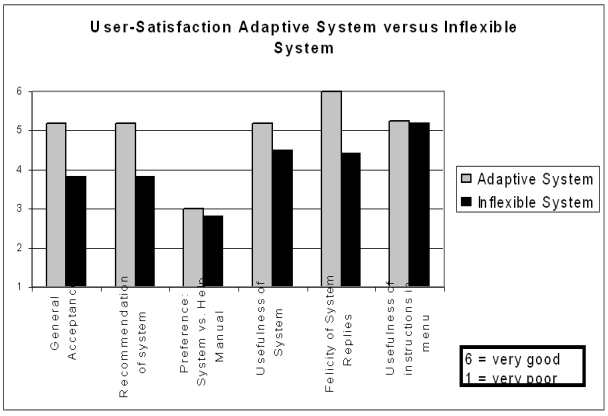

Figure 7: User Satisfaction

7. CONCLUSION

In this work we have shown in the Multiple and Operator and Subject study 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.

Our human-human data show that in nearly all of the dialogues the subjects followed the operator's choice of vocabulary slightly more often than the other way around. However, in crucial moments the operator did adopt the subject's choice of terms. For example, in one of the dialogues the subject appeared to be rather lost and was not able to figure out how to move the mouse across the screen when she tested the TV's Internet mode. When using the Internet mode, the mouse is moved by means of a trackpoint on the remote control or via the remote's keyboard. In several trials the most common denotation for the trackpoint was *little black round key* and *the black knob*. After several unsuccessful attempts to explain how to solve the task, the operator adopted the term introduced by the subject. The subject in this dialogue referred to the trackpoint as *nipple*. In fact this term is a suitable description for the trackpoint on the remote control, but not for the one on the keyboard. However, when the operator noticed that the subject had difficulties in finding the trackpoint on the keyboard, she used the term that had been introduced before. While the statistical analysis shows that the subjects adopted the terms introduced by the operator more often than vice versa, the example shows that in crucial moments the operator does adopt the subject's term. The crucial aspect of those moments is that the subject shows signs of insecurity and that task completion is at risk if a term is not adopted.

Task completion is the most crucial factor to multimodal systems. If users cannot even complete tasks, acceptance will be extremely low. Our findings also verify that acceptance and especially felicity rankings rise with the level of cooperation. All of this shows that entrainment is highly relevant for the development of multimodal systems and especially the natural language processing components. It is highly speculative how the concept of entrainment - or more general alignment and its effectiveness - can be transfered to other modalities such as gesture. Technologically first prototypes such as described in Kopp *et al.* [18] provide the capabilities perform analogous examinations. Furthermore, current finding and research on gesture clearly show its important role in interaction, from turn taking strategies to the transmission of meaning (e.g. [23]).

Our findings also suggests that a mere quantitative measuring of the occurrence of entrainment does not fully capture the importance thereof. As long as everything is going well, as was the case for the experts in all tasks and for the non-experts in the easier tasks, entrainment can be regarded as an optional feature that increases user satisfaction, but is by no means critical. This, however, changes drastically in situations where the operators sensed insecurity and consequently adopted their subject's terms. For a successful application of entrainment, we, therefore, need to endow multimodal systems with the capability to detect such critical moments. This can already be achieved by combining an ensemble of methods such as proposed by Walker *et al.* [27] and Forbes-Riley and Litman [13]. In our minds the stage is - at least - technically set to enable multimodal systems to implement strategies for adapting appropriately to their users by means of entrainment, which, as we have shown, is not always necessary or critical, but depends on the actual dialogical situation.

In future work will concern the implementation of entrainment capabilities for state of the art multimodal systems (e.g. [26]) and unimodal spoken interfaces. To achieve this goal systematic ways for handling the recognition of appropriate dialogical situations and a dynamic interface to the needed morpho-syntactic and semantic resources. Based on such prototypical implementations further studies can be performed to test and tune such adaptive interfaces to optimize their effectiveness as well as the satisfaction they provide to their users.

Acknowledgments

This work has been partially funded by the German Federal Ministry of Research and Technology (BMBF) as part of the SmartWeb project under Grant 01IMD01E and by the Klaus Tschira Foundation. The authors would like to thank Vanessa Micelli for her additional annotations and the reviewers for their valuable comments.

8. REFERENCES

[1] ALLEN, J. F., MILLER, B., RINGGER, E., AND SIKORSKI, T. A robust system for natural spoken dialogue. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (Santa Cruz, USA, 1996).

[2] ALTMAN, D. *Practical Statistics for Medical Research*. Oxford University Press, Oxford, 1990.

[3] BAILLY, G., CAMPBELL, N., AND MÖBIUS, B. Isca special session: Hot topics in speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology* (Geneva, Switzerland, 2003).

[4] BRENNAN, S. Lexical entrainment in spontaneous dialogue. In *Proceedings of the International Symposium on Spoken Dialogue* (Philadelphia, USA, 1996), pp. 41–44.

[5] BRENNAN, S. Processes that shape conversation and their implications for computational linguistics. In *Proceedings of ACL* (Hong Kong, 2000).

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

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

[8] CORE, M. G., AND MOORE, J. D. Robustness versus fidelity in natural language understanding. In *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding* (Boston, Massachusetts, USA, May 2 - May 7 2004), R. Porzel, Ed., Association for Computational Linguistics, pp. 1–8.

[9] COX, R., KAMM, C., RABINER, L., SCHROETER, J., AND WILPON, J. Speech and language processing for next-millenium communications services. *Proceedings of the IEEE 88*, 8 (2000).

[10] DARVES, C., AND OVIATT, S. 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).

[11] ELTING, C., ZWICKEL, J., AND MALAKA, R. 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.

[12] FELLBAUM, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass., 1998.

[13] FORBES-RILEY, K., AND LITMAN, D. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. In *Proceedings of 6th SIGdial Workshop on Discourse and Dialogue* (2005), Lisbon, Portugal.

[14] FRANCONY, J.-M., KUIJPERS, E., AND POLITY, Y. Towards a methodology for wizard of oz experiments. In *Third Conference on Applied Natural Language Processing* (Trento, Italy, March 1992).

[15] FRASER, N. Sublanguage, register and natural language interfaces. *Interacting with Computers 5* (1993).

[16] FURNAS, G., LANDAUER, T., AND DUMAIS:, G. The vocabulary problem in human-system-communication: an analysis and a solution. *Communications of the ACM 30(11)* (1987), 964–971.

[17] GARROD, S., AND ANDERSON, A. Saying what you mean in dialog: A study in conceptual and semantic co-ordination. *Cognition 27* (1987).

[18] KOPP, S., SOWA, T., AND WACHSMUTH, I. Imitation Games with an Artificial Agent: From Mimicking to Understanding Shape-Related Iconic Gestures. In *Gesture-Based Communication in Human-Computer Interaction, International Gesture Workshop, Genua, Italy April 2003* (2004), A. Camurri and G. Volpe, Eds., LNAI 2915, Springer, pp. 436–447.

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

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

[21] PAEK, T. Empirical methods for evaluating dialog systems. In *Proceeding of the 2nd SIGdial Workshop on Discourse and Dialogue* (Aalborg, Denmark, 2001), pp. 100–107.

[22] PARIS, C. L. *User Modeling in Text Generation*. Pinter, London, 1993.

[23] PARRILL, F., AND SWEETSER, E. What we mean by meaning: Conceptual integration in gesture analysis and transcription. *Gesture 4* (2004), 197–219.

[24] PORZEL, R., AND BAUDIS, M. The Tao of CHI: Towards effective human-computer interaction. In *HLT-NAACL 2004: Main Proceedings* (Boston, Massachusetts, USA, May 2 - May 7 2004), D. M. Susan Dumais and S. Roukos, Eds., Association for Computational Linguistics, pp. 209–216.

[25] SCHU, J. Formen der Elizitation und das Problem der Natürlichkeit von Gesprächen. In *Text- und Gesprchslinguistik. Ein internationales Handbuch zeitgenössischer Forschung* (2001), K. Brinker, G. Antos, W. Heinemann, and S. Sagere, Eds., Springer, pp. 1013–1021.

[26] WAHLSTER, W., REITHINGER, N., AND BLOCHER, A. Smartkom: Multimodal communication with a life-like character. In *Proceedings of the 7th European Conference on Speech Communication and Technology* (2001).

[27] WALKER, M., LANGKILDE, I., WRIGHT, J., GORIN, A., AND LITMAN, D. Learning to predict problematic situations in a spoken dialogue system: Experiments with how may i help you? In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)* (2000), Seattle, WA, pp. 210–217.

[28] WALKER, M. A., KAMM, C. A., AND LITMAN, D. J. Towards developing general model of usability with PARADISE. *Natural Language Engeneering 6* (2000).

[29] WOOFFITT, R., GILBERT, N., FRASER, N., AND MCGLASHAN, S. *Humans, Computers and Wizards: Conversation Analysis and Human (Simulated) Computer Interaction.* Brunner-Routledge, London, 1997.

[30] ZOEPPRITZ, M. Computer talk? Tech. rep., IBM Scientific Center Heidelberg Technical Report 85.05, 1985.