Modality Convergence in a Multimodal Dialogue System

Linda Bell[1], Johan Boye[2], Joakim Gustafson[1] and Mats Wirén[2]

[1]Centre for Speech Technology, KTH
Drottning Kristinas väg 31, S-100 44 Stockholm, Sweden
bell@speech.kth.se, jocke@speech.kth.se

[2]Telia Research
S-123 86 Farsta, Sweden
johan.boye@trab.se, mats.wiren@trab.se

Abstract

When designing multimodal dialogue systems allowing speech as well as graphical operations, it is important to understand not only how people make use of the different modalities in their utterances, but also how the system might influence a user's choice of modality by its own behavior. This paper describes an experiment in which subjects interacted with two versions of a simulated multimodal dialogue system. One version used predominantly graphical means when referring to specific objects; the other used predominantly verbal referential expressions. The purpose of the study was to find out what effect, if any, the system's referential strategy had on the user's behavior. The results provided limited support for the hypothesis that the system can influence users to adopt another modality for the purpose of referring.

1. Introduction

1.1. The problem

When participants in a dialogue refer to specific objects on successive occasions, they typically converge towards using the same terms in their referential expressions (Brennan and Clark 1996). Such *lexical convergence* in human–human interaction has a counterpart in human–computer interaction in the sense that human dialogue participants tend to adopt the terms of the system when referring to various concepts (Brennan 1996).

In this paper, we set out to investigate whether there is a more general form of convergence in human–computer interaction in multimodal dialogue systems. In the systems that will be of interest to us here, both the user and the system have the option of using either graphical operations or verbal expressions (or both) as they refer to specific objects in the dialogue. Given that users can choose to communicate by using speech or by using a pointing device to select objects on the screen, the question was to what extent they would be affected by the system's behavior as they constructed references.

1.2. Motivation

Apart from being a problem which is interesting in its own right, we believe that the results obtained from such an investigation will have important practical consequences for the design of multimodal human–computer dialogue systems. In order to create a system that performs well, it is crucial to have a good understanding of how the system should behave, so as to increase the chances of correctly interpreting the user's input. In particular, if we can find a systematic correspondence between the feedback strategy of the system on the one hand, and the user's choice of modality in her utterances on the other (i.e. what the user expresses in words and what she expresses by means of graphical operations), a lot can be gained. The present study is a step towards pursuing this goal.

1.2.1. Modality switching as an error handling strategy

Errors can occur on all levels of a dialogue system, but in domains where many of the words in the recognition lexicon are similar sounding, or where there is a large morphological overlap, the problem of recognition errors may become especially difficult. Experiments by Oviatt and VanGent (1996) have shown that there is a tendency for users to switch from one modality to another when their interaction with a multimodal system becomes problematic. In these semi-simulated experiments, users were subjected to errors which required them to repeat their input up to six times. Many users went from speech to graphical input after already having repeated and rephrased their spoken input to the system several times. It appears as if people use modality switching to recover from errors after having been subjected to a series of failures in communication by a noncooperative system.

It should be interesting to examine whether it is possible for a cooperative system to promote the use of one modality rather than another without explicitly asking the user to alternate or ceasing to 'understand' the user's input. Ultimately, the goal would be to design a multimodal system with the ability to predict and prevent the occurrence of longer error sequences. A low confidence score from the speech recognizer or an error indication from another part of the system could be used by the dialogue manager as a signal to encourage a user to switch to the graphical input mode. In this way, it would perhaps be possible to avoid a succession of errors and a resulting spiral of miscommunication.

1.3. The setting

This research has been carried out within the Adapt project, whose principal aim is to study various aspects of multimodal human-computer interaction in the context of an apartment-seeking domain. The practical goal of the project is to create a multimodal dialogue system which will help users find an apartment in the city of Stockholm.

The apartment domain is highly useful for studying multimodal interaction. An apartment is a complex object

that has properties suitable for graphical presentation (e.g. its location in the city), as well as properties suitable for verbal presentation (price, description of interior details, etc). Furthermore, it is not always obvious which modality is preferable for a referential construction.

For the purpose of the experiment described here, we use a simulation system where the key functionalities of the intended system are handled by a "wizard" (namely, analysis of multimodal user input, dialogue management and multimodal response generation).

2. Background

2.1. Lexical entrainment

In spontaneous human-human dialogue, participants frequently use referential expressions as a way of making the interaction efficient and concise. Clark and Wilkes-Gibbs (1986) have demonstrated that participants in a dialogue collaborate in the making of references. This collaborative effort is a sort of negotiation, where one of the interlocutors suggests a way of using a noun phrase to refer to a certain object, and the other accepts, rejects or postpones the decision. Once the participants have found a mutually acceptable way of referring to the object in question, they tend to use the term agreed on. Garrod and Anderson (1987) have established that people who repeatedly refer to the same objects in a dialogue often start using the same terms. They called this phenomenon *lexical entrainment*. Brennan and Clark (1996) have argued that lexical entrainment can be understood in terms of shared conceptualizations that are established between people engaged in conversation. After a conceptual pact has been established, speakers are sometimes overinformative in subsequent references instead of introducing a new term.

Brennan (1996) has argued that there is a phenomenon corresponding to lexical entrainment in human–computer interaction. Human dialogue participants tend to mimic the terms introduced by a spoken language system, something Brennan calls *lexical convergence*. Since computer programs generally are not constructed to negotiate about terminology, entrainment in Brennan and Clark's sense is not really possible in human–computer interaction. However, there appears to be a unidirectional influence by which the terminology of a natural language system is likely to influence the user's choice of vocabulary.

2.2. Multi-modal human-computer dialogue systems

Multimodal interfaces are potentially more flexible, powerful and effective than unimodal interfaces. Experiments in map-based simulation environments have demonstrated that a pen/voice interface can be more efficient and user-friendly than either a speech-only interface (Oviatt 1997) or a graphics-only interface (Cohen, Johnston et al. 1998). Studies of how users integrate the different input modes in multimodal dialogue systems have been previously reported in (Oviatt and Olsen 1994; Oviatt and VanGent 1996; Oviatt, DeAngeli et al. 1997). In a study where speech or pen input could be used to interact in a simulated map system (Oviatt, DeAngeli et al. 1997), it was demonstrated that people use the spoken and written modalities in a complementary way, rather than provide redundant information. Adaptable multimodal systems offer many possible advantages over unimodal interfaces, such as greater expressive power. However, if these systems are to become useful, we need to put greater efforts into studying how people use different modalities and alternate between them.

3. Method

3.1. Hypotheses

Our conjecture when embarking on this experiment was that when both system and user may choose the modality in which to construct a reference, the system will, to some extent, affect the user to enter into "modality convergence" with itself. More specifically, we were interested in testing two hypotheses with respect to modality convergence:

"Strong convergence": the user converges on the system's behavior while abandoning his previously adopted modality behavior.

"Weak convergence": the user converges on the system's behavior while retaining and integrating it with his previously adopted modality behavior.

Essentially, the weak hypothesis states that the system can "entrain" the user to adopt new behaviors. The strong hypothesis additionally states that the user can be retrained and made to abandon old behaviors.

We take it that it would be possible to achieve strong convergence if the system is suitably "uncooperative", for example, if it explicitly tells the user to switch modality or if it ceases to understand a certain behavior. However, rather than trying to affect the user by putting restrictions on the system's capabilities, we were interested in investigating to what extent a cooperative system could influence the user's behavior merely by changing its own way of constructing references.

The experimental task used to test these hypotheses involves the construction of deictic references to specific apartments on a map. Subjects who referred to apartments had the option of using either graphical or verbal means, or both. The question was then to what extent the subjects' construction of deictic (and other) references would be influenced by the behavior of the system.

3.2. Simulation system

The basic vehicle for the experiment was a Wizard-of-Oz simulation tool which provided information about available apartments in downtown Stockholm. The tool included a map showing names of streets, major neighborhoods, parks, etc., an overview map allowing the user to scroll the detailed map, and an animated agent speaking with a synthesized voice (see Figure 1). For each displayed icon, limited information about the corresponding individual apartment was provided in the row of a table. Here, the apartment's address, size and listed price were displayed. Icons on the map that represented apartments at adjacent or identical positions were only allowed to overlap to a limited extent in order to keep them simultaneously visible to the user.

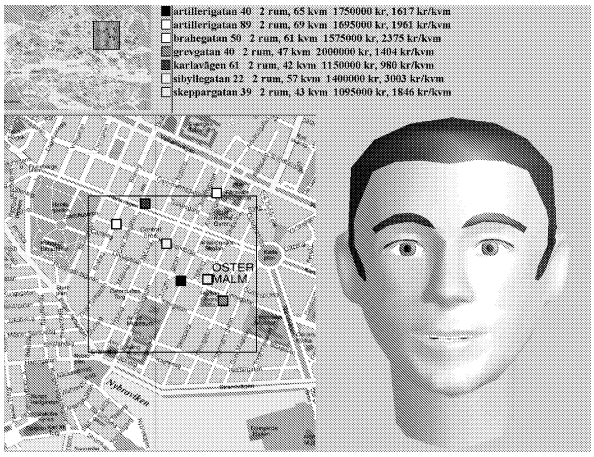

Figure 1. The graphical user interface.

The user's input was sent to the wizard interface where a human operator controlled the system's response. Much care was devoted to design the wizard interface to allow rapid system response times (typically between one and two seconds), thus giving users the impression of a fully functional system. The wizard chose his answer from a button menu, where information about specific apartments from the database was included in one of a number of possible answer templates.

To investigate convergence effects, the experiment focused on two equivalent ways of forming deictic references using different modalities, namely, graphics (point-and-click) and verbal expressions. To this end, the simulations mimicked two versions of a system, called System G ("graphics-oriented") and System S ("speech-oriented"), which behaved identically except for the way the deictic references were constructed. Thus, both versions used square-formed icons to indicate apartment positions on the map. The icons were color-coded so that each displayed icon had a unique color. The sole difference between the two simulated systems was that System G, while using a deictic utterance ("This apartment has a tiled stove"), let the corresponding icon on the map "shake" in a highly perceptible way for a fixed number of seconds (1.5, to be exact). In contrast, System S constructed apartment deictic references by using a verbal expression that exploited the color-coding ("The yellow apartment has a tiled stove"), but without shaking or otherwise changing the appearance of the icon in any way. Throughout the dialogues, the two systems retained their way of referring to the individual apartments.

Because of the difficulty of verbally distinguishing a large number of colors, and in order to help focus the dialogues on a limited number of objects which could be systematically compared, both of the simulated systems displayed at most seven apartment icons at any given time. Thus, as long as the current set of apartments to match the user's constraints was larger than seven, no icons were shown on the map. The animated agent would then prompt the user to narrow down the search by saying something like, "There are too many apartments to show. Are there any particular features you'd like your apartment to have?"

To make it straightforward for the user to associate table rows with the corresponding apartment icons, each row was preceded by a color-coded icon similar to the one on the map.

3.3. Experiment

To collect the data needed to test the hypothesis, a between-subjects design was selected. 16 participants were randomly assigned to a task/system sequence and each completed two tasks. For each task order (A-B, B-A), there was a corresponding system order (G-S, S-G), resulting in four unique sequences of two tasks (AG-BS, BG-AS, AS-BG, BS-AG), aimed at counterbalancing sequence effects. Each of these sequences was completed by eight persons, and a total of 32 dialogues were thus recorded.

Each task involved finding an apartment that fulfilled certain criteria. In solving the tasks, the subjects were invited to take their time looking around, and to contrast individual apartments in order to arrive at a suitable alternative. Before subjects started an experimental session, they were asked to try the functionalities of the system. In this way, the experimenter could make sure each user knew how to carry out the various operations.

As can be seen in Figure 2, task A and B both included a map of Stockholm where different areas had been shaded. These were the designated areas in which the users were to look for an apartment in their respective scenarios. In addition, the number of rooms the apartment should have and an approximate time period for the construction of the building were indicated on scales. Pictures of interior and exterior details were also added to each task. The subjects were informed that these details (stucco and a balcony, for instance) were merely suggestions, and that they were free to ask the system about other things that might interest them.

Subjects were instructed that they could communicate with the system using an open microphone and two graphical operations with respect to the map, namely, the selection of a position by point-and-click and the selection of a rectangular area of arbitrary size. The subjects' graphical operations were echoed in the same way by the two system versions; in particular, a point-and-click on an apartment icon was echoed by highlighting the icon. 16 subjects, all volunteers, participated in the experiment. Eight of the subjects were female and eight were male, and their ages ranged from 17 to 55. The subjects were all native speakers of Swedish, and while a few of them were staff at the Department of Speech, Music and Hearing,

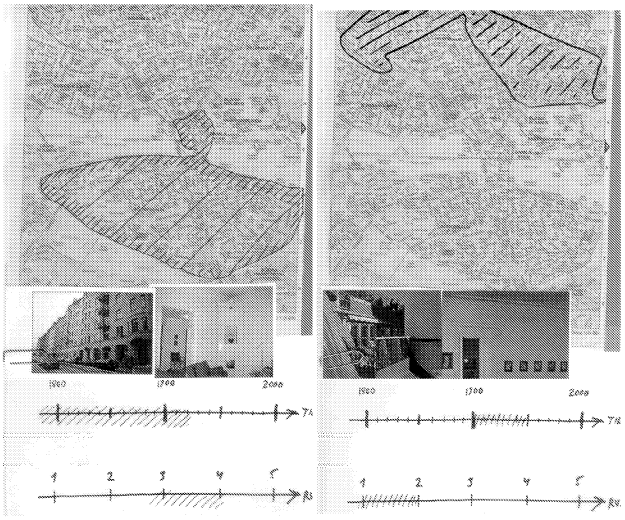

Figure 2. The scenarios, A on the left and B on the right.

none were working in the field of speech technology. All subjects reported to be familiar with computers, most of them regularly used word processing software and browsed the web, but only a couple claimed to have any significant programming skills. Each experiment session (including the introduction to the system and the post-experimental interview) lasted for approximately 30 minutes. During the post-experimental interviews, subjects were asked to give feedback on the interface and comment on their own modality choices during the dialogues. Finally, the experimenter verified that all subjects had been unaware of the fact that they had been interacting with a Wizard-of-Oz simulation rather than a real system. Those subjects who had been recruited from outside the department were rewarded with a movie voucher.

4. Data coding and analysis

4.1. Coding

Figure 3 shows an excerpt from a typical dialogue obtained during the experiment. A dialogue normally proceeds as follows: The subject begins by indicating a city area and providing a number of preferences in order for the system to come up with a set of matching apartments. The subject then asks the system about various features of the individual apartments displayed. This procedure may be iterated for a number of areas.

The dialogue excerpt shown in Figure 3 begins at the point where the system has just displayed a set of apartment icons on the map (with accompanying information in the table, as seen in Figure 1). The subject uses color references with a deictic function when he shifts focus from one apartment to another (that may or may not have been referred to previously in the dialogue).

…	
(User.44)	**Does the black apartment have a ceiling with stucco?** /subject uses mouse to select the black apartment/
(Sys.45)	The black one has stucco
(User.45)	Does the black one have a balcony?
(Sys.46)	Yes
(User.46)	Does the black one have a terrace?
(Sys.47)	I don't know
(User.47)	When was the building constructed?
(Sys.48)	The black house was built in 1928
(User.48)	When was it refurbished?
(Sys.49)	In 1994
(User.49)	Is it a quiet apartment?
(Sys.50)	I don't know
(User.50)	Which floor was that on?
(Sys.51)	The black apartment is on floor one
(User.51)	Facing the street or the back?
(Sys.52)	The street
(User.52)	**The red apartment which floor?**
(Sys 53)	The red apartment is on floor two
…	

Figure 3. A translated excerpt from one of the dialogues, using System S. Focus shifts are indicated using bold fonts.

After a focus shift, subjects typically use pronominal expressions to refer to the apartment under discussion. To a lesser extent, subjects continue to make use of constructions with a deictic function during turns subsequent to focus shifts. (Examples of this occur in Figure 3 at turns User.45 and User.46.) In these cases, however, they typically retained the particular reference construction used at the focus shift.

Our data analysis thus showed that the critical points at which modality convergence can be investigated correspond to focus shifts. The coding was therefore guided by the need to track user references to apartments made at these points. The references occurring at user turns other than focus shifts were not tagged.

References at focus shifts were tagged along two dimensions:

1. A category for each primitive type of reference construction used by the subject. We distinguished between four types:
   - g – graphical reference (that is, point-and-click);
   - c – color reference (for example, "the yellow one", "the black apartment");
   - a – address reference (a street name optionally followed by a street number, such as "Swedenborgsgatan 7");
   - m – miscellaneous (for example, "this one", "the apartment with a sauna").
2. A tag indicating whether the focus shift was initiated by the user or the system (user-init and system-init, respectively).

As an example of this, the dialogue excerpt shown in Figure 3 contains two focus shifts which are tagged as follows:

- (User.44) cg; user-first
- (User.52) c; user-first

The notation "cg" means that the subject used an integrated color and graphical reference by making a point-and-click operation in connection with a verbal utterance. Each reference categorized as "cg" was counted as one "c" and one "g", in addition to being counted as one "cg".

Focus shifts that occur initially in the dialogues, before the system has had any chance of entraining the subjects, have not been included in the count. However, we still coded them, since they could tell us something about the subjects' a priori preferences with respect to reference constructions at focus shifts.

4.2. Analysis

As previously stated, our main objective was to investigate if and how the subjects were influenced by the system in their way of referring to individual apartments. System G consistently referred to apartments using a graphical operation; system S consistently used color codes; hence we were primarily interested in the subjects'

behavior in this regard, which was reflected by the values of the "g" and "c" categories. Two "g" values were calculated for each subject, one for the number of "g" references in dialogue 1 and one for the number of "g" references in dialogue 2. Analogously, two "c" values were calculated for each subject.

In order to enable meaningful comparisons between subjects, we normalized each "g" value ("c" value) by dividing it with the total number of coded references in that dialogue. We used the notation "gNorm" ("cNorm") to refer to the normalized "g" values ("c" values).

As described in Section 3.3, the 16 test subjects were divided into four groups, each group corresponding to a unique sequence of scenario-system pairs (AG-BS, BG-AS, AS-BG, and BS-AG). The first test performed was to investigate whether the scenario had any significance for the behavior of the subjects. We therefore compared the values for the "gNorm" and "cNorm" parameters for the AG-BS group with those of the BG-AS group, and similarly for the AS-BG and BS-AG groups. As we found no significant differences, we collapsed the AG-BS and BG-AS groups into one group called G-S (corresponding to the eight subjects who used system G first and system S second). The AS-BG and BS-AG groups were collapsed into another group called S-G (corresponding to the eight subjects who used system S first and system G second).

The next step was to compare the values of the parameters "gNorm" and "cNorm" between and within the G-S and S-G groups. More specifically, we were interested in the relations indicated by the arrows in Table 1 below. The horizontal arrows in the table correspond to possible changes in referential behavior within the same group, but between the subject's first and second dialogue. The vertical arrows correspond to possible differences between the two groups either in the subjects' first dialogue, or in their second dialogue. In Table 1-4 below, "D1" and "D2" denote the first and second dialogue, respectively.

Table 2 shows the relations that should hold for the data to support the weak convergence hypothesis of Section 3.1. The value of the "gNorm" parameter should be higher for the G-S group than for the S-G group in dialogue 1, since at that point in time the S-G group had not yet been subjected to "graphical" behavior from the system. Similarly, within the S-G group, the "gNorm" value should be higher in dialogue 2 (when the system starts to behave "graphically") than in dialogue 1. An analogous line of reasoning gives the required relations indicated in the "cNorm" part of Table 2.

Table 3 shows the additional relations, apart from those of the weak convergence hypothesis, that should hold for the data to support the strong convergence hypothesis. The value of the "gNorm" parameter should be higher for the S-G group than for the G-S group in dialogue 2, since in the second dialogue the system behaved "graphically" towards the S-G group but not towards the G-S group. Similarly, within the G-S group, the "gNorm" value should be higher in dialogue 1 (when the system behaves "graphically") than in dialogue 2. An analogous line of reasoning gives the required relations indicated in the "cNorm" part of Table 3 below.

Table 1 Relevant data relations

	gNorm			cNorm		
	D 1		**D 2**	**D 1**		**D 2**
G-S	?	↔	?	?	↔	?
	↕		↕	↕		↕
S-G	?	↔	?	?	↔	?

Table 2. Relations that would support the weak convergence hypothesis

	gNorm			cNorm		
	D 1		**D 2**	**D 1**		**D 2**
G-S	?			?	<	?
	>			<		
S-G	?	<	?	?		

Table 3. Additional relations that would support the strong convergence hypothesis

	gNorm			cNorm		
	D 1		**D 2**	**D 1**		**D 2**
G-S	?	>	?			?
			<			>
S-G			?	?	>	?

5. Results

The most important results of the experiment are summarized in Table 4 below.

Table 4. Mean values for the "gNorm" and "cNorm" parameters

	gNorm		cNorm	
	D 1	**D 2**	**D 1**	**D 2**
G-S	0.16	0.32	0.04	0.48
S-G	0.04	0.03	0.13	0.21

If we begin by examining the four relations relevant for testing the weak convergence hypothesis (cf. Table 2), we see our data supports the hypothesis in three cases. Only the decrease from 0.04 to 0.03 for the S-G group's "gNorm" parameter is inconsistent with the hypothesis (but on the other hand, the total number of graphical references is indeed very small in those dialogues). The other three relevant relations are consistent with the weak hypothesis. However, only the increase from 0.04 to 0.48 for the G-S group's "cNorm" parameter proved to be statistically significant using a correlated t-test ($t(7)= -3.39$, $p<0.012$), as well as a Wilcoxon signed rank test ($W_+=1$, $p<0.028$). We therefore conclude that we have found limited support for the weak convergence hypothesis.

In contrast, the strong convergence hypothesis is not supported at all by the data. Of the four relations indicated in Table 3, only the difference between the two groups for the "cNorm" parameter for the second dialogue (0.48 vs. 0.21) is consistent with the strong hypothesis, however not significantly so. There is even an almost significant difference ($t(14)=2.12$, $p<0.053$) between the two groups for the "gNorm" parameter for the second dialogue (0.32 vs. 0.03), something which speaks against the strong hypothesis.

The strong hypothesis is also contradicted by the tendencies within the groups between the first and second dialogues. The group which started out using System S increased their proportion of color references in their second dialogue (from 0.13 to 0.21), even though the system had changed its behavior. The same tendency could be shown for the group that started out using System G (0.16 to 0.32), i.e. the subjects amplified the behavior adopted in their first dialogue rather than allowing themselves to be "retrained".

The group who started using System G had a higher proportion of graphical references during both dialogues when compared to the other group (almost significantly so in the second dialogue, as discussed above). This might be seen as a "delayed" convergence effect from their first dialogue. However, a closer look at the data reveals that out of the 17 graphical references by this group in the second dialogue, ten are integrated with color ("cg"). Thus, the increased use of graphics did not occur at the expense of color references, but rather "hand in hand" with these.

Looking at the subjects' behavior across the two dialogues, what we have said above might be summarized as follows: Rather than the subjects replacing one type of behavior with another as an effect of modality convergence, their "converging" behavior in the first dialogues was amplified in the second dialogues. In addition, they showed clear potential for taking up and integrating a new form of converging behavior in one of the second dialogues, namely, with the system that used color references.

Another way of formulating this is that the added proportions of color and graphical references increased from dialogue 1 to 2 for both groups. In other words, there was a tendency for subjects to gradually converge to the two kinds of reference construction that the system used, ("c" and "g") at the expense of the other kinds of reference construction ("m" and "a", mentioned in Section 4.1 above).

An interesting observation is that none of the subjects had color ("c") as their a priori preference in their first dialogue; still, color ended up being the altogether most used reference construction.

6. Discussion

Our post-experimental interviews indicated that the function of mouse clicks was not entirely obvious to the subjects. One subject said: "I preferred to speak since what I could do with the mouse seemed so limited" and another reported: "He [the animated agent] understood what I said, but not what I meant by clicking". The post-experimental interviews also revealed that the graphical input mode was perceived by several subjects as being less efficient and concise: "The question one asks with a mouse click seems rather undefined", "I preferred to speak, it was easy", "It was faster (I think) to speak directly to the animated agent."

Intuitively, it seems that in order for the system to maximize its chances of successfully entraining the user, the manifestatons of the input and output reference constructions should be as similar or "symmetric" as possible. In our experiment, such a symmetry was trivially achieved for verbal references (through the spoken manifestations of the user and system), but less so for graphical references: User clicks on apartment icons were echoed by highlighting the selected icon, whereas the system's graphical references were indicated by shaking the icon. Because of this, the connection between the graphical output and input might not have been obvious to the subjects. One way of clarifying this connection might be for the system to produce a characteristic short sound as each icon is highlighted. The same sound could then be repeated as the user clicks on one of the icons on the screen.

Furthermore, it is worth noting that the dialogues, generally speaking, were quite short. Since the tasks given were deliberately vague, some of the subjects chose to speak about no more than a couple of different apartments. A tendency in our data was that those subjects who persisted in interacting with the system for a longer time were more likely to be affected by the system's behavior. Longer dialogues would most certainly have given us more datapoints, and possibly also more clear-cut entrainment effects.

Acknowledgements

The authors would like to thank the other members of the Adapt group at the Centre for Speech Technology. We also thank Nils Dahlbäck for helpful methodological advice in the process of writing this paper. We are grateful to Bo Delling and Mattias Heldner for their invaluable assistance with the statistical computations.

7. References

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*: 41-44.

Brennan, S. E. and H. H. Clark (1996). Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition* **22**(6): 1482-1493.

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

Cohen, P. R., M. Johnston, et al. (1998). The efficiency of multimodal interaction: A case study. *Proceedings of the International Conference on Spoken Language*.

Garrod, S. and A. Anderson (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* **27**: 181-218.

Oviatt, S., A. DeAngeli and K. Kuhn. (1997). Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. *Proceedings of CHI '97*. Atlanta: 415-422.

Oviatt, S. and E. Olsen (1994). Integration Themes in Multimodal Human-Computer Interaction. *Proceedings of the International Conference on Spoken Language Processing*. Volume 2: 551-554.

Oviatt, S. and R. VanGent (1996). Error resolution during multimodal human-computer interaction. *Proceedings of the International Conference on Spoken Language Processing*: 204-207.

Oviatt, S. L. (1997). Multimodal interactive maps: Designing for Human Performance. *Human Computer Interaction* **12**: 93-129.