# COORDINATION OF REFERRING EXPRESSIONS IN MULTIMODAL HUMAN-COMPUTER DIALOGUE

*Gabriel Skantze*

Centre for Speech Technology
Department of Speech, Music and Hearing, KTH
gabriel@speech.kth.se

## ABSTRACT

This study examines coordination of referring expressions in multimodal human-computer dialogue, i.e. to what extent users' choices of referring expressions are affected by the referring expressions that the system is designed to use. An experiment was conducted, using a semi-automatic multimodal dialogue system for apartment seeking. The user and the system could refer to areas and apartments on an interactive map by means of speech and pointing gestures. Results indicate that the referring expressions of the system have great influence on the user's choice of referring expressions, both in terms of modality and linguistic content. From this follows a number of implications for the design of multimodal dialogue systems.

## 1. INTRODUCTION

One of the great advantages with advanced dialogue systems – the possibility for the user to express herself in a "natural" way – is also a source of potential problems: the large variations in the system input [1]. Multimodal interaction could reduce this problem by giving the user means to replace complex verbal expressions with non-verbal expressions, especially in a spatial domain [2]. The number of misunderstandings and disfluences could thereby be reduced [3].

It is not certain, however, that the user will always choose expressions that are the easiest for the system to handle. It should therefore be interesting to identify some of the factors underlying the user's verbal behaviour. Several studies have shown that the system's expressions will affect the user's expressions [4, 5]. Coordination of language use is a natural part of human-human communication, and this must be considered when designing a human-computer dialogue system. In this study, an experiment was conducted on a multimodal dialogue system to see if the users' choice of referring expressions would coordinate with the system output. One version of the system used deictic expressions for referring, and one used definite descriptions.

### 1.1. The system used

The experiment was conducted on a system called ADAPT – a multimodal dialogue system for apartment seeking in Stockholm, developed at KTH [6]. At the time the experiment was conducted, no input analysis or dialogue management was implemented in the system, so a semi-automatic, Wizard-of-Oz setup was used, with an operator (the "Wizard") controlling the missing parts of the system. The users believed they interacted with a real system. This also allowed "perfect" speech recognition.

The user interface in ADAPT consists of an interactive map, presenting the apartments, and on which the user can point to formulate deictic referring expressions. There is also an animated talking head (using speech synthesis), and a table where some information about the apartments is presented. The user can interact with a mouse and by speech input.

### 1.2. Motivation and hypothesis

Knowledge about how the system's output will affect the user's behaviour, i.e. the system's input, could be used as a guideline to choose the system behaviour that will give the best performance. Results from this study could be applied in this way. Furthermore, such knowledge could be used to dynamically change the user's behaviour in order to avoid errors when problematic situations are coming up. This study will, however, not give answers to whether this is possible. It should be the subject for further studies.

The hypothesis in this study is that users of a system with a consistent set of referring expressions, will adapt to those expressions, both in references to apartments and areas; two conceptually very different types of objects. Apartments are well-defined objects represented as boxes on the map. Areas are more ill defined objects, and there is less mutual knowledge between the speakers about the meaning (i.e. extension) of such referents.

## 2. BACKGROUND

### 2.1. Coordination of referring expressions

As Clark and Wilkes-Gibbs [7] have demonstrated, referring is a collaborative process between speaker and hearer. In constructing the referring expression, the speaker tries to get the hearer to identify the object that he has in mind. But since the speaker and the hearer will inevitably have somewhat different beliefs about the world, the hearer might not be able to identify the object from this description. Therefore, the speaker and hearer will engage in a negotiation in order to get a mutual understanding of the referent.

To make the conversation efficient, the speakers coordinate their use and interpretation of referring expressions. In order to reduce the amount of negotiation, the speakers will strive to establish a common ground. This common ground, built up between speakers, will not only consist of common beliefs, but also of common means of referring to those beliefs, resulting in a coordination of referring expressions. As shown by Garrod

and Andersson [8] in experiments on human-human conversation, speakers build conceptual pacts during the dialogue, which are specific for the discourse shared between them.

### 2.2. Modelling the user's expressions

The tendency for humans to coordinate their expressions with the other participant in a conversation could be utilized in human-computer dialogue systems in order to reduce the vocabulary. Zoltan-Ford [5] tested the hypothesis that the user's vocabulary and phrase length could be shaped, by manipulating the computer's output. The results showed that the users, both during spoken and written interaction, adopted the phrase length (but not the vocabulary) of the system's expressions. The phenomenon was called *verbal shaping*.

Brennan [4] showed in an experiment on a natural language dialogue system that also the vocabulary could be modelled (calling it *lexical convergence*). For some of the user's referring expressions, the system corrected the user by using the expression that was to be modelled in an implicit or explicit confirmation. These methods were called *embedded* and *exposed* modeling. It is the embedded modeling that is used in this study (see Figure 1 for an example).

An experiment similar to this has previously been conducted on ADAPT [9], concerning the user's choice of modality in referring to apartments. The study gave some support for the convergence of the use of descriptions between user and system. In this study, we have looked at references to both apartments and areas. We have also looked at both modality and linguistic content, analyzing both local and global discourse. Quantitative measures of the users' experiences have also been taken into the analysis.

## 3. METHOD

### 3.1. The two versions of the system

For the experiment, the system was modified to behave in two different ways, resulting in one version that used definite descriptions in referring to apartments and areas (from now on called the *description* version), and one that used deictic expressions (called the *deictic* version). The only difference between the two versions was their means of referring; otherwise they were identical. The apartments were represented as colored boxes on the map. To make pointing gestures, the system could blink with an apartment box or highlight an area on the map. Definite descriptions consisted of the apartment's color and/or the name of the area. Figure 1 (a) shows a possible dialogue with the description version, and (b) the same dialogue with the deictic version.

(a) U: Are there any apartments around Karlaplan?
S: There are five apartments around Karlaplan. [highlights Karlaplan, shows five apartments]
U: What does this apartment cost? [points at an apartment]
S: The red apartment at Karlaplan costs 3750000 crowns.

(b) U: Are there any apartments around Karlaplan?
S: There are five apartments in this area. [highlights Karlaplan, shows five apartments]
U: What does this apartment cost? [points at an apartment]
S: This apartment costs 3750000 crowns. [the apartment blinks]

*Figure 1:* An example of a possible dialogue with the two versions of the system.

### 3.2. Experimental design

16 subjects, 6 male and 10 female, participated in the experiment. All subjects had at least some basic knowledge about the geography of Stockholm and limited experience of telephone based speech recognition applications. None of them had used ADAPT, or any other multimodal dialogue system, before. A between-subjects design was used; the subjects were divided in two groups, balancing gender and experiences. One group used the description version of the system and the other group used the deictic version.

The subjects were given some initial oral instructions on how to use the system, but no detailed instructions on how they should use the different modalities.

Because the purpose of this study was to see how the user's expressions were affected by the system's expressions, it was very important not to let the task descriptions influence the users verbally. Therefore pictorial scenarios were used. The user's task was to find two apartments that should match some visual requirements presented on the screen; see figure 2.

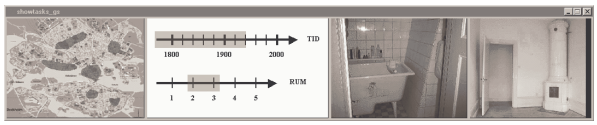


*Figure 2*: The scenario presented to the subjects, showing location, age, rooms, a bathtub and a tiled stove.

In order to make the conversations longer and to get more data, the whole scenario was not presented at once; the last two pictures were added after 10 and 20 minutes respectively. Another purpose of this method was to promote the users to go back and ask about apartments that had previously been mentioned, by using global anaphora.

To distract the subjects as little as possible and create a natural setting, no videcameras were used and the subject was alone in the room during the experiment.

After the experiment, the subjects were given a questionnaire about their experience of the system. This was done before they were told about the Wizard-of-Oz setup.

### 3.3. Data analysis

The recorded dialogues were transcribed and all referring expressions were tagged for statistical analysis. Only referring expressions that shifted focus from one object to another, or looked like they could shift focus, were counted. For example, descriptions like "the red apartment" or expressions like "this apartment" (including a mouse-click) were counted in the analysis, since they could be used to shift focus, but expressions like "it" (without a mouse-click) were not counted.

Every referring expression was tagged with the referring features that it contained: color description, other description, deictic term (e.g. "this", "that") and/or pointing gesture. Note

that one referring expression may contain more than one feature, for example both a color description and a pointing gesture. The references were then classified based on the referent (apartment or area). The number of features for each user was counted and divided with the user's total number of references (that could shift focus), to calculate a percentage. An average value for all subjects in each group, based on these percentages, was then calculated for every feature.

# 4. RESULTS

The 16 dialogues contained a total of 512 referring expressions, of which 300 were references to areas.

## 4.1. References to apartments

The mean percentage of features in the references to apartments is shown in Figure 3. A star (*) denotes a significant difference between the groups (t=3.3, 5.2, 3.6, 2.3 respectively; dF=14; p<0.05; one-tailed t-tests).

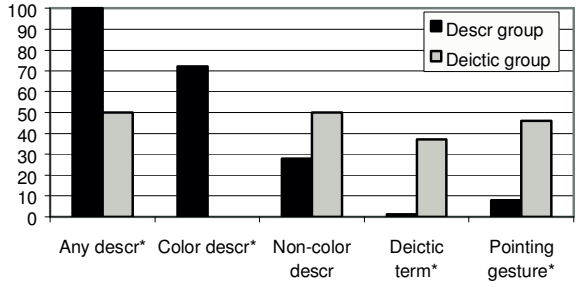


*Figure 3:* Mean percentage of features in the referring expressions to apartments.

The figure shows that the users in the two groups did adapt to the system's way of referring to apartments. All users using the description version of the system used some sort of description in their referring expressions, compared to only 50% in the deictic group. Pointing gestures were much more common in the deictic group. Colors were common in the description group, but never occurred in the deictic group.

### *4.1.1. Alternation of referring expressions*

The users in the deictic group used deictic expressions in 50% of their referring expressions, but that doesn't mean that they alternated between different types of expressions; instead different users used different types consequently. Users that mostly used descriptions only deviated from that behaviour on the average 4.6% of the times. For users that preferred deictic expressions the average deviation was about 8%.

This means that some users mostly used the same type of expressions as the system did (thus adapting to the system), and some users did not adapt at all.

### *4.1.2. Global anaphora*

There were only 8 global anaphoric references to apartments (i.e. references to apartments that were not currently under discussion, but had been so at a previous stage). Colors were never used in such references, only street addresses. Unlike color descriptions and deictic expressions, street addresses give apartments a globally unique label. There was a significant correlation between the number of local references by street address, and the total number of global anaphora (in general) for each user (r=0.44; t(14)=1.86, p<0.0.5), suggesting that referring by addresses helps the user to go back in the dialogue and compare apartments that are currently under discussion with those previously mentioned.

## 4.2. References to areas

The mean percentage of features in the references to areas is shown in Figure 4. There were no significant differences between the groups.

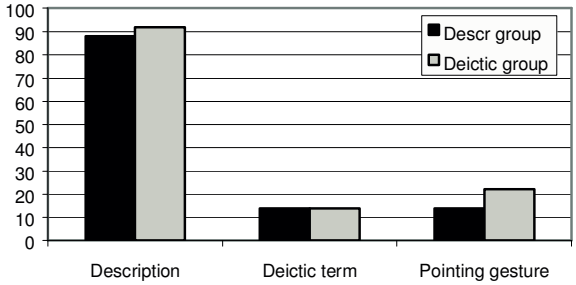


*Figure 4:* Mean percentage of features in the referring expressions to areas.

This shows that deictic references were very rare in both groups, and that the system's way of referring to areas didn't affect the users at all.

## 4.3. Subjective measures

The results from the final survey are presented in Figure 5. A scale from 1 to 7 was used, where 1 means "doesn't agree at all" and 7 means "fully agree". There were no significant differences between the groups.

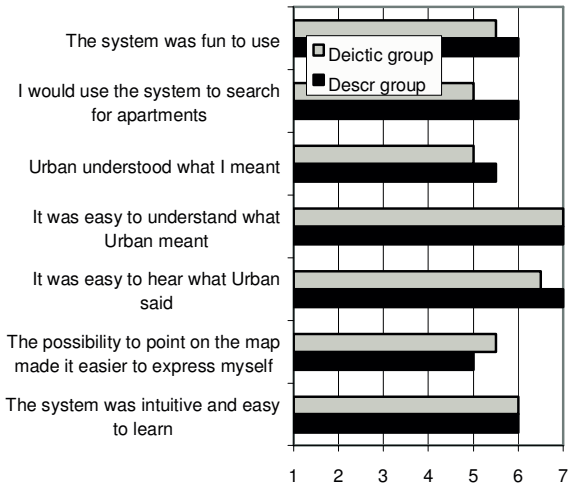


*Figure 5:* Median of the users' ratings in the two groups, after using the system.

# 5. CONCLUSIONS & DISCUSSION

Users adopt the system's means of referring to apartments, but not to areas. This may be due to the fact that the apartment

icons more looked like traditional "clickable" objects. If the user could draw areas on the map, the use of the pointing gestures for areas may also be greater. The value of deictic references to areas may also be greater for users with less geographical knowledge.

Referring by means of definite descriptions, rather than by deictic expressions, is more effective as a way of affecting the user's expressions. The adaptation to the description version was stronger than that to the deictic version. A possible explanation may be that the system's pointing gesture (blinking) doesn't resemble that of the user's (pointing with the mouse).

Some users adapt to the system's means of expressions, some don't. Previous studies [2] have also shown that user's of multimodal systems prefer one modality to another, instead of alternating between them. This may make it hard to dynamically model a user's expressions.

The user is more inclined to use verbal expressions that she has heard or seen being used by the system. Colors were only used if the computer had referred to apartments using color names; deictic terms were almost never used if the computer hadn't done so. Both groups used street addresses, and those were also printed in the table. When users referred by descriptions of areas, these consisted almost exclusively of names printed on the map. This suggests that the user may not be sure about the system's capabilities, and that the user relies on the (seen or heard) verbal output of the system. One could relate this to Donald Norman's notion of "affordances" in design, i.e. the functions that the user of an artifact perceives are supported, and that should be apparent in the user interface [10].

Users that use globally unique expressions to refer to an object will be able to use global anaphoric references, and will also do so to a greater extent. This is important to note when choosing which expressions the system will use, in order to give the user means of comparing objects in a global context. The description version used area and color in referring to apartments (e.g. "The red apartment at Karlaplan"), which could be used as a globally unique label. No users adopted theses expressions for global references, instead they used the apartment's street addresses.

There were no differences between the users' experiences of the two systems, despite their different behaviors. The ratings seem to be relatively high, which may have given some ceiling effects.

#### *5.1.1. Future work and application of the results*

Referring expressions must be used to at least implicitly confirm the user's utterance and ensure common ground. As shown here and elsewhere, how these expressions are chosen will have an effect on the input the system will get back. The user will not automatically choose the expressions that will be easiest for the system to handle. Of the options tested here, referring to apartments with colors seems to give the most consistent input (with 72% of the utterances containing colors). If the users had adapted to a greater extent to the deictic version, this mode may have been preferable; deictic expressions should be less error-prone. This study does not suggest, however, that this is possible.

Neither colors nor deictic expressions seem to provide means for global references, and this must be considered as a weakness. Although street addresses are error-prone, this way of referring may for this reason be preferable. This approach was not, however, explored in this study.

If we want users to use verbal expressions in referring, these should be provided, visually or spoken (as "affordances"). The dialogue system should also be able to handle all verbal expressions that the system agent uses, or that are printed somewhere in the graphical user interface.

The results do not tell us if a user's behaviour could be dynamically modelled. This could be a subject for future studies. As Oviatt and VanGent [11] have shown, users tend to switch modality after repeated system errors. To speed up this adaptation, the system could change the means of referring when grounding the user's utterance, if it detects a possible error (i.e. from the speech recognition confidence scores).

## 6. ACKNOWLEDGEMENTS

This research was carried out at the Centre for Speech Technology, a competence center at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The author would like to thank the other members of the ADAPT group, and Arne Jönsson, Dept. of Computer and Information Science, Linköping University, for helpful comments.

## 7. REFERENCES

[1] Furnas, G. W., Landauer, T. K., Gomez, L. M. & Dumais, S. T. (1987). The Vocabulary Problem in Human-System Communication. In *Communications of the ACM*, **30**, 964-971.

[2] Oviatt, S. L. (1999). Ten Myths of Multimodal Interaction. In *Communications of the ACM*, **42**(11), 74-81.

[3] Cohen, P. R., Johnston, M., McGee, D. Oviatt, S. L., Clow, J. & Smith, I. (1998). The Efficiency of Multimodal Interaction: A Case Study. In *Proceedings of ICSLP*, **2**, 249-252.

[4] Brennan, S. E. (1996). Lexical Entrainment in Spontaneous Dialog. In *Proceedings of ISSD*, 41-44.

[5] Zoltan-Ford, E. (1991). How to get People to Say and Type what Computers can Understand. *International Journal of Man-Machine Studies*, **34**(4), 527-547.

[6] Gustafson, J., Bell, L., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén, M. (2000) AdApt – a multimodal conversational dialogue system in an apartment domain. In *Proceedings of ICSLP*, **2**, 134-137.

[7] Clark, H. H. & Wilkes-Gibbs, D. (1992). Referring as a Collaborative Process. In H. H. Clark (Ed.), *Arenas of Language Use*, 107-143. Chicago: University Press.

[8] Garrod, S. & Andersson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition*, **27**, 181-218.

[9] Bell, L., Boye, J, Gustafson, J & Wirén, M. (2000) Modality Convergence in a Multimodal Dialogue System. *Proceedings of Götalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*, 29-34.

[10] Norman, D (1988). *The Design of Everyday Things*. London: MIT Press.

[11] Oviatt, S. L. & VanGent, R. (1996). Error Resolution during Multimodal Human-Computer Interaction. In *Proceedings of ICSLP*, **1**, 204-207.