Active and Passive Gestures - Problems with the Resolution of Deictic and Elliptic Expressions in a Multimodal System

Michael Streit
Siemens AG c/o German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3
66123 Saarbrücken
streit@dfki.uni-sb.de

Abstract

This paper deals with aspects of the resolution of deictic and elliptic expressions that are related to gestures. It discusses different approaches to distinguish between deictic pointing and manipulative gestures. We compare two strategies of combining natural multimodal communication with direct manipulation. The first approach uses click free mouse gestures for deictic pointing, while manipulative gestures are performed by using mouse button events as is usual in graphic interfaces. The second approach uses a touchscreen as gestural input device.

1 Introduction

This paper deals with aspects of the resolution of deictic and elliptic expressions that are related to gestures. It discusses two approaches to distinguish between deictic pointing and manipulative gestures. The resolution methods for both approaches have been implemented in a project for the development of a multimodal interface for route-planning, traffic information and driver guidance MOFA (German acronym for "Multimodale Fahrerinformation"- "Multimodal Driver Information"). [1] We proved the reusability of the methods and the architecture of MOFA in a completely different domain by implementing a prototype for multimodal calendar management (the system TALKY).

The input modalities supported by our system are spoken and written natural language, deictic pointing gestures and the interaction methods known from direct manipulation systems. As input devices for gestures we use either a mouse or touch screen technology. The latter allows the user to perform deictical pointing in an (almost) natural way.

Our approach to multimodal systems aims at a smooth integration of a conversational communication style with the features of direct manipulation. By conversational style of multimodal communication we understand the use of natural speech supported by deictical gestures, as in the combination of the utterance "an *dem* Tag geschäftlich in Kaiserslautern", ("at *this* day in Kaiserslautern for business") with a pointing act to a graphical presentation of the day under consideration. In this example the gesture acts only passively, supplying the spoken utterance with additional information. In contrast gestures used in direct manipulation are active, they are expected to trigger some action by the system, e.g. clicking at an icon which represents "1st of May" opens a representation of this day. Verbal utterances that accompany or follow such gestures may be related to this gesture:

- If in this context the user utters "um 12 Uhr in Raum 17" ("at 12 o'clock in room 17") it is most likely that he specifies an appointment at the 1st of May. In this case the elliptic verbal utterance can be resolved by the manipulative gesture.
- But if the user utters "at 2nd of May meeting with Mr. X" he performs an unrelated tasks.

2 Overview of the System Architecture of MOFA

In our architecture the MOFA interface is organized into four main components that are realized as independent processes that communicate via sockets:

- The Graphical Interface (GI) for gestural input and graphical output,
- the Speech Recognizer (SR),

[1] For a description of MOFA see [Streit 96].

- the multimodal interpretation, dialogue and presentation planning process (MIDP)
- and the speech synthesis component (SYN).

The back end application (AP) may be integrated in the MIDP or realized as separate process (in case of MOFA, the AP is the route planner). This organization allows for parallel input and output and also an overlapp of user input, interpretation and presentation.

The GI sends both gestural events like pointing to an object or selecting a menu item, and written input to the MIDP. It also realizes graphical presentations, which are planned by the MIDP. The speech recognizer is activated by the user via a speech-button which can be operated from the GI. We do this to avoid problems with recognition results that are not intended by the user (see [Gauvain 96]). Both the results of the speech recognition and the input processed by the GI are augmented with time information.

The MIDP consists of the following components:

1. The input controller checks if modalities are used in parallel or subsequently and determines temporal relations between events.
2. The interpretation module is responsible for robust parsing and plan recognition. Parsing and plan recognition work in close interaction and produce a task oriented representation. The interpretation depends on the dialogue state as specified by the clarification or the continuation module. The interpretation module also handles gestures that the input controler recognizes as monomodal input.
3. Deictic and elliptic expressions are resolved in the resolution module.
4. A plan library and a simple domain model support plan recognition and resolution.
5. The clarification module checks in a case based manner the partly instantiated task hypotheses. It first tries silently to repair certain inconsistencies and then starts a clarification dialogue or calls the back end application. The clarification module has an interface to the user model which proposes fillers for parameters which are not specified by the user.
6. The continuation module decides if the system keeps the initiative or if it simply waits for further user input. The modules (5) and (6) also control speech recognition. Depending on the dialog state they select the appropriate language model.
7. The presentation module plans the presentation of the result of an application call. It controls the output of SYN and GI.
8. The user model follows a memory based learning approach. The user model is an optional module, since it can be left out from the system without any changes to modules.

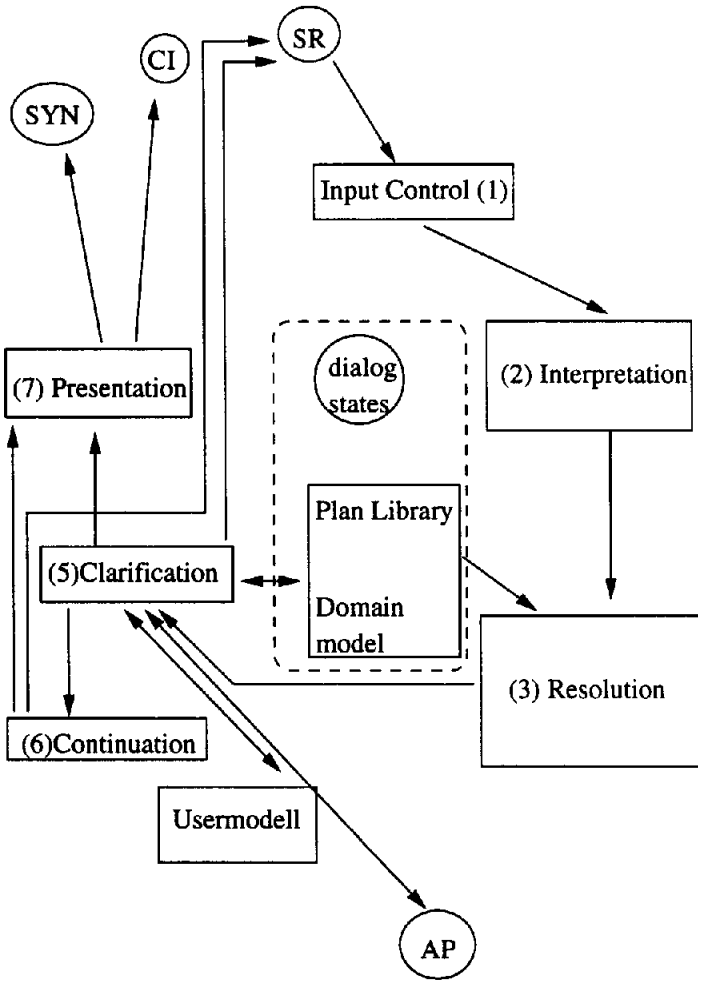

Figure 1: Architecture

3 Two implemented Applications: Sample Interactions with MOFA and TALKY

In MOFA, the interaction style is mainly conversational, menus are only used for commands on the meta level (e.g starting a trip simualtion along a planned route, or introducing some external events

into the trip simulation (e.g. invoking a traffic jam). TALKY is much more a mixture between manipulative interaction and multimodal conversation. If the system is running on a SUN Ultra Sparc, speech recognition is real time, if the user speaks slowly.

In the following we give three examples which are processed by our systems MOFA and TALKY. In these examples, natural language is spoken, unless otherwise stated. System reactions are given in spoken and written form. In our sample dialogs '↗' indicates a pointing gesture, '↗(Y)' means that the gesture is unambiguously pointing to some object, where Y refers to an object on the screen.

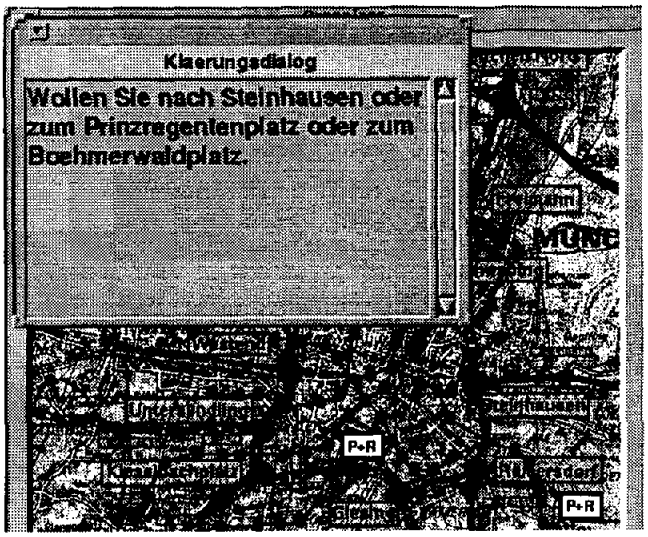

Figure 2: MOFA

Example 1: Route Planning

- USER: "Von hier" ("From here") ↗(A)
- MOFA: "Sie wollen nach A. Bitte zeigen Sie mir Ihr Ziel." ("You want to start in A. Please show me your destination.")
- MOFA: highlights A.
- USER: "Karte nach Giesing" ("move map to giesing")
- MOFA: moves map.
- USER: " Zeige mir diesen Platz" ("show me this place") ↗ (X)
- MOFA: shows a picture
- USER: okay
- MOFA: removes the picture
- USER: "Dorthin" ("to this location") ↗ "Über diese Kreuzungen" ("via these crossings") ↗(B), ↗(C)
- MOFA: "Wollen Sie nach Steinhausen oder zum Prinzregentenplatz oder zum Boehmerwaldplatz ?" (Is your destination Steinhausen or Prinzregentenplatz or Boehmerwaldplatz ?")
- USER: "Boehmerwaldplatz" (Boehmerwaldplatz is a tube-station)
- MOFA: "Von A über B nach C habe ich die gelbe Route gefunden. Von B aus führe ich Sie zu dem Park-und-Ride Platz U. Für diese Route gibt es einen Detailplan. Sie haben dort die folgenden U-Bahnverbindungen ... (etc) " ("I found the yellow route from A via B to C. From B I lead you to the park-and-ride parking U. That route is shown on a more detailed map. There, you have the following tube connection ... (etc)") (this information is only given as text)

Example 2: Route Planning

- USER: "Ab hier ↗ (A) zum Kiesselbachplatz" ("from here to Kiesselbachplatz")
- MOFA: ...presents route ...
- USER: "dorthin" ("to this location") ↗ (B)
- MOFA: "Wollen Sie eine neue Route definieren?" ("Do you want to define a new route?")
- (MOFA does not know, if the user wants to perform a modification task referring to the route previously defined, or if he wants to define a new route)
- USER: "Nein" ("No")
- MOFA: ... presents route from A via Kiesselbachplatz to B ...

Example 3: Appointment Scheduling

- USER: ↗ *1st of May button*"
- USER: "Morgen von viertel vor vier bis sechs mit Maier" (Tomorrow at a quarter to four until six with Maier")
- TALKY: presents the appointment in a structured graphic dialog box with an okay button (cf Figure 3). The box contains default information, that is proposed by the user model.

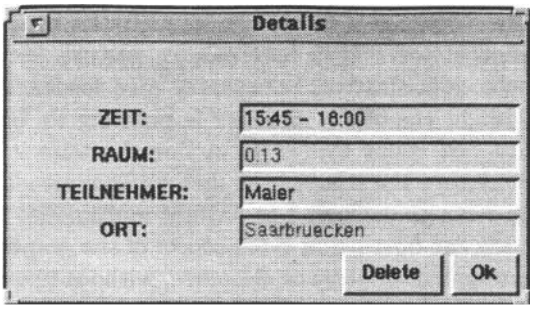

Figure 3: TALKY

- USER: "Im Konferenzraum, bis sieben Uhr" ("in the conference room, until seven o'clock")
- TALKY: Adds the information that the Meeting Room is *Konferenzraum* and sets a new end time (cf. Figure 4). The continuation module propses as follw up task the information of the participants of the meeting. To avoid a clarification dialog, the system assumes, as long the user has not confirmed a proposal, he will still further modify the appointment.
- USER: ↗ (okay button)
- TALKY: removes Dialog box
- USER: "von zwei bis drei" ("from two to three")
- TALKY: presents a new appointment presentation box

Figure 4: TALKY

4 Problems with the integration of direct manipulation and natural multimodal dialog

In direct manipulation gestures lead to an immediate reaction. The referent of a gesture is always unambiguous: Either there is a single object selected or the gesture is not successful at all. In this process of selection only the gesture and the objects are involved. In natural communication deictic gestures may be vague or ambiguous. They have to be interpreted by considering context and the natural language utterances that occur together with the gesture. The possibility of modifying a pointing gesture by spoken information is not a weakness but a valuable feature multimodal communication, that makes it easier to refer to structured objects or to closely assembled tiny objects.

For multimodal systems we are faced with the problem that speech recognition and natural language analysis takes some time, which is somewhat contrary to the immediate reaction expected from direct manipulative gestures. The problem cannot be completely solved by making analysis faster because the user may want to perform some manipulation and see the result while he is speaking a longer utterance.

If we could distinguish the manipulative or deictic nature of the gesture by analysing it's form and the object referred to we could avoid waiting for linguistic analysis. In the following we will discuss some approaches for a solution of this problem.

5 Active and Passive gestures

Without the support of other modalities, an active gesture determines an action of the system (in case of a graphical user interface) or it causes the dialog partner to perform an action (in case of natural communication). A passive or "merely referential" pointing gesture serves only as a reference to objects and does not call for any reaction, aside from recognizing the act of reference. The gesture may be ambiguous or vague. Passive gestures are not always communicative in nature (e.g someone may point at a map to support his own perception without any intention to communicate an act of reference).

If a gesture is active depends on the form of the gesture (e.g. moving the mouse to some object is a passive form, pressing a mouse button at an object is an active form), but also on the object, which being referred to with the gesture. E.G. a mouse click performed on a menu item will start an action, while clicking on a picture may be without any result. We will now give a short definition of passive and ac-

tive gesture forms and of passive and active objects. Then we analyse possible combinations.

Passive gesture forms are not used to trigger actions, they may be used non-communicatively. Active forms are always intended communication, they may be used to trigger an action. Passive objects serve only as potential referents of referential acts, while active objects react if they are activated by an active gesture form without the support by other modalities. There may be mixed objects as well, that behave actively using certain active gesture forms and passively with others. With passive gesture forms every object behaves passive by definition.

There are six possible cases for a combination of objects with gesture forms:

1. Passive gesture forms performed on passive object
2. Passive gesture forms performed on mixed object
3. Passive gesture forms performed on active objects
4. Active gesture forms performed on passive object
5. Active gesture forms performed on mixed object
6. Active gesture forms performed on active objects

We consider cases (1) to (4) as passive gestures, while (6) is considered an active one. If (5) is active or passive depends on the concrete gesture form. Passive gestures are candidates for conversationaly used deictic gestures, while manipulative gestures must be active. In the following, we will discuss two approaches to distinguish between deictic and manipulative uses of gestures.

5.1 Distinction between Deictic and Manipulative Gestures by Active and Passive Gesture Forms

To allow for a coexistence of natural communication style and direct manipulative interaction in one multimodal interface we dedicate

- (1),(2) and (3) to natural communication
- and (4),(5) and (6) to graphical interaction ((4) could be seen as a communication failure or as an attempt to perform natural communication by manipulative gestures).

This results in a clear cut between the communication styles: GUIs take passive gesture forms as non-communicative. They may give feedback and highlight the object the user is pointing to, but we must not count this as an action, because it does not change the state of the dialog. This means that the gestures dedicated to natural multimodal interaction can operate on every object of the graphically represented universe of discourse, without unwanted manipulative effects. Another advantage of this approach is, that the user can keep with the graphic interaction style, he is familiar with from graphical interfaces. There is no conflict with the selection process (i.e. the direct manipulative principle of reference resolution). We can introduce an additional process, which combines ambiguous or vague information from deictic pointing with natural language information. Furthermore, passive pointing gestures with the mouse are much more convenient than active ones if they are performed in parallel with speech. We noticed that the coordination of mouse clicks with deictic expressions requires high concentration on the side of the user and frequently leads to errors. This is quite different with touchscreens. We followed the approach, presented in this section, in an earlier version of MOFA (cf. section 6 *Experience with Active and Passive Gesture Forms - MOFA with the Mouse as Input Device*). We observed two problems with that approach.

- It may be difficult to decide between communicative and non-communicative uses of passive gesture forms.
- If we use other input devices than the mouse, the distinction between passive and active gesture forms may be not available , or only be achievable by an artificial introduction of new gestures.

5.2 Distinction between Deictic and Manipulative Gestures by different Active Gesture Forms

If we make all graphically represented objects mixed or passive we arrive again at a clear cut between styles, by distinguishing between certain types of active gesture forms. The advantage of this approach is, that there is no problem with non-communicative uses of gestures. But with this approach we have to change the usual meaning of gestures and the usual behaviour of objects, that are not mixed (e.g. menu items or buttons are active objects). Gestures will also tend to become more complicated (in some cases we need double clicks instead of simple clicks to activate an action).

If we do not change the normal behaviour of objects, we stay with objects for which we cannot decide if a gesture is meant deicticly or manipulatively. In particular, this means that graphical selection may prevent pointing gestures from being modified by speech.

There is another small problem with active gesture forms. The user may expect that using them will cause some action even with passive objects, if the context or the nature of the objects suggests how to interpret such gestures.

We will elaborate on these problems in sections 8.1 *Deictic Expressions and Pointing Gestures* and 8.3 *Deictic Pointing Gestures as Autonumuos Communicative Acts.*

6 Experience with Active and Passive Gesture Forms - MOFA with the Mouse as Input Device

This version is implemented with mouse-based gestural communication. We used active gesture forms to achieve manipulative effects and passive ones for deictic pointing. Because the mouse has to move across the screen to point to a certain referent, it is very likely that objects are touched without intention. This is especially important for route descriptions, where several pointing acts occur during one spoken command. Different filters are used to take out unintended referents.

- First, the search for referents is restricted to a time frame which is a bit longer than the interval within the user is speaking.
- Next, type restrictions are applied to the possible referents. Type restrictions are inferred from deictic expressions e.g. "diese Strasse" - " this street","diese U-Bahnstation" - "this tube station", but also from inherent restrictions concerning the recognized task(s).
- Finally, we exploit the fact that deictic expressions and deictic gestures are strictly synchronized in natural speech. The problem with this approach is that speech recognizers usually do not deliver time information on the word level. Therefore time stamps on the word level are interpolations. There is only a rudimentary analysis of the track and the temporal course of the mouse movement. Such an analysis would certainly improve referent resolution, though we noticed that pointing was sometimes not marked and on the other hand, users made pauses during mouse movement without pointing intentionally.

In this MOFA version we can only use linguistic information to identify a task. The identification of referents by gesture analysis alone is to uncertain to use the type of the referents for task recognition. This is different in the recent touchscreen based MOFA version, which we will describe in the following.

7 MOFA and TALKY - the version for touchscreen input

The recent versions of MOFA and TALKY are implemented with touchscreen technology as input device. The version also works with a mouse as input device, but in this case the user must perform mouse clicks for deictic pointing.

With touchscreens, every pointing to the screen is normally mapped to mouse button events. There are no passive gestures at all. The distinction of deictic gestures must rely on active gesture forms. Furthermore, the problem of vague or ambiguous pointing becomes more prominent: In the usual implementation of touchscreens, pointing with the finger will be mapped to some exact coordination, but these coordinates are only vaguely related to the point the user wanted to refer to. A big advantage of touchscreen technology is that there is no problem with unintended pointing. Although there is additional vagueness in pointing, we can use type information that we get from referents much easier than with passive mouse gestures. Also, active pointing at the touchscreen is completely natural in combination with speech.

8 Reference Phenomena in MOFA and TALKY

The system handles temporal and spatial deixis, and also certain phenomena at the borderline of deictic and anaphoric reference. It is able to treat elliptic expressions, in which gestures supply missing arguments (deictical ellipsis). The system resolves elliptic expressions and anaphora that occur in in modification and follow up tasks. Also dialog ellipsis and elliptic utterances that introduce new tasks are handled. Many expressions including temporal deixis are not resolved by deictical gestures, but by computation of values, depending on the speaking time (e.g. heute (today)). Because of the graphic representation of time objects especially in the calendar application, there are also deictic gestures referring to temporal entities. Deictic expressions occur as demonstrative NPs (e.g. diese Kreuzung (this crossing)) definite NPs (die Route (the route)) and also as adverbs (dorthin (there), dann (then), heute (to-

day)). The NPs and some of the adverbs are also used anaphoricly.

- In MOFA the universe of discourse is the map. The objects on the map are not of the active type. The interaction with the map is not manipulative, but conversational. There are also some menu items, to which the user will hardly refer deicticly.
- In TALKY there are many active objects. The user may navigate in the calendar by speech or by manipulating graphical interaction elements. In contrast to MOFA there are manipulative objects, that are likely to be referred to also by deictic gestures.

8.1 Deictic Expressions and Pointing Gestures

Reference resolution for gestures as usual in graphical user interfaces works in an unambiguously way. To handle vague pointing we introduce transparent fields that constitute neighbourhoods for clusters of objects. The selection of these fields or of one object on such a fields makes the neighbouring objects salient as referents. (cf. section 3 *Two implemented Applications: Sample Interactions with MOFA and TALKY*, example 1 Route Planning). Now type information is used as a filter. In example 1 every object that is a possible starting point for a car route and also every tube-stations is of appropriate type. If there remains more than one referent, or there is no referent left, a clarification dialog is invoked, the dialog must not be performed by natural language only, zooming on the cluster is also a possible reaction.

- (1) " zu dieser U-Bahnstation ("to that tube-station") ↗

In (1) referent resolution applies the type restriction *tube-node* before any clarification dialog is invoked.

- (2) "dorthin" ("to there") ↗(U1) "von dort"(from there") ↗(U2)

In (2) the system first recognizes an abstract task route planning. If the two referents of the gestures are tube stations the system does that know by an unambiguous gesture or after a clarification dialog. The system will use these type constraints to recognize the concrete task "find a tube connection".

8.2 Elliptic Expression and Pointing Gestures

- (3) mit der U-Bahn (by tube) ↗,↗

(3) is an example for an elliptic expression related to deictic gestures. The gestures supply additional arguments to the task "plan a tube connection". In account of the order of the arguments, MOFA will guess the first referent is the start, the second is the destination of the route. The task is identified by analyzing the elliptic expression. The Task now imposes type restriction on the arguments.

8.3 Deictic Pointing Gestures as Autonumuos Communicative Acts

We recall the fact, that we use active gesture forms as deictic (i.e. passive) gestures in the touchscreen version of MOFA. As mentioned in section 5.2 this may give the user the idea to use them actively to communicate without speech. It is very natural to order a ticket just by e.g. saying "Saarbrücken München" with the first town as starting point and the second town as destination. Similar one can describe a route on a map by pointing first to the start and then to the destination. In contrast, if the user is pointing only once, it is most likely, that he means the destination of a route. Such pointing acts communicate the same informations as speaking the names of the locations. This way to communicate is a sort of natural communication, that does not fit to direct manipulation. The interpretation of the first pointing depends on the fact if there is a second one, which is not the way as direct manipulation works. We handle these *natural gestural* communication by the following steps.

- The input control checks if the speech channel is active.
- If speech is not active it waits a short time until timeout.
- If there is a second gesture before timeout it waits again a short time.
- Otherwise the interpretation module is called for monomodal gesture interpretation.
- The interpretation module proceeds this input like a sequence of names, perhaps after a clarification dialog to resolve ambiguous reference.

8.4 Are Temporal Relations necessary for the Resolution of Deictic Expressions

[Huls 1996] proposes instead of analysing temporal relation to solve the problem of parallel gestures by incrementally parsing and resolving deictic expressions. We think that approach will not work with spoken input and in certain cases will also not work for text input.

1. With spoken input the deictic expressions contained in the utterance can be analysed in general only after all deictic gestures are performed, because speech recognizers do not provide incremental input. Therefore the temporal order has to be accounted for.

2. A deictic gesture may be refer to an object, that is not of appropriate type by an error of the user. From the temporal synchronization of deictic expressions and deictic gestures, we infer, that this gesture was intend to communicate the referent of the deictic expression. But if we apply anaphora resolution methods, it is very likely that we exclude by type considerations the referent of the gesture from the set of possible referents of the deictic phrase. Perhaps we may instead find some other referent, from which we could now by temporal consideration, that it is not a possible referent.

3. If deictic gestures are used as in section 8.3 *Deictic Pointing Gestures as Autonumuos Communicative Acts*, it is obvious that we need the temporal order of the gestural events for interpretation. This argument applies also to elliptic utterances as (4).

- (4) "mit der U-Bahn" ("by tube") ↗,↗

In (4) we must now the order of the gestures to distribute the referent in the right order to the arguments the route planning task.

9 Open Question

As mentioned in section 8 *Reference Phenomena in MOFA and TALKY* some active objects in TALKY could appropriately be used for deictic reference. If the user opens a day-sheet in the calendar by manipulation and defines an appointment, without specifying a day, by speech, TALKY will schedule the appointment to the day, opened by the user. The effect is the same as with an deictical reference to a passive representation of the day under consideration.

- (5) "Urlaub von dann ↗ bis dann. ↗" ("Holidays from then to then")

In (5) it is doubtful, if the user really wants to open the two day-sheets he is referring to. Is there a solution for this problem, when you have a touchscreen as input device? Does the problem mean, that we must use or invent other gestural input technics?

References

C. Huls, E. Bos, W. Claassen, "Automatic Referent Resolution of Deictic and Anaphoric Expressions", Computational Linguistics, 1996.

J.L. Gauvain, J.J. Gangolf, L. Lamel, "Speech Recognition for an Information Kiosk", Proc. ICSLP 96, Philadelphia, 1996.

M. Streit, A. Krueger, "Eine agentenorientierte Architektur fuer multimediale Benutzerschnittstellen", Online 96 - Congressband VI, Hamburg, 1996.