In: J. W. Sullivan and S. W. Tyler (eds.): Intelligent User Interfaces, New York : ACM Press , pp. 45-67.

USER AND DISCOURSE MODELS FOR MULTIMODAL COMMUNICATION

WOLFGANG WAHLSTER

Computer Science Department
German Research Center for Artificial Intelligence
University des Saarlandes[*]

ABSTRACT

In face-to-face conversation humans frequently use deictic gestures parallel to verbal descriptions for referent identification. Such a multimodal form of communication is of great importance for intelligent interfaces, because it simplifies and speeds up reference to objects in a visual context. Natural pointing behavior is very flexible, but possibly ambiguous or vague, so that without a careful analysis of the discourse context of a gesture there would be a high risk of reference failure. The subject of this paper is how the user and discourse models of an intelligent interface influence the comprehension and, production of natural language with coordinated pointing, and conversely how multimodal communication influences the user model and the discourse model.

After a brief description the deixis analyzer of our XTRA system, which handles a variety of tactile gestures, including different granularities, inexact pointing gestures, and pars-pro-toto deixis, we present some empirical results of an experiment that investigates the similarities and differences between natural pointing in face-to-face communication and simulated pointing using our system. This paper focuses on consequences of this investigation for our present work on an extended version of the deixis analyzer and a gesture generator currently under development. We show how gestures can be used to shift focus and how focus can be used to disambiguate gestures. Finally, we discuss how the user model affects the decision of the presentation planning component to use a pointing gesture, a verbal description, or both, for referent identification.

3.1 INTRODUCTION

In face-to-face conversation humans frequently use *deictic gestures* (e.g., the index finger points at something) parallel to verbal descriptions for referent identification. Such a *multimodal* form of communication can improve human interaction with machines, because it simplifies and speeds up reference to objects in a visual world.

The basic technical prerequisites for the integration of pointing and natural language are fulfilled by high-resolution bit-mapped displays and window systems for the presentation of visual information; various pointing devices such as mouse, light-pen, joystick, and touch-sensitive screens for deictic input; and the DataGlove™[Zimmerman87] or even image sequence analysis systems for gesture recognition. But the remaining problem for artificial intelligence is that explicit meanings must be given to natural pointing behavior in terms of a formal semantics of the visual world.

Unlike the usual semantics of mouse clicks in direct manipulation environments, in human conversation the region at which the user points is not necessarily identical with the region to which he or she intends to refer. Following the terminology of Clark, we call the region at which the user points, the *demonstratum*; the descriptive part of the accompanying noun phrase, the *descriptor* (which is optional); and the region to which he

[*] This research was partially supported by the German Science Foundation (DFG) in its Special Collaborative Programme on AI and Knowledge-Based Systems (SFB 314).

or she intends to refer, the *referent* [Clark83]. In conventional systems there exists a simple one-to-one mapping of a demonstratum onto a referent, and the reference resolution process does not depend on the situational context. Moreover, the user is not able to control the granularity of a pointing gesture, since the size of the predefined mouse-sensitive region specifies the granularity.

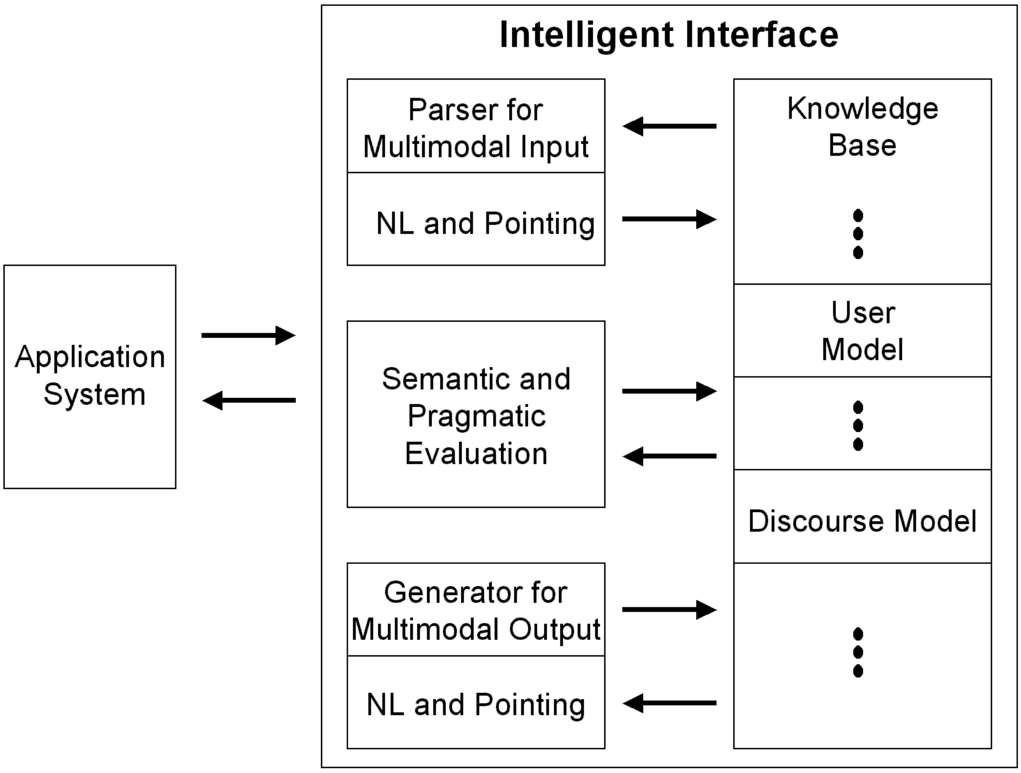
Compared to that, natural pointing behavior is much more flexible, but also possibly ambiguous or vague. Without a careful analysis of the *discourse context* of a gesture there would be a high risk of reference failure, as a deictic operation does not cause visual feedback from the referent (e.g., inverse video or blinking as in direct manipulation systems).

The subject of this paper is how the user and discourse models of an intelligent interface influence the comprehension and production of natural language with coordinated pointing to objects on a graphics display and conversely how multimodal communication influences the user and discourse models.

Figure 3.1 outlines the basic architecture of an intelligent interface with multimodal input and output. For the sake of clarity, we have omitted all aspects of processing components and knowledge base that are not relevant to the topic of this paper.

Before we review previous research on the combination of natural language and pointing and describe some current approaches related to our work, let us briefly introduce the basic concepts of user and discourse models.

FIGURE 3.1
THE BASIC ARCHITECTURE OF AN INTELLIGENT MULTIMODAL INTERFACE



3.2 USER MODELS AND DISCOURSE MODELS

A reason for the current emphasis on user and discourse models [Wahlster86, Kobsa89] is the fact that such models are necessary prerequisites in order for a system to be capable of exhibiting a wide range of intelligent and cooperative dialogue behavior. Such models are required for identifying the objects to which the dialogue partner is referring, for analyzing a nonliteral meaning and/or indirect speech acts, and for determining what effects a planned utterance will have on the dialogue partner. A cooperative system [Wahlster84] must certainly take into account the user's goals, plans, and prior knowledge about the domain of discourse, as well as misconceptions the user may possibly have concerning the domain.

We use the following definitions [Wahlster88]:

- A *user model* is a knowledge source containing explicit assumptions about all aspects of the user that may be relevant to the dialogue behavior of the system.

- A *user modeling component* is that part of a dialogue system that performs the following functions:

    1. To incrementally build up a user model
    2. To store, update, and delete entries in it
    3. To maintain the consistency of the model
    4. To supply other components of the system with assumptions about the user

- A *discourse model* is a knowledge source that contains the system's description of the syntax, semantics, and pragmatics of a dialogue as it proceeds.

- A *discourse modeling component* is that part of a dialogue system that performs the following functions:

    1. To incrementally build up a discourse model
    2. To store and update entries in it
    3. To supply other components of the system with information about the structure and content of previous segments of the dialogue

It seems commonly agreed upon that a discourse model should contain a syntactic and semantic description of discourse segments, a record of the discourse entities mentioned, the attentional structure of the dialogue including a focus space stack, anaphoric links, and descriptions of individual utterances on the speech act level. However, there seem to be many other ingredients needed for a good discourse representation that have not yet been determined in current discourse theory.

An important difference between a discourse model and a user model is that entries in the user model must often be explicitly deleted or updated, whereas in the discourse model entries are never deleted (except for phenomena related to forgetting). Thus according to our definition above, a belief revision component is an important part of a user modeling component.

This does not imply that the discourse model is static and only the user model is dynamic. The discourse model is also highly dynamic (consider, e.g., focus shifting), but it lacks the notion of logical consistency that is important for belief revision and default reasoning in a user modeling component. The discourse model is like an annotated trace of the various levels of the system's processing involved in understanding the user's utterances and generating its own dialogue contributions.

3.3 RELATED WORK ON DEICTIC INPUT

Although in an intelligent multimodal interface the "common visual world" of the user and the system could be any graphics or image, most of the projects that combine pointing and natural language focus on business forms or geographic maps.

To the best of our knowledge, Carbonell's work on SCHOLAR represents the first attempt to combine natural language and pointing in an intelligent interface [Carbonell70]. SCHOLAR, a tutoring system for geography, allowed simple pointing gestures on maps displayed on the terminal screen. NLG [Brown79] also combined natural language and pointing using a touch screen to specify graphics with inputs like (1).

(1) Put a point called Al here <touch>.

Woods and his co-workers developed an ATN editor and browser that can be controlled by natural-language commands and accompanying pointing gestures at the networks displayed on the screen [Woods79].

In SDMS [Bolt80] the user can create and manipulate geometric objects by natural language and coordinated pointing gestures. The first commercially available multimodal interface combining verbal and nonverbal input was NLMenu [Thompson86], where the mouse could be used to rubber band an area on a map in sentences like (2).

(2) Find restaurants, which are located here <pointing> and serve Mexican food.

All approaches to gestural input mentioned so far in our brief review were based on a simple one-to-one mapping of the demonstratum onto a referent and thus have not attacked the central problems of analyzing pointing gestures.

Recently, several research groups have more thoroughly addressed the problems of combining nonverbal and verbal behavior. Several theoretical studies and empirical investigations about the combination of natural language and pointing have been published [Hayes86, Hinrichs87, Reilly85]. Working prototype systems have been described, which explore the use of complex pointing behavior in intelligent interfaces.

For example, the TACTILUS subcomponent[1] of our XTRA system [Kobsa86], which we will describe below in more detail, handles a variety of tactile gestures, including different granularities, inexact pointing gestures, and pars-pro-toto deixis. In the final case, the user points at an embedded region when actually intending to refer to a superordinated region.

In the DIS-QUE system [Wetze187] the user can mix pointing and natural language to refer to student enrollment forms or maps. The deictic interpreter of the T3 system [Scragg87] interacts with a natural language interpreter for the analysis of pointing gestures indicating ship positions on maps. In addition it can utilize continuing or repeated deictic input. CUBRICON [Neal9l] is yet another system that simultaneously handles input in natural language and pointing to icons on maps, using language to disambiguate pointing and using pointing to disambiguate language.

While the simultaneous utilization of both verbal and nonverbal channels provides maximum efficiency, most of the current prototypes do not use truly parallel input techniques, since they combine typed natural language and pointing. In these systems the user's hands move frequently back and forth from the keyboard to the pointing device. Note, however, that multimodal input makes even natural-language interfaces without speech input more acceptable (fewer keystrokes) and that the research on typed language forms the basis for the ultimate speech-understanding system.

3.4 A CLASSIFICATION OF TACTILE POINTING GESTURES

For a study of the semantics and pragmatics of pointing, it is important to distinguish between two types of gestures:

- Pointing at *graphic models of objects* in the domain of discourse (e.g., geographic maps, icons for an office environment). In this case, the *detailed structure* of an icon is not relevant for the interpretation of a pointing gesture. For example, pointing at the lid of the trash icon causes the same effect as pointing at the can.

- Pointing at *objects* of a visual domain (e.g., forms, texts, graphics, formulas, images). In principle, *every pixel* on the screen can be a *separate reference object* in this case. For example, in our XTRA system (see Section 3.5 following) gestures can refer to all parts of the tax forms.

Today, most multimodal interfaces combining natural language and pointing belong to the first category (see Section 3.3 preceding). In this case, the interpretation and generation of pointing gestures are much easier than in the second category.

On the other hand, many of the subtleties of natural pointing come into their own only in the second case (see also [Schmauks87]). Moreover, that category covers a much wider range of possible applications.

Another fundamental distinction, which is independent from the classification introduced above, is whether the system deals with a static or a dynamic visual domain:

[1] In 1984 in the proposal for the XTRA project, I described the basic architecture of a flexible multimodal interface with a gesture analysis component. Since 1985, we have been working on the integration of pointing and natural language. The current version of TACTILUS was designed and implemented by J. Allgayer.

- Pointing at *fixed* and *static* visual objects on the screen (e.g., an icon for an airport on a map, a region of a tax form). In this case pointing gestures refer to *directions*, *locations*, or *objects*.

- Pointing at *dynamic* and *animated* visual objects on the screen (e.g., an animated ship icon on a map, a moving car in an image sequence). The pointing gestures can refer to events (e.g., "This U-turn was not allowed").

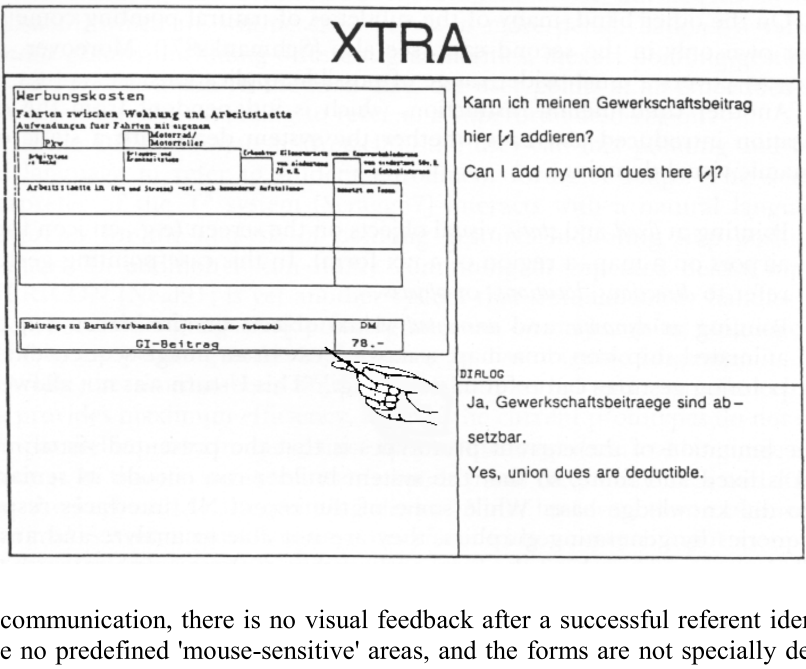
One limitation of the current prototypes is that the presented visual material is fixed and finite, so that the system builder can encode its semantics into the knowledge base. While some of the recent NL interfaces respond to queries by generating graphics, they are not able to analyze and answer follow-up questions about the form and content of these graphics, since they do not have an appropriate representation of its syntax and semantics. Here one of the challenging problems is the *automatic formalization of synthetic visual information* as a basis for the interpretation of gestural input.

3.5 XTRA: AN INTELLIGENT MULTIMODAL INTERFACE TO EXPERT SYSTEMS

XTRA (eXpert TRAnslator) is an intelligent multimodal interface to expert systems that combines natural language, graphics, and pointing for input and output. As its name suggests, XTRA is viewed as an intelligent agent, namely a translator that acts as an intermediary between the user and the expert system. XTRA's task is to translate from the high-bandwidth communication with the user into the narrow input/output channel of the interfaces provided by most of the current expert systems.

The present implementation of XTRA provides natural language access to an expert system, which assists the user in filling out a tax form. During the dialog, the relevant page of the tax form is displayed on one window of the screen, so that the user can refer to regions of the form by tactile gestures. As shown in Figure 3.2, there are two other windows on the right part of the display, which contain the natural language input of the user (upper part) and the system's response (lower part). An important aspect of the communicative situation realized in XTRA is that the user and the system share a common visual field - the tax form.

FIGURE 3.2
THE COMBINATION OF NATURAL LANGUAGE, GRAPHICS, AND POINTING IN XTRA



XTRA

Werbungskosten

Fahrten zwischen Wohnung und Arbeitsstaette

Aufwendungen fuer Fahrten mit eigenem

Pkw

Motorrad/
Motorroller

Arbeitsstaette in

Beitraege zu Berufsverbaenden

GI-Beitrag

DM

78.-

Kann ich meinen Gewerkschaftsbeitrag
hier [↗] addieren?
Can I add my union dues here [↗]?

DIALOG

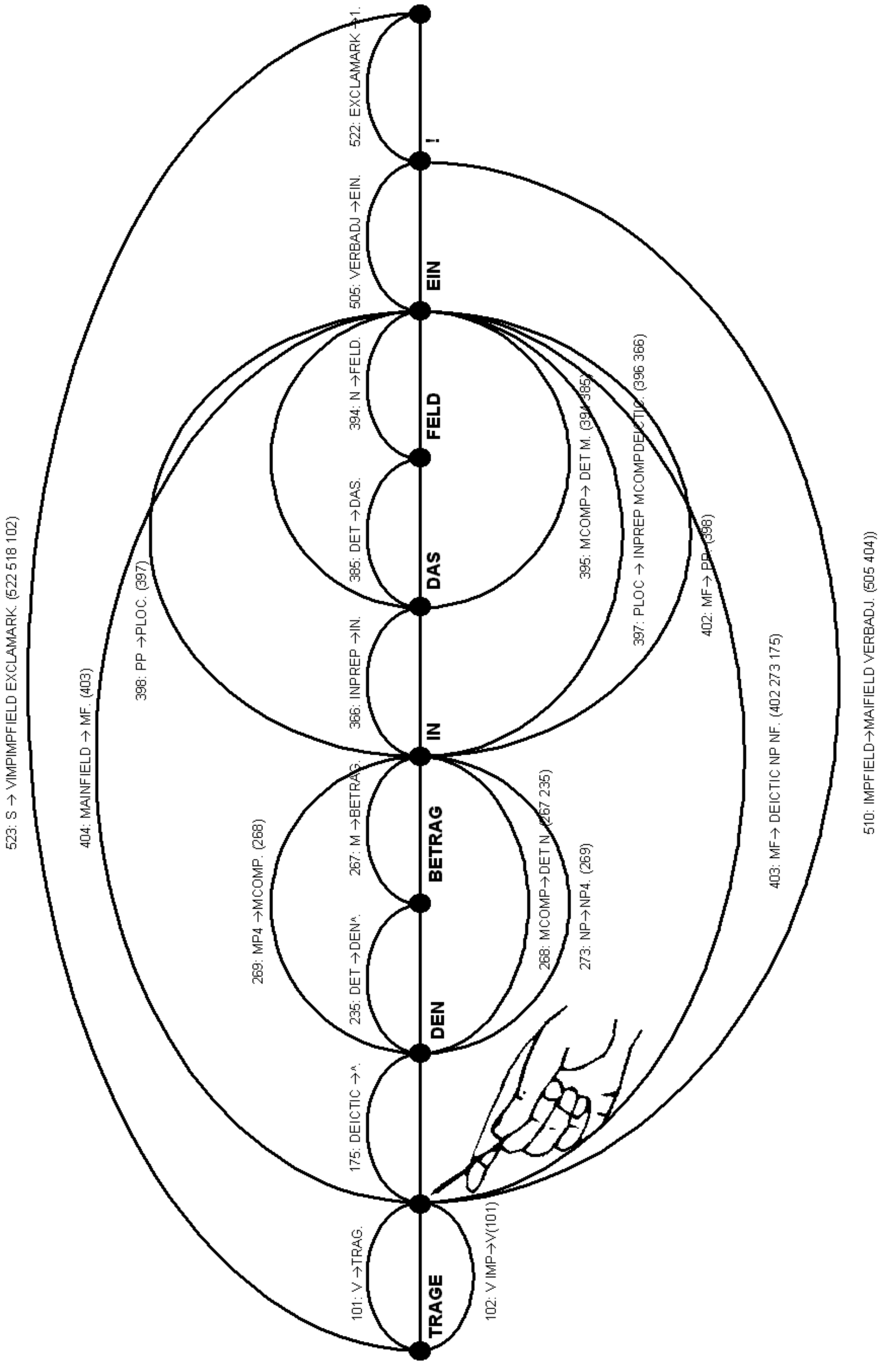
Ja, Gewerkschaftsbeitraege sind ab-
setzbar.
Yes, union dues are deductible.

As in face-to-face communication, there is no visual feedback after a successful referent identification process. Moreover, there are no predefined 'mouse-sensitive' areas, and the forms are not specially designed to simplify gesture analysis. For example, the regions on the form may overlap, and there may be several subregions embedded in a region of the form.

XTRA uses a unification-based parser for German, which is distinguished from similar parsers in that it is able to parse multimodal input. As Figure 3.3 shows, XTRA's parser treats pointing gestures as terminal symbols in the input stream. These symbols are then mapped onto the preterminal category "deictic".

In its full generality, the parsing of multimodal input is a complicated subject in its own right, and even a modest exposition of this topic would be beyond the scope of the present paper.

FIGURE 3.3
A CHART PRODUCED BY XTRA'S PARSER FOR MULTIMODAL INPUT



The syntax and semantics of the tax form are represented as a directed acyclic graph, called *organization graph*. It contains links to concepts in a terminological knowledge base encoded in SB-ONE, a representation language in the KL-ONE paradigm. The nodes of the organization graph represent various types of regions of the form, and the edges describe relations such as "geometrically embedded" or "conceptual part of". Four types of nodes are used in this graph:

- *Value regions*, where data can be entered by the user (e.g., the region where the number 78 has been typed in; see Figure 3.2)
- *Label regions*, which provide captions for value regions and framed regions (e.g., the string DM above the number 78 in Figure 3.2)
- *Framed regions*, which highlight rectangular parts of the form (e.g., the box containing the string DM in Figure 3.2)
- *Abstract regions* as aggregations of conceptually related, but not necessarily adjacent parts of the form (e.g., the column of three boxes above the DM box in Figure 3.2)

In addition to the direct interpretation of a gesture, where the demonstratum is simply identical to the referent, TACTILUS provides two other types of interpretation. In a pars-pro-toto interpretation of a gesture the demonstratum is geometrically embedded within the referent. In this case, the referent is either a framed region that contains smaller regions, or an abstract region. An extreme case of a pars-pro-toto interpretation in the current domain of XTRA is a situation where the user points at an arbitrary part (pars in Latin) of the tax form intending to refer to the form as a whole (pro toto in Latin). Another frequent interpretation of gestures is that the demonstratum is geometrically adjacent to the referent: the user points, for instance, below or to the right of the referent. Reasons for this may be inattentiveness or the attempt to gesture without covering up the data in a field.

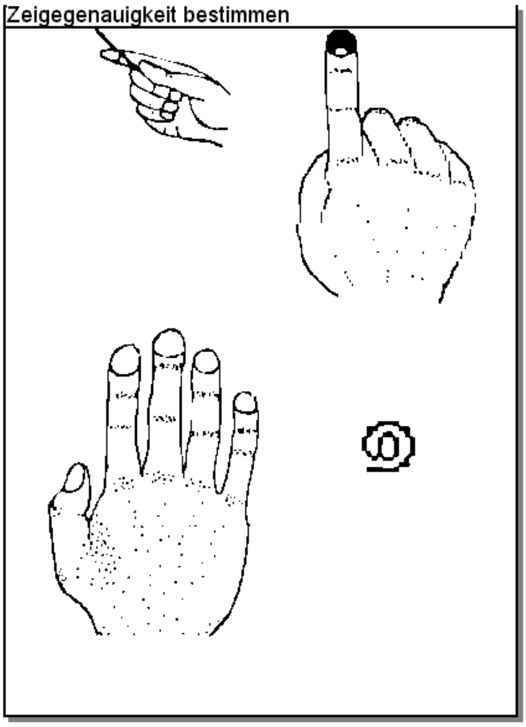
The user first chooses the granularity of the intended gesture by selecting the appropriate icon from the pointing mode menu or by pressing a combination of mouse buttons and then performs a tactile gesture with the pointing device symbolized by the selected mouse cursor. The current implementation supports four pointing modes (see Figure 3.4):

- Exact pointing with a pencil
- Standard pointing with the index finger
- Vague pointing with the entire hand
- Encircling regions with the '@'-sign

The deixis analyzer of XTRA is realized as a *constraint propagation* process on the organization graph described above. A pointing area of a size corresponding to the intended granularity of the gesture is associated with each available pointing mode. A plausibility value is computed for each referential candidate of a particular pointing gesture according to the ratio of the size of the part covered by the pointing area to the size of the entire region. The result of the propagation process is a list of referential candidates consisting of pairs of region names and plausibility values.

Since pointing is fundamentally ambiguous without the benefit of contextual information, this list often contains many elements. Therefore, TACTILUS uses various other knowledge sources of XTRA (e.g., the semantics of the accompanying verbal description, case frame information, the dialog memory) for the disambiguation of the pointing gesture (see [Allgayer86] and [Kobsa86] for further details).

FIGURE 3.4
THE POINTING MODE MENU



3.6 NATURAL VERSUS SIMULATED POINTING: SOME EMPIRICAL RESULTS

In order to evaluate the strengths and limitations of the deixis analyzer described above, an experiment[2] was carried out. The main objective of the experiment was to investigate the *similarities* and *differences* between the following:

- *Natural pointing* in face-to-face communication with an advisor
- *Simulated pointing* using the TACTILUS component of the XTRA interface

In this experiment, 32 subjects were asked to fill in two pages of the German income tax forms using data about a fictitious person. The information about this person was provided by the experimenter in textual form.

While the first page of the tax form was presented as a hard copy and was filled out using a pencil, the second page was displayed on the screen of a Lisp machine and was filled out using TACTILUS. The complete experiment consisting of 16 hours of dialog sessions was video- and audiotaped.

The tape transcriptions consist of an analysis of both the spoken and typed or written expressions and the accompanying gestures along with their temporal interdependency.

1200 gestures were identified and classified along the following dimensions (selection only):

device ::= pencil | finger | hand | mouse arrow (screen only)
movement ::= point | underline | encircle
exactness ::= precise | borderline | vague
directness ::= tactile | visual (hardcopy only)
location ::= exact | above | below | left | right

[2] It should be noted that in what follows we present only some preliminary results and that the final evaluation of all the data obtained from the experiment is not yet available. The experiment was designed by M. Wille with the help of D. Schmauks and Th. Pechmann.

Considering first the data for which there was no marked difference between natural and simulated pointing, two main results of the experiment should be noted:

- The low frequency (< 1%) of the following types of pointing gestures: using the hand/hand icon as a pointing device encircling

- The high frequency of pointing *below* the demonstratum.

Let us now turn to the results of the first part of the experiment, where the subjects used a pencil to fill in a hard copy of the tax form. The most important findings for natural pointing were the following:

- The high frequency of underlining (about 30%). The data showed the following order of frequency for the dimension "type of movement": point > underline > encircle.

- The preference of the subjects for using the pencil as a pointing device.

- The high frequency of pointing at the borderline of the demonstratum (about 36%).

- The frequency of using pointing device for focusing (see Section 3.7).

- The large percentage of visual pointing gestures (about 60%) as compared to tactile pointing gestures.

An encouraging result of the experiments with TACTILUS was that after a short training period (1-2 mins) even subjects without any computer experience were able to use the system to perform the specified task. There were two important observations in this part of the experiment:

- The low frequency of underlining (1.6%)
- A greater number of gestures (830) than in the natural setting (370)

It is quite clear that the higher frequency of pointing in the dialog sessions TACTILUS can be explained by the fact that in this setting the subjects had the additional task of positioning the input cursor, which required extra pointing.

Most of the design decisions for TACTILUS were supported by the findings from the experiment. It became evident that the important prerequisites for truly natural interaction in a multimodel mode are the following:

- *Context-sensitive interpretation* of pointing gestures (i.e., no one-to-one mapping of the demonstratum onto the referent)
- A *user model* and *discourse model* together with *assertional* and *terminological knowledge* for the interpretation of *ambiguous* pointing gestures

In addition, it was concluded that the ability to deal with a variety of tactile gestures, including different granularities and inexact pointing, to evaluate pointing gestures below the demonstratum and to cope with pars-pro-toto s is a positive feature of the current implementation that should be extended in future versions of the system.

On the other hand, the data suggest that in our current work on an roved version of TACTILUS the hand icon should be removed and that we need not make an effort to extend the mechanisms for the interpretation encircling gestures (e.g., by allowing circles around arbitrary polygons), these options were used extremely seldom. It also became clear that the interpretation of focusing gestures must be included in an improved version of the system, since this use of pointing was often observed in the natural setting but could not be handled by TACTILUS.

The high frequency of visual pointing gestures on the 2D tax forms showed that the 3D analysis of the position and orientation of the pointing device (e.g., by using a DataGlove™) is a promising direction for further improvements of the current system. With two DataGloves and the option to type with the gloves or to use speech input, truly parallel input becomes possible.

Finally, some comments are in order concerning the extent to which these findings may be generalized. The situation investigated in this study was highly restricted. The study was limited to 2D demonstrata with a permanent location. 3D and moving objects were excluded as targets for pointing actions. Another limitation of the present study concerned the nature of the experimental task, which was basically data entry. There is a large variety of situations in which people use deictic gestures that could not be studied in the present experiment.

3.7 THE INFLUENCE OF POINTING GESTURES ON THE DISCOURSE MODEL

In the experiment reported above, pointing was used not only for referent identification but also to mark or change the *dialogue focus*, for example, to control or shift *attention* during comprehension. As we noted in section 3.2, focus is an important notion in a discourse model, since it influences many aspects of language analysis and production. For example, focus can be used to disambiguate definite descriptions and anaphora [Grosz81].

Figure 3.5 gives an example of the disambiguation of a definite description using a focusing gesture. Without focus the definite description ‘the A’ is ambiguous in the given visual context, since three objects are visible which could be referred to as 'A' (one in each row of the table displayed in Figure 3.5). Together with the gesture of pointing at row Y, which marks this row as a part of the immediate focus, the definite description can be disambiguate since there is only one 'A' in the focused row.

As in the case of gestures for referent identification, the effect of a focusing gesture can also be produced by a *verbal paraphrase*. For the example presented in Figure 3.5, a meta-utterance like 'Now let's discuss the entries in row Y' would have the same effect on the discourse model and help disambiguate the subsequent definite description.

As we noted earlier, without a discourse context most pointing gestures are ambiguous. In the example above, we have seen that a discourse context can be established not only by verbal information but also by gestures. Thus there is a twofold relation between gestures and focus. Gestures can be used to shift focus, and focus can be used to disambiguate gestures.

From this it follows that in *simultaneous pointing actions* two communicative functions of pointing can be combined: focus shifting and reference. The following two types of simultaneous pointing can be identified:

- One-handed input:

  Focusing act: For example, the pencil is put down on the form, so that it points to a particular region on the form.

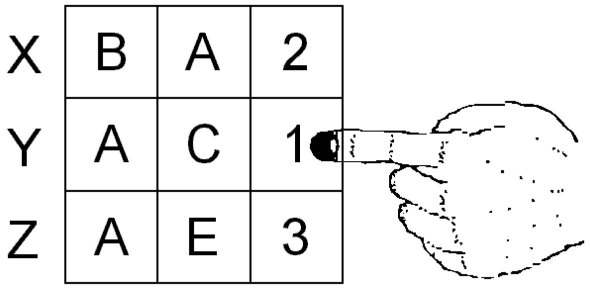
  Referential act: A subsequent pointing gesture refers to an object in the marked region.

- Two-handed input (see also [Buxton86]):

  Focusing act: For example, the index finger of one hand points to a region of the form.

  Referential act: The index finger of the other hand points to an object in the marked region.

FIGURE 3.5
FOCUSING GESTURE DISAMBIGUATING THE QUESTION "WHY SHOULD I DELETE THE ‘A’ ”



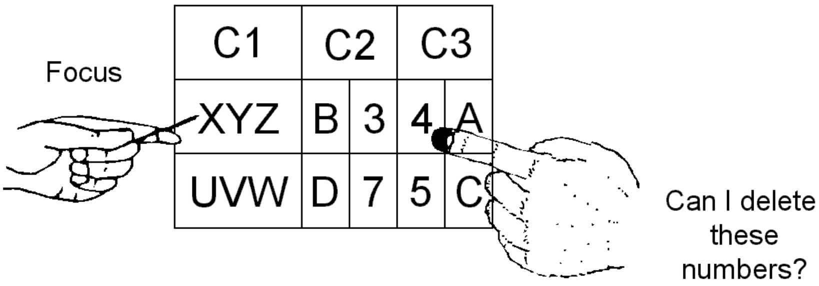
X	B	A	2
Y	A	C	1
Z	A	E	3

Figures 3.6 and 3.7 illustrate the use of focusing gestures for the disambiguation of referential gestures. Note that in both situations displayed in Figures 3.6 and 3.7 the index finger points at the same location on the

form and that the utterances combined with these referential gestures are identical. The cases shown in both figures differ only in the location of the pencil used for focusing.

Let us explore the processing of these examples in detail. Since the referential gesture with the index finger is relatively inexact, TACTILUS computes a large set of possible referents. The head noun, "numbers," of the verbal description that accompanies the pointing gesture imposes two restrictions on this set of possible referents. Since there are only four numbers displayed on the part of the form shown in Figures 3.6 and 3.7, the semantics of the noun restricts the solution space to the power set of {3, 4, 7, 5}, and the plural implies that only sets with at least two elements are considered in this power set. Finally, the position of the index finger on the form makes the interpretations {3, 7, 5}, {3, 4, 5}, {3, 4, 7} and {4, 5, 7} implausible, so that the resulting set of plausible referential readings becomes {{3, 4}, {4, 5}, {3,4,7,5}}, where {3, 4, 7, 5} is a typical example of a pars-pro-toto reading.

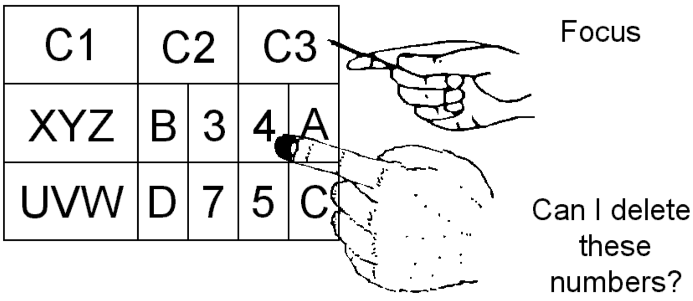
FIGURE 3.6
SIMULTANEOUS POINTING GESTURES



C1	C2		C3	
XYZ	B	3	4	A
UVW	D	7	5	C

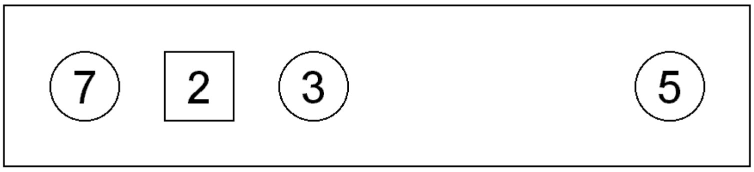
This means that there remain three possible interpretations before we consider the focusing gesture. It is worth noting that this is one of the cases where the combination of verbal and nonverbal information in one reference act does not lead to an unambiguous reading. Here information from the discourse model helps to clarify what is meant. In Figure 3.6 the pencil points at the row beginning with XYZ, so that this row and all its parts become focused. Now the intersection of the set of plausible referents and the currently focused objects results in the unique interpretation {3, 4}. Similarly, in Figure 3.7 the pencil is pointing at the block of columns called ‘C3’, so that the intersection of the focused elements with the results of the referential analysis is again a unique interpretation, namely {4, 5}, but it differs from the set of referents found for the gestural input shown in Figure 3.5. These examples once again emphasize the basic premise of our work, that pointing gestures must be interpreted in a highly context-sensitive way and that all approaches supposing a one-to-one mapping of the demonstratum onto the referent will fail in complex multimodal interactions.

FIGURE 3.7
SIMULTANEOUS POINTING GESTURES WITH DIFFERENT FOCUS



C1	C2		C3	
XYZ	B	3	4	A
UVW	D	7	5	C

FIGURE 3.8
INTRINSIC INTERPRETATION OF 'REPLACE THE BOX BY THE RIGHT CIRCLE'

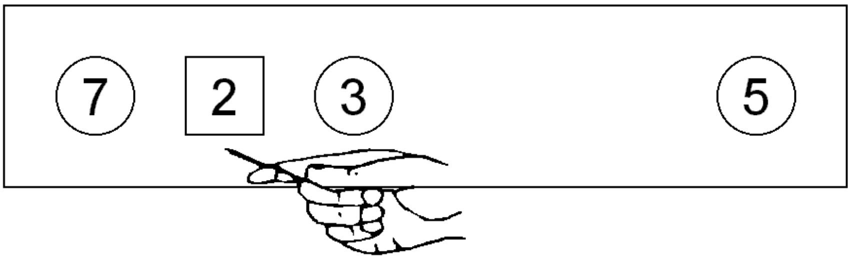


As we have seen, each focusing gesture modifies the discourse model. Another example of the impact of such focus information on the comprehension process is the effect of gestures on the selection of the *intrinsic* or *extrinsic use of spatial relations*. It is well known that the interpretation of spatial expressions depends on the selected frame of reference (for a more complete discussion of the intrinsic versus extrinsic use of spatial prepositions see [Retz-Schmidt88]).

One way to establish a reference frame is to use an intrinsic orientation. For example, consider the interpretation of the definite description 'the right circle' in Figure 3.8. Since there are three circles in the shared visual world of the user and the system, the interpretation of 'right' is crucial for finding the correct referent. In this case, the normal reading direction selecting from left to right forms the basis for an intrinsic interpretation as a default, selecting Circle 5 as the referent of the noun phrase.

Another way to establish a frame of reference is the use of a certain point of view for the extrinsic interpretation of spatial relations. The pointing gesture at Box 2 shown in Figure 3.9 overrides the default interpretation used for Figure 3.8. The focus information in the discourse model resulting from the gesture should cause the system to favor an interpretation where 'the right circle' refers to Circle 3. In this example, the pointer induces a reference frame for the interpretation of the spatial description.

FIGURE 3.9
EXTRINSIC INTERPRETATION OF 'REPLACE THE BOX BY THE RIGHT CIRCLE'



Note that in this situation the pointing gesture at Box 2 is redundant with respect to the referent identification process for the noun phrase 'the box'. Because only one box is visible, a unique referent can be determined without considering discourse information.

3.8 USER MODELING FOR PRESENTATION PLANNING

As we noted at the outset, an intelligent interface should be able not only to analyze multimodal input, but also to generate multimodal output. The design of XTRA's generator allows the simultaneous production of deictic descriptions and pointing actions [Reithinger87]. Because an intelligent interface should try to generate cooperative responses, it has to exploit its user model to generate descriptions tailored to users with various levels of expertise.

One important decision that a multimodal presentation planner has to make is whether to use a pointing gesture or a verbal description for referent identification. Let us use an example from our tax domain to explore the impact of the user model on this decision.

Suppose the system knows the concept 'Employee Savings Benefit' and an entry in the user model says that the current dialog partner seems to be unfamiliar with this concept. When the system plans to refer to a field in the tax form, which could be referred to using 'Employee Savings Benefit' as a descriptor, it should not use this technical term but a pointing gesture to the corresponding field. This means that in the conversational context described (3) would be a cooperative response, whereas (4) would be uncooperative.

(3) You can enter that amount here [↗] in this [↗] field.
(4) You can enter that amount as employee savings benefit.

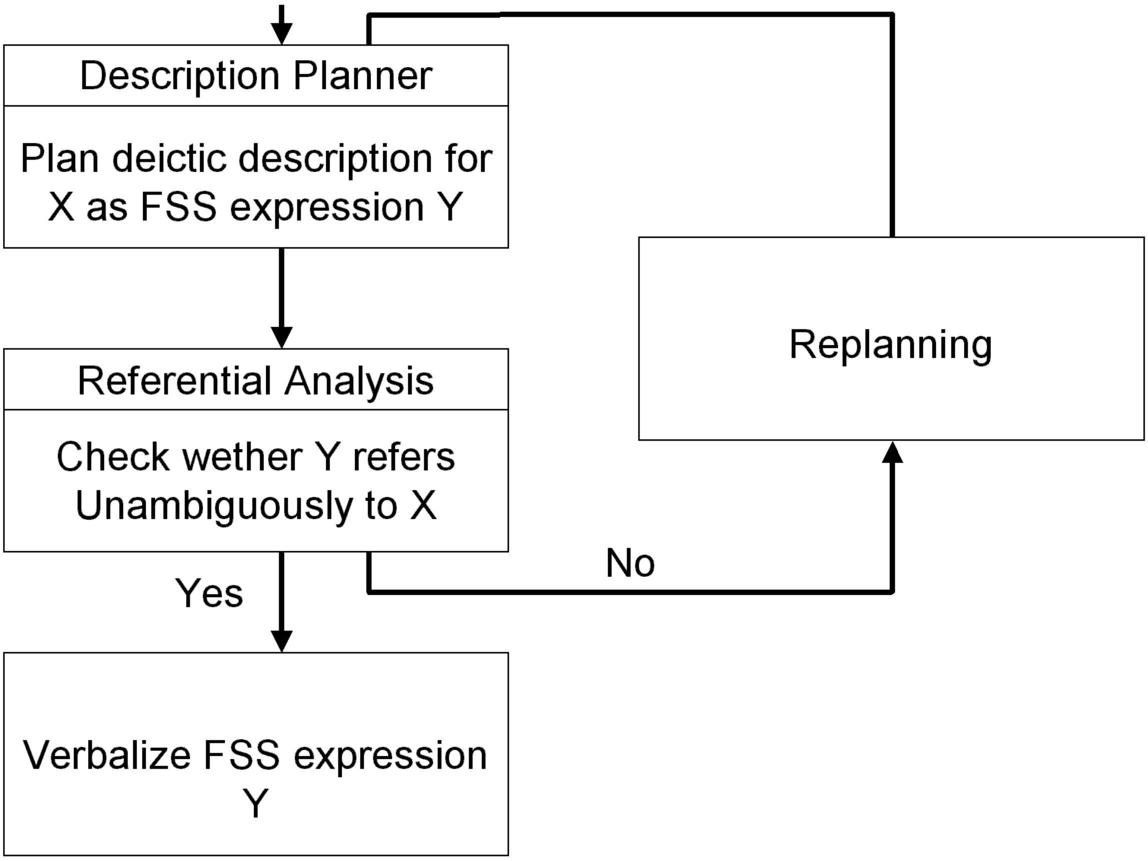
To summarize that point, if the system knows that a technical term which could be used to refer to a particular part of the tax form visible on the screen is not understandable to the user, it can generate a pointing gesture, possibly accompanied by a mutually known descriptor.

In the following, we discuss a particular method of user modeling, called *anticipation feedback*, that can help the system to select the right granularity of pointing when generating multimodal output. Anticipation feedback loops involve the use of the system's comprehension capability to simulate the user's interpretation of a communicative act that the system plans to realize [Wahlster86]. The application of anticipation feedback loops is based on the implicit assumption that the system's comprehension procedures are similar to those of the user. In essence, anticipation on the part of the system means answering a question like (5).

(5) If I had to analyze this communicative act relative to the assumed knowledge of the user, then what would be the effect on me?

If the answer to this question does not match the system's intention in planning the tested utterance, it has to replan its utterance, as in a generate-and-test loop. Figure 3.10 shows an extremely simplified version of a multimodal description planning process with an anticipation feedback loop for user modeling. Let us assume that the generator decided to plan a deictic description of an object X, to which the systems intends to refer. The result of the description planning process is a an expression Y of the functional-semantic structure (FSS) together with planned gesture. The FSS is a surface-oriented semantic representation language used on one of the processing levels of the how-to-say component of XTRA's generator.

FIGURE 3.10
AN ANTICIPATION FEEDBACK LOOP FOR PRESENTATION PLANNING


Description Planner
Plan deictic description for X as FSS expression Y
Referential Analysis
Check wether Y refers Unambiguously to X
Yes
No
Replanning
Verbalize FSS expression Y

This preliminary deictic description is fed back into the system's analysis component, where the referent identification component together with the gesture analyzer TACTILUS try to find the intended discourse object. If the system finds that the planned deictic description refers unambiguously to X, the description is fed into the final transformation process before it is outputted. Otherwise, an alternative FSS and/or pointing gesture has to be found in the next iteration of the feedback process (Figure 3.10).

Now let us use a concrete example to follow the feedback method as it goes through the loop. Suppose that the system plans to refer to the string 'Membership Fees' in the box shown in Figure 3.11. Also assume that the presentation planner has already decided to generate an utterance like 'Delete this [↗]' together with the pointing gesture shown in Figure 3.12.

For a punctual pointing gesture the system chooses the pencil as a pointing device. In this case, the exact position of the pencil was selected according to XTRA's default strategy described in [Schmaucks88]: the pencil is below the entry, so that the symbol does not cover it.

When this pointing gesture is fed back into the gesture analyzer of the referent identification component, the set of anticipated reference candidates might be {‘Fees’, ‘e’, ‘Membership Fees’} containing only elements that can be 'deleted' (the current version of TACTILUS does not deal with characters or substrings of a string). Since the system has detected that the planned gesture is ambiguous, it starts replanning and then selects the index finger icon as a pointing gesture with less granularity (Figure 3.12). This time, the result of the feedback process is unambiguous, so that the system can finally perform the pointing action.

FIGURE 3.11
PLANNED POINTING GESTURE

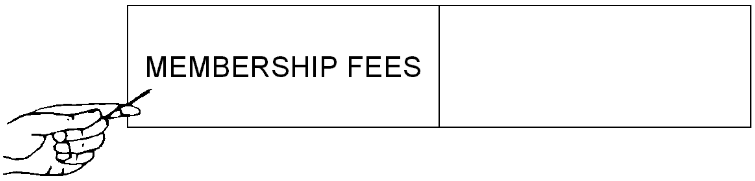
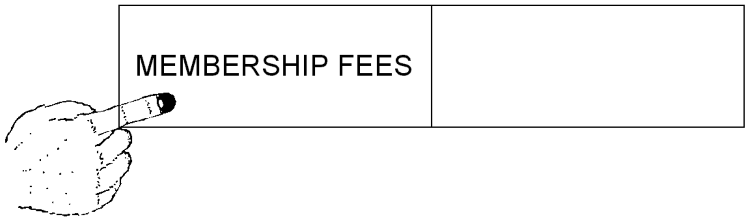


FIGURE 3.12
POINTING GESTURE AFTER REPLANNING



3.9 CONCLUSIONS

We have shown how the user and discourse models of an intelligent interface influence the comprehension and production of natural language with coordinated pointing to objects on a graphics display, and conversely how multimodal communication influences the user model and the discourse model.

First, we described XTRA as an intelligent interface to expert systems that handles a variety of tactile gestures, including different granularities, inexact pointing, and pars-pro-toto deixis, in a domain- and language-independent way. Then we discussed several extensions to the XTRA's deixis analyzer and presented our approach to generating multimodal output.

We showed how gestures can be used to shift focus and focus can be used to disambiguate gestures, so that simultaneous pointing actions combine two communicative functions: focus shifting and reference. We explored the role of user modeling for presentation planning and described how the user model can be exploited to generate multimodal descriptions tailored to the user's level of expertise.

Finally, we discussed anticipation feedback as a particular method of user modeling that can help the system to select the right granularity of pointing when generating multimodal output.

Some of the questions that have to be answered through future research on intelligent multimodal interfaces are the following:

- How can we deal with pointing gestures which refer to events? When there are dynamic and animated objects on the screen, the restriction of current prototypes, that is, that the presented visual material is fixed and static, is no longer viable.

- How can we handle pointing in 3D space? In the current systems the deictic space is two-dimensional and all objects are completely visible, so that tactile pointing is always possible.

- How can we cope with complex pointing actions, for example, a continuous movement of the index finger (underlining something, specifying a direction or a path) or a quick repetition of discrete pointing acts (emphatic pointing, multiple reference)?

REFERENCES

[Allgayer86] Allgayer, J., and Reddig, C. 1986. Processing Descriptions Containing Words and Gestures - A System Architecture. In Rollinger, C.-R. (ed.), *Proceedings GWAI/ÖGAI 1986*. Berlin: Springer, pp. 119-130.

[Bolt80] Bolt, R. A. 1980. Put-That-There: Voice and Gesture at the Graphics Interface. *Computer Graphics*, 14, 262-270.

[Brown79] Brown, D. C., Kwasny, S. C., Chandrasekaran, B., and Sondheimer, N. K. 1979. An Experimental Graphics System with Natural-Language Input. *Computer and Graphics*, 4, 13-22.

[Buxton86] Buxton, W., and Myers, B. A. 1986. A Study in Two-Handed Input. In *Proc. CHI'86 Human Factors in Computing Systems*, New York: ACM, pp. 321-326.

[Carbonell70] Carbonell, J. R. 1970. *Mixed-Initiative Man-Computer Dialogues*. BBN Report No. 1971. Cambridge, MA: Bolt, Beranek and Newman.

[Clark83] Clark, H. H., Schreuder, R., and Buttrick, S. 1983. Common Ground and the Understanding of Demonstrative Reference. *Journal of Verbal Learning and Verbal Behavior*, 22, 245-258.

[Grosz81] Grosz, B. 1981. Focusing and Description in Natural Language Dialogues. In Joshi, A., Webber, B., and Sag, I. (eds.) *Elements of Discourse Understanding*. New York: Cambridge Univ. Press, pp. 84-105.

[Hayes86] Hayes, P. J. 1986. Steps towards Integrating Natural Language and Graphical Interaction for Knowledge-based Systems. *Proceedings 7th European Conference on Artificial Intelligence*, Brighton, Great Britain, pp. 436-465.

[Hinrichs87] Hinrichs, E., and Polanyi, L. 1987. Pointing The Way: A Unified Treatment of Referential Gesture in Interactive Discourse. *Papers from the Parasession on Pragmatics and Grammatical Theory at the 22nd Regional Meeting*, Chicago Linguistic Society, Chicago, pp. 298-314.

[Kobsa86] Kobsa, A., Allgayer, J., Reddig, C., Reithinger, N., Schmauks, D., Harbusch, K., and Wahlster, W. 1986. Combining Deictic Gestures and Natural Language for Referent Identification. In *Proceedings 11th International Conf. on Computational Linguistics*, Bonn, Germany, pp. 356-361.

[Kobsa89] Kobsa, A., and Wahlster, W. (eds.), 1989. *User Models in Dialog Systems*. New York: Springer.

[Neal91] Neal, J. G., and Shapiro, S. C. 1991. Intelligent Multi-Media Interface Technology. In present volume.

[Reilly85] Reilly, R., (ed.) 1985. *Communication Failure in Dialogue: Techniques for Detection and Repair*. Dublin, Ireland: Deliverable 2, Esprit Project 527, Educational Research Center, St. Patrick's College.

[Reithinger87] Reithinger, N. 1987. Generating Referring Expressions and Pointing Gestures. In Kempen, G. (ed.) *Natural-Language Generation*, Dordrecht: Kluwer, pp. 71-81.

[Retz-Schmidt88] Retz-Schmidt, G. 1988. Various Views on Spatial Prepositions. In *AI Magazine*, 9 (2) 95-105. Also appeared as: Report No. 33, SFB 314, University of Saarbrücken, Computer Science Department.

[Schmauks87] Schmauks, D. 1987. Natural and Simulated Pointing. In *Proceedings 3rd European ACL Conference*, Copenhagen, Denmark, pp. 179-185.

[Schmauks88] Schmauks, D., and Reithinger, N. 1988. Generating Multimodal Output - Conditions, Advantages, and Problems. In *Proceedings 12th International Conference on Computational Linguistics*, Budapest, Hungary, pp. 584-588.

[Scragg87] Scragg, G. W. 1987. *Deictic Resolution of Anaphora*. Unpublished paper, Franklin and Marshall College, P.O. Box 3003, Lancaster, PA 17604.

[Thompson86] Thompson, C. 1986. Building Menu-Based Natural Language Interfaces. *Texas Engineering Journal*, 3, 140-150.

[Wahlster84] Wahlster, W. 1984. Cooperative Access Systems. *Future Generation Computer Svstems*, 1, 103-111.

[Wahlster86] Wahlster, W., and Kobsa, A. 1986. Dialog-Based User Models. In Ferrari, G. (ed.), *Proceedings of the IEEE*, 74 (7), 948-960.

[Wahlster88] Wahlster, W. 1988. Distinguishing User Models from Discourse Models. In Kobsa, A., and Wahlster, W. (eds.) *Computational Linguistics*. Special Issue on User Modeling, 14 (3), 101-103.

[Wetzel87] Wetzel, R. P., Hanne, K. H., and Hoepelmann, J. P. 1987. *DIS-QUE: Deictic Interaction System-Query Environment*. LOKI Report KR-GR 5.3/KR-NL 5, Stuttgart, Germany: Fraunhofer Gesellschaft, IAO.

[Woods79] Woods, W. A., et al. 1979. *Research in Natural Language Understanding*. Cambridge, MA: Annual Report, TR 4274, Bolt, Cambridge, MA: Beranek and Newman.

[Zimmermann87] Zimmermann, T. G., Lanier, J., Blouchard, C., Bryson, S., and Harvill, Y. 1987. A Hand Gesture Interface Device. *Proceedings CHI'87 Human Factors in Computing Systems*, New York: ACM, pp. 189-192.