Generating Multimodal Output – Conditions, Advantages and Problems

Dagmar SCHMAUKS, Norbert REITHINGER
SFB 314: Künstliche Intelligenz – Wissensbasierte Systeme
Universität des Saarlandes, FB 10 – Informatik IV
D-6600 Saarbrücken 11
Federal Republic of Germany
Net: schmauks@sbsvax.uucp, bert@sbsvax.uucp

Abstract

In natural communication situations, multimodal referent specification is frequent and efficient. The linguistic component are deictic expressions, e.g. 'this' and 'here'. Extralinguistic devices in dialogs are different body movements, mainly pointing gestures. Their functional equivalent in texts are means like arrows and indices.

This paper has two intentions. First, it discusses the advantages of multimodal reference in interhuman communication which motivate the integration of extralinguistic "pointing" devices into NL dialog systems. The generation of multimodal output poses specific problems, which have no counterpart in the analysis of multimodal input. The second part presents the strategy for generating multimodal output which has been developed within the framework of the XTRA system (a NL access system to expert systems). XTRA allows the combination of verbal descriptions and pointing gestures in order to specify elements of the given visual context, i.e. a form displayed on the screen. The component POPEL generates referential expressions which may be accompanied by a pointing gesture. The appearance of these gestures depends on several factors, e.g. the type of referent (whether it is a region or an entry of the form) and its complexity.

Acknowledgements

The work presented here is being supported by the German Science Foundation (DFG) in its Special Collaborative Program on AI and Knowledge-Based Systems (SFB 314), project N1 (XTRA). We would like to thank our colleagues of the XTRA project for their helpful comments on an earlier version of this paper.

1. Introduction

In face-to-face communication, speech and communicative body movements are performed simultaneously. A prime example ot this *multimodality* are deictic actions which specify elements of a shared visual world by the combination of deictic expressions ('this', 'there' etc.) and extralinguistic devices like pointing gestures. The advantages of this *multimodal deixis* motivate the integration of extralinguistic means for referent specification into natural language (NL) dialog systems. Starting point of the following considerations is the system XTRA, an NL access system for expert systems, which is under development at the University of Saarbrücken. In its current application domain, it assists the user in filling out a tax form which is visible on the screen. Elements of this form can be specified not only by (typed) verbal descriptions, but also by combining descriptions and simulated pointing gestures. Some problems of *multimodal input* and solutions in XTRA have already been treated in detail (cf. /Allgayer, Reddig 86/, /Allgayer et al. 88/, /Schmauks 86a, 87/).

Multimodal output is no simple mirror image of multimodal input. Rather, it has to deal with different problems the investigation of which has been missing till now (for a first impression see /Reithinger 87a/). Because of the novelty of the task, one cannot claim to offer ultimate solutions. Instead, we wish to outline several approaches for the realization of multimodality, present our strategy and give reasons for the choice.

In section 2, we present the means, conditions and advantages of multimodal deixis within natural communication situations. Topics of section 3 are the different strategies for realizing multimodality in NL dialog systems and some of the problems arising. Section 4 sketches the framework of the XTRA system and the types of gestures occuring in this domain. Section 5 presents the generation component POPEL[1], focussing on its global strategy for generating multimodal output. Subtopics are POPEL's architecture and its methods for simulating different types of pointing gestures. In section 6, some alternative strategies for generating multimodal output are briefly discussed.

1) POPEL is the acronym for *Production Of (Perhaps, Possible, ..) Eloquent Language.*

2. Deixis in natural communication situations

Deictic reference occurs in dialogs as well as in texts. In both situations, the objects referred to can be linguistic entities (sentences, chapters etc.) or non-linguistic objects (cats, tables etc.). For the following considerations, only those types of deixis are relevant which specify non-linguistic entities. They can be performed by combining linguistic expressions with extralinguistic devices.

Dialogs are characterized by the possibility of turn-taking. If both participants are present, they can specify elements of their common visual world by combining deictic expressions and body movements, mainly pointing gestures. If a speaker can point to objects, s/he can use shorter, simpler and even referentially insufficient descriptions. In particular, pointing facilitates reference if the speaker doesn't know how to describe the object in question. One example is the utterance

THIS [☞] is broken.

while pointing at some part of the engine of one's car[2].

Successful reference by pointing has some preconditions, for instance the receiver's visual attention. S/he has to face the speaker in order to notice his/her gesture and then has to follow this gesture with his/her gaze. The first step can fail by visual inattentiveness, the latter by wrong direction of gaze. *Feedback* is received by speakers via two channels. On the one hand, a speaker controls the *nonverbal reaction* of the receiver and can therefore immediately request attentiveness or correct a wrong direction of gaze. On the other hand, s/he gets delayed feedback by the *verbal reaction*.

Communication by *text* normally implies a spatial and temporal dissocation of sender (=writer) and receiver (=reader). Therefore, the sender can *deictically* refer only to non-linguistic entities which are visible also for the receiver. This condition is fulfilled if the text is combined with non-linguistic representations (pictures, diagrams, maps etc.). In these cases, the sender can refer to elements of this 'visual context' by combining linguistic expressions and extralinguistic means (arrows, indices etc.). The latter represent a functional equivalent to pointing gestures within dialogs and have the same advantages. But, like the text itself, they don't require attentiveness on the reader's side during the period of their production.

3. Deixis in NL dialog systems

The type of dialog considered here is a *consultation dialog:* the system (= expert) assists the user (= non-specialist) in filling out his/her tax form. The system has not only more expert knowledge about the domain, but also more knowledge concerning content and structure of the graphics displayed on the screen.

Due to these differences in knowledge, the *analysis component* has to deal with shortcomings in the user's input. His/her pointing gestures may be imprecise because s/he doesn't know the structure of the presented graphics. Ignorance of technical terms results in inadequate descriptions. In theses cases, additional knowledge sources are needed for referent identification, e.g. case frame analysis and dialog memory (/Allgayer, Reddig 86/, /Allgayer et al. 88/). In contrast, the *generation component* can always produce precise pointing gestures as well as exact descriptions. But the latter capability may be in conflict with the task of generating system reactions which are communicatively adequate. If the user doesn't know certain technical terms, then the combination of underspecified description and precise gesture is more comprehensible than a totally specified description.

2) Pointing gestures are represented by the sign ' [☞] '
Capital letters highlight the correlated phrase.

Another problem is the different perceptual capabilities of user and system. Humans are 'multichannel systems' which receive information about objects through a great variety of channels. In contrast, the perceptible world of all systems developed to date is only a small subset of the user's world. Normally, systems with more general application domains are only able to process textual and graphical input. In particular, these systems cannot "see" the user's nonverbal behavior and therefore cannot request attention if necessary. Also, wrong user reactions cannot serve as an indication of his/her visual inattentiveness, because they can be caused by several other factors. For example, it might be the case that the user has correctly identified the field in question but enters a wrong amount because s/he has confused some technical terms. During natural pointing, the sound which occurs when the speaker touches the form may cause the hearer to pay attention to his/her gestures. But in the case of simulated pointing, the generation of a specific audible signal in parallel to each pointing gesture implies a rather "unnatural" situation.

The design of multimodal interfaces is one central topic of recent research. It has to be emphasized that the term 'multimodal input/output' covers a great variety of heterogeneous phenomena from the manipulation of simulated objects within an "artificial reality" (e.g. the *DataGlove*, see /Zimmerman et al. 87/) to the use of different pointing devices.

The goal 'multimodal referent specification' can be achieved by various strategies. If one wants to *simulate natural pointing*, the pointing device should correspond to natural gestures. A touch-sensitive screen allows highly natural gestures, but pointing by means of a so-called 'mouse cursor' can also simulate some aspects of natural pointing. The latter strategy is chosen in the XTRA system. If, in contrast, one wants to offer *functional equivalents*, there exists a great variety of devices. It is possible to adapt the extralinguistic deictic means which occur in texts, e.g. arrows and indices. Furthermore, the computer offers several specific devices, which have no model in natural pointing, such as framing, highlighting or inverting the referent. The choice depends on several factors, for example which types of objects are to be referred to.

4. Form deixis in XTRA

The given visual context of the XTRA system is the form displayed on the screen. In order to specify its elements, several types of pointing actions occur (cf. /Allgayer 86/, /Schmauks 86a, 86b, 87/):

- *Punctual pointing* indicates one singular point on the form and can be produced in order to specify primitive objects, i.e. individual regions and individual entries. Another possibility is the reference to a complex region by pointing to a part of it (*pars-pro-toto deixis*).
- During *non-punctual pointing*, the pointing device performs a complex motion, e.g. underlines an entry or gives the borders of a larger region.
- *Multiple pointing* means, that one utterance is accompanied by more than one pointing gesture. These complex pointing actions specify elements of sets, for example several instances of one concept.

One aim of XTRA is the use of multimodal referent specification techniques in input as well as in output. *Multimodal input* is performed by combining typed NL descriptions and simulated pointing gestures. The latter are currently realized by means of a mouse cursor. They simulate natural pointing with regard to two aspects: the user can select the accuracy of gesture, and the relation between the gesture and the object referred to depends on context /Allgayer 86/. For example, if the user points at a region which is already filled out, descriptor analysis is necessary in order to decide whether s/he refers to the region itself or to its actual entry.

The generation component has to reckon with different problems concerning pointing actions. If it also realizes gestures by movements of a mouse cursor, their perception may be hampered by the user's visual inattentiveness. In the case of multiple pointing, for example, s/he might fail to notice one of the pointing gestures and consequently may not identify the referent. This causes the whole utterance (e.g. 'THIS AMOUNT [☞], you could also enter HERE [☞]') to become incomprehensible.

5. Generation of pointing actions with POPEL

5.1 Architecture of POPEL

The task of POPEL, the natural language generation component of XTRA, is to select and verbalize those parts of the conceptual knowledge base that are to be uttered. The structure of the component follows the well-known division into a "what-to-say" and a "how-to-say" part /McKeown 85/: POPEL-WHAT, which *selects* the content, and POPEL-HOW, which *verbalizes* it (cf. /Reithinger 87b/). Contrary to most other systems, the information flow between these two sub-modules is not unidirectional from the selection part to the verbalisation part. Rather, both parts communicate while processing the output of the system (cf. /Hovy 87/).

A second essential feature of POPEL's architecture is the parallel processing approach to generation: the different stages of selecting and realizing the output proceed in a parallel cascade. In this way, it is possible to go ahead with the selection processes inside POPEL-WHAT, while a previously selected part of the utterance is already verbalized in POPEL-HOW. As one consequence, restrictions to the selection arising out of the verbalization process can be taken into account.

Currently, a first prototype of POPEL is under development. The processor for the parallel cascade has already been implemented. The emphasis was placed on information propagation both upwards and downwards and on the definition of the syntax and semantics of the transition rules. The next step will be the encoding of knowledge within this framework. POPEL is implemented on a Symbolics 3640 Lisp machine running Zetalisp.

5.2 Pointing gestures as special cases of descriptions

5.2.1 Selection of descriptions

Selection of descriptions is one essential interaction point between the two components. Decisions which concern POPEL-WHAT are:

- "Givenness" of an object: the description of an object depends on whether that object is known in the (implicit ot explicit) context of the user. In general, POPEL-HOW selects definite phrases for known objects and indefinite phrases for unknown objects, but the required knowledge as to "givenness" is stored in the user model which is accessed by POPEL-WHAT.
- "Pointability" of an object: the so called 'form hierarchy' represents the structure of the form. It links the regions of the form to the respective representations in the conceptual knowledge. If an object is selected for verbalization, the link from the concept of the object to the form hierarchy provides the information that a pointing gesture can be generated.
- Situation-dependency of a description: the contextual knowledge bases contain structure and content of the previous dialog. They allow the determination of differently detailed descriptions, depending on the current context. If necessary, meta-communicative or text-deictic attributes can be added.

POPEL-HOW makes the following decisions:

- Generation of a description: whether an object in the conceptual knowledge base is to be realized as a description depends on the language-related structure that has already been determined.
- Language-dependent constraints: the possible surface structures remaining for a description depend on the extent to which the sentence has already been verbalized. In German, for instance, it is hardly possible to generate a pronominal NP if there is already a lexical NP or PP after the finite verb and the pronominal NP is to follow this phrase (cf. /Engel 82/).

The sequence of these decisions is intertwined. For example, the inquiry of POPEL-WHAT, as to whether an object is available in the context makes sense only after POPEL-HOW has decided to generate a description at all (for an outline see /Reithinger 87a/).

5.2.2 When to point

From the viewpoint of an NL dialog system, pointing actions are descriptions which are accompanied by a pointing gesture. They focus the user's visual attention and can therefore localize visible objects. In the XTRA domain, pointing actions can refer to three types of objects:

- A *form region*, e.g. 'You can enter your donations HERE [☞].'
- An *entry*, e.g. 'THESE 350 DM [☞] are travel expenses.'
- A correlated *concept*, e.g. 'Can I deduct SUCH DONATIONS [☞]?'

All elements of the form are in the shared visual context; therefore, they can be referred to by definite descriptions. No serious problems arise if an utterance is accompanied only by *one* pointing gesture. In contrast, the simulation of *multiple* pointing requires further considerations (cf. section 4) and has therefore not been treated in this paper.

If the system's reaction contains more than one description which allows pointing, only one possibility will be realized. The others are reduced to purely verbal descriptions. The sentence (1) for example allows the reductions (1a) and (1b):

(1) THIS AMOUNT [☞], you have to enter HERE [☞].

(1a) *The donations of 15 DM*, you have to enter HERE [☞].

(1b) THIS AMOUNT [☞], you have to enter *in the line 'donations'*.

Because sentence generation is performed incrementally, POPEL-WHAT doesn't know the whole content of the utterance at the moment it has to decide whether to use a pointing gesture or not. Therefore, the decisions have to be based on heuristics and may be "suboptimal". One of these heuristics is: do not use a pointing gesture if the object in question can also be specified by a short referential expression, for example a pro-word. Then the pointing gesture remains available to reduce a complex description if it follows in the same utterance.

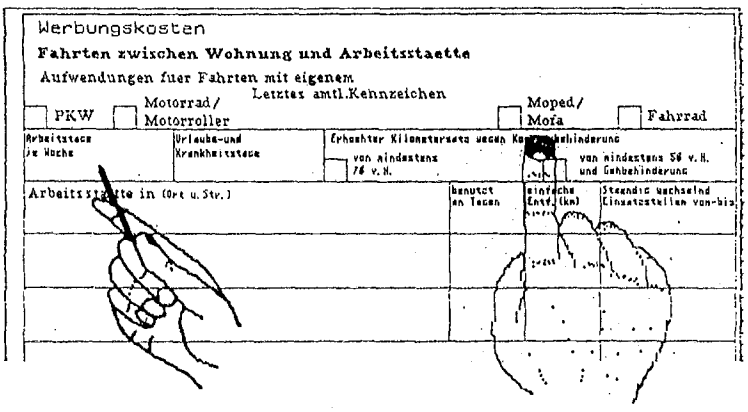
5.2.3 How to point

Following the simulation-oriented strategy of XTRA, pointing gestures are realized by positioning a mouse cursor on the screen. This is a close approximation of the type of movements a human performs when pointing with his/her finger. Furthermore, different degrees of accuracy are simulated by different shapes of the cursor. POPEL performs the pointing gesture parallel with verbalizing the correlated phrase and presenting it on the screen.

5.2.3.1 Punctual pointing gestures

During a *punctual pointing gesture*, the cursor doesn't move on the form. This type of gesture is used both for the localization of primitive objects as well as for pars-pro-toto deixis. Because a gesture can refer either to a field of the form or to its content (i.e. a string in our domain), the linguistic information (e.g. 'this field' vs. 'this amount of money') has to disambiguate between these possibilities. A hand which holds a pencil is used as the symbol for this type of gesture (see figure 1/symbol A). The exact position depends on the type of the object. The default strategy is as follows: if the pointing action refers to a field, the pencil is *in the middle of the field*, if it refers to an entry, the pencil is *below the entry*, so that the symbol doesn't cover it. Additionally, the user model takes effect: if the user requested another position of the gesture repeatedly (e.g. 'Take away the finger, I cannot read that!'), the pointing strategy has to be changed.

Each time the speaker-hearer roles are reversed, the current pointing symbol changes to a neutral symbol (i.e. the standard mouse cursor). In this way, the user's visual attention doesn't remain fixed to the location of the last pointing gesture. If the system generates a new pointing gesture, it first changes the neutral symbol into the choosen pointing symbol. Then it moves the symbol to the new pointing location. This method mimics the functionality of the movements of the hand during natural pointing, which already direct the hearer's visual attention to the target location.



A B

Figure 1: Different types of pointing gestures

Furthermore, punctual pointing gestures are used to realize pars-pro-toto deixis, which refers to greater parts of the form. In this case too, the ambiguity of the gesture has to be compensated by linguistic information. In our domain, unambiguous descriptions are 'row' and 'column'. Ambiguous expressions like 'region' can be disambiguated by additionally naming the referent, e.g. 'the region of DEDUCTIBLES'.

Delayed perception of a punctual pointing gesture doesn't hamper referent identification. The pointing symbol changes only when the user takes initiative in the dialog again. Until then, the information of the gesture remains visually available. There exists an equivalent in natural pointing: it might happen that a speaker leaves his/her forefinger extended, until the dialog partner recognizes the gesture.

5.2.3.2 Non-punctual pointing gestures

Non-punctual pointing, for example the encircling of a whole area, poses much greater problems. After the movement of the cursor ceased, the actual cursor position indicates only the final point of the gesture. If the user was inattentive, s/he cannot reconstruct the course of movement. This loss of information can be partially avoided by providing exact descriptions.

Standard candidates for non-punctual pointing actions are composite objects, for example rows, columns or larger regions. However, a *non-punctual pointing gesture* that has not been noticed does not deliver any more information than the combination of *punctual pars-pro-toto deixis* and an exact linguistic description.

Non-punctual pointing gestures can be realized by various means. In a first release of POPEL, the gesture is performed with another symbol (hand with stretched-out forefinger, see figure 1/symbol B). The movement should be both "natural" as well as relatively precise. Further research has to evaluate POPEL's current strategy with respect to various features, for example the efficiency of the pointing strategy and its acceptance by the user.

6. Alternative concepts for multimodal input/output and future requirements

In the case of non-punctual and multiple pointing actions, the possible inattentiveness of the user and the current "blindness" of the system may lead to a loss of information. This danger increases with the temporal complexity of the gestures. The usage of "lasting" pointing techniques would be one possibility of dealing with this problem.

One strategy is to "freeze" the track of non-punctual pointing gestures. This is similiar to underlining or encircling with a pencil. The track remains visible on the form until the next change in dialog control. One can imagine two variants of this strategy: the first is the *successive* drawing of the line, which is similiar to a human-made gesture. Also the drawing speed could be adopted from natural drawing. The second variant is to produce the whole line *simultaneously*.

But this *freezing method* has the essential shortcoming that the additional lines muddle the screen. Therefore, the functionally similar but "unnatural" means of referent specification (framing, underlaying, blinking, inverting etc.) seem to be more advantageous. They preserve the form's structure since it is not blurred by additional lines. Furthermore, these methods specify form regions, i.e. rectangular objects, more exactly than circular lines. On the other hand, however, this *framing approach* cannot simulate the context-dependency of natural pointing.

One unsolved problem remains to be emphazised: all the aforementioned methods *alone* cannot solve the problems of multiple pointing. If the sequence of the gestures must be known in order to understand the utterance, the frames etc. have to be combined with additional means. One solution could be the adaptation of methods used in texts in order to refer to elements of graphics (e.g. indices, cf. section 2).

A highly user-adapted generation of pointing actions would require the storage of information about pointing in the user model. On the one hand, these are facts about the *user's pointing behavior*, including frequency and accuracy of gestures and possible systematic deviations (e.g. pointing consistently beside or below the intended referent). On the other hand, the generation component has to take into account the *user's reaction to the system's pointing actions*. If s/he repeatedly misunderstands such an action, the system has to modify its pointing strategy and switch to the fixation method or to the framing approach, for example.

References

Allgayer, J. (1986): Eine Graphikkomponente zur Integration von Zeigehandlungen in natürlichsprachliche KI-Systeme. Proceedings der 16. GI-Jahrestagung. Berlin: Springer.

Allgayer, J. and C. Reddig (1986): Processing Descriptions containing Words and Gestures – A System Architecture. In: C.-R. Rollinger (Hrsg.): GWAI/ÖGAI 1986. Berlin: Springer.

Allgayer, J., K. Harbusch, A. Kobsa, C. Reddig, N. Reithinger, D. Schmauks (1988): XTRA – A Natural-Language Access System to Expert Systems. Technical Report, SFB 314, FB Informatik, Universität des Saarlandes, Saarbrücken.

Engel, U. (1982): Syntax der deutschen Gegenwartssprache. Berlin: Erich Schmidt.

Hovy, E.H. (1987): Generating Natural Language Under Pragmatic Constraints. Ph.D. Dissertation, Yale University.

McKeown, K.R. (1985): Text generation. Cambridge: Cambridge University Press.

Reithinger, N. (1987a): Generating Referring Expressions and Pointing Gestures. In: G. Kempen (ed.): Natural Language Generation. Dordrecht: Kluwer.

Reithinger, N. (1987b): Ein erster Blick auf POPEL -- Wie wird was gesagt? In: K. Morik (ed.): GWAI 87. Berlin: Springer.

Schmauks, D. (1986a): Formulardeixis und ihre Simulation auf dem Bildschirm. Ein Überblick aus linguistischer Sicht. Memo Nr.4, SFB 314, FB Informatik, Universität des Saarlandes, Saarbrücken.

Schmauks, D. (1986b): Form und Funktion von Zeigegesten. Ein interdisziplinärer Überblick. Report Nr. 10, SFB 314, FB Informatik, Universität des Saarlandes, Saarbrücken.

Schmauks, D. (1987): Natural and Simulated Pointing. Proceedings of the 3rd European ACL Conference, Kopenhagen, Danmark. Also: Report Nr. 16, SFB 314, FB Informatik, Universität des Saarlandes, Saarbrücken.

Zimmerman, T.G. et al. (1987): A Hand Gesture Interface Device. Proc. CHI'87 Human Factors in Computing Systems. ACM, New York.